



UNIVERSIDAD NACIONAL DE SAN AGUSTÍN
ESCUELA DE POSGRADO
UNIDAD DE POSGRADO DE LA FACULTAD DE INGENIERÍA DE PRODUCCIÓN Y
SERVICIOS

GENERACIÓN GENÉTICA MULTIOBJETIVO DE BASES
DE CONOCIMIENTO FUZZY PARA CLASIFICACIÓN EN
BASES DE DATOS NO BALANCEADAS USANDO EL
ENFOQUE ITERATIVO

Tesis presentada por el Magister Edward Hinojosa Cárdenas
para la obtención del grado de Doctor en Ciencias de la Computación

Orientador: Dr. Yván Jesús Túpac Valdivia
Co-Orientadora: Dra. Heloisa de Arruda Camargo

Arequipa, Agosto del 2015



UNIVERSIDAD NACIONAL DE SAN AGUSTÍN
ESCUELA DE POSGRADO
UNIDAD DE POSGRADO DE LA FACULTAD DE INGENIERÍA DE PRODUCCIÓN Y
SERVICIOS

GENERACIÓN GENÉTICA MULTIOBJETIVO DE BASES
DE CONOCIMIENTO FUZZY PARA CLASIFICACIÓN EN
BASES DE DATOS NO BALANCEADAS USANDO EL
ENFOQUE ITERATIVO

Tesis presentada por el Magister Edward Hinojosa Cárdenas
para la obtención del grado de Doctor en Ciencias de la Computación

Orientador: Dr. Yván Jesús Túpac Valdivia
Co-Orientadora: Dra. Heloisa de Arruda Camargo

Arequipa, Agosto del 2015

Agradecimientos

A Dios, por darme la vida, salud, fuerza y fe que necesitamos en cada momento para seguir adelante y ser felices.

A mis padres, Edwin y Reina, que me han amado, cuidado y apoyado cada instante de mi vida.

A mi hermano Edwin, que es el compañero de mi vida, el me ha querido y cuidado desde siempre.

A mi novia Vanessa, con la que he pasado momentos lindos e inolvidables, y deseo que me acompañe el resto de mi vida.

A mi cuñada Eveline y a mis sobrinos Dianita, Edwincito y Fabianita, que llegaron para alegrar nuestro hogar y con cada sonrisa nos dan fuerzas para seguir adelante.

A mi tía Luz, a mi abuelita Doris (Do) y a mi abuelito Alberto (Bate - en memoria) , que junto a mis padres, me han amado, cuidado y apoyado desde que nací.

A mi abuelita Aurora (mamá Aurora - en memoria) y mi abuelito Grimaldo (papá Grimaldo - en memoria), que en tierra y ahora desde el cielo me guían y protegen.

A mi orientador Yván Tupac y a mi co-orientadora Heloisa Camargo, que me orientaron atenciosa y efectivamente para la realización de este trabajo, y de los cuales aprendí, aprendo y aprenderé mucho.

Al Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica del Perú (CONCYTEC) y al Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT), por el apoyo y financiamiento de este trabajo. El documento que sustenta el apoyo financiero es el Contrato de Subvención N° 315-2012-CONCYTEC-OAJ en el marco de la Cátedra CONCYTEC - UNSA, Arequipa, que mediante Resolución de Presidencia N° 161-2012-CONCYTEC-P, fui ganador de una de las becas de doctorado en "Tecnología de la Información y Comunicación - orientada al desarrollo de Software" de la Universidad Nacional San Agustín de Arequipa.

Atte,

Edward Hinojosa Cárdenas

Agosto 2015

Resumen

Los sistemas fuzzy han sido utilizados para resolver diversos tipos de problemas, entre ellos los problemas de clasificación, dado que pueden obtener alta precisión e interpretabilidad. Los algoritmos genéticos, métodos adaptativos definidos así porque se inspiran en la evolución biológica y tiene base genético-molecular, son uno de los tipos de algoritmos evolutivos más utilizados para generar u optimizar alguno(s) o todos los componentes de los sistemas fuzzy de forma automática a partir de datos o información previa. Dentro de esa área de investigación, definida como sistemas fuzzy genéticos, existen diferentes tendencias de investigación, entre ellas tenemos dos de interés para este trabajo: (i) el uso de algoritmos genéticos multiobjetivo para generar sistemas fuzzy con equilibrio entre precisión e interpretabilidad. Los algoritmos genéticos multiobjetivo son una extensión de los algoritmos genéticos y se caracterizan porque la búsqueda de soluciones que balancean objetivos contradictorios está embebida en el mismo algoritmo; y (ii) utilizar los sistemas fuzzy para resolver el problema de clasificación en bases de datos no balanceadas. Estas bases de datos se caracterizan porque las clases no son representadas por la misma o similar cantidad de ejemplos, y son de particular interés porque muchos problemas de clasificación del mundo se caracterizan por tener una gran cantidad de ejemplos típicos o normales (clase negativa) y un pequeño porcentaje de ejemplos atípicos o relevantes (clase positiva); esa diferencia entre las clases hace que el aprendizaje de modelos de clasificación sea una tarea difícil y sea un tema en abierto en la comunidad de investigación de aprendizaje de máquina. Conforme a las dos tendencias mencionada anteriormente, este trabajo propone un método

de generación genética multiobjetivo de la base de conocimiento fuzzy, de sistemas fuzzy basados en reglas (un tipo de sistemas fuzzy), para la clasificación en bases de datos no balanceadas usando el enfoque iterativo. El enfoque iterativo es utilizado específicamente en la generación de reglas fuzzy, dado que una regla fuzzy es obtenida en cada ejecución del algoritmo genético multiobjetivo. El método propuesto se basa en tres fases: (i) el pre procesamiento, en este caso balanceamiento, de los datos no balanceados utilizando técnica de sobremuestro, submuestreo e híbridos; (ii) la generación genética multiobjetivo iterativa de reglas fuzzy considerando dos objetivos, la precisión o tasa de clasificación y la interpretabilidad o cantidad de condiciones de cada regla fuzzy; (iii) y la optimización genética multiobjetivo de los conjuntos fuzzy considerando dos objetivos, la precisión o tasa de clasificación y un índice de interpretabilidad semántica basada en el desplazamiento y la diferencia de las áreas de los conjuntos fuzzy originales e optimizados. Los resultados obtenidos muestran que el método propuesto tiene un buen desempeño, superando en precisión a otros métodos que realizan la misma tarea encontrados en la literatura. La principal contribución del método propuesto, incluso mayor que la precisión, es la menor cantidad total de condiciones obtenidas en comparación a los otros métodos, y por lo tanto una menor complejidad de los sistemas de clasificación fuzzy obtenidos para bases de datos no balanceadas.

Abstract

The fuzzy systems have been used to solve different types of problems, for example, classification problems, with high accuracy and interpretability. Genetic algorithms, adaptive methods inspired by biological evolution, are a type of evolutionary algorithms used to automatically learn or tune one, some or all components of the fuzzy systems from previous data or information. Within that research area, defined as genetic fuzzy systems, there are different research trends, for example, (i) the use of multi-objective genetic algorithms to learn or tune fuzzy systems with trade-off between accuracy and interpretability. Multi-objective genetic algorithms are an extension of genetic algorithms where the search for solutions that balance conflicting objectives is embedded in the algorithm itself; and (ii) the use of fuzzy systems to solve the classification problem on imbalanced datasets. These datasets are characterized by classification categories not equally represented and are particularly interesting since lots of real world classification problems are often associated with data samples which consist of a large percentage of normal examples (negative classes) and a small percentage of atypical or relevant examples (positive classes). To learn classification models from imbalanced datasets is still an open issue in the machine learning research community. Based on the two research trends mentioned above, this work proposes a method to learn fuzzy knowledge base, of fuzzy rule-based systems (a type of fuzzy systems), using multi-objective genetic algorithms for classification in imbalanced datasets by means of the iterative approach. The iterative approach is used specifically to learn fuzzy rules, since a fuzzy rule is obtained after a execution of the multi-objective genetic algorithm.

The proposed method follows three phases: (i) the pre-processing, in this case, to balance the imbalanced dataset using oversampling, under-sampling and hybrid methods; (ii) iterative fuzzy rule learning using multi-objective genetic algorithms considering two objectives, accuracy or classification rate and interpretability or number of conditions of each rule fuzzy; (iii) tuning of the fuzzy sets using multiobjective genetic algorithms considering two objectives, accuracy or classification rate and a semantic-based interpretability index based on displacements and difference area between original and optimized fuzzy sets. The results show that the proposed method has good performance, the obtained accuracy is better than other state-of-the-art methods. The main contribution of the proposed method, even on the accuracy, is the smaller number of conditions than the other state-of-the-art methods, therefore less complexity of fuzzy classification systems obtained for imbalanced datasets.

Índice general

Índice de figuras	xi
Índice de tablas	xv
1 Introducción	1
1.1 Justificativa y Relevancia	3
1.2 Objetivos	4
1.3 Organización del trabajo	5
2 Bases de Datos No Balanceadas	6
2.1 Métodos a Nivel de Datos: Oversampling	9
2.1.1 Synthetic Minority Oversampling TEchnique	10
2.1.2 Borderline-Synthetic Minority Oversampling TEchnique .	13
2.1.3 ADaptive SYNthetic sampling	14
2.1.4 Safe Level Synthetic Minority Oversampling TEchnique .	17
2.2 Métodos a Nivel de Datos: Undersampling	20
2.2.1 Tomek Link	20
2.2.2 One Sided Selection	21
2.2.3 Neighborhood CLeaning rule	23
2.2.4 underSampling Based on Clustering	24
2.3 Métodos Híbridos	26
2.3.1 Synthetic Minority Oversampling TEchnique + Tomek Links	26
2.3.2 Synthetic Minority Oversampling TEchnique + Edited Nea- rest Neighbor	27

ÍNDICE GENERAL

2.4	Medidas de Evaluación	28
2.4.1	Área Bajo la Curva ROC o Area Under the ROC Curve . .	28
2.5	Consideraciones Finales	30
3	Computación Evolutiva	31
3.1	Algoritmos Evolutivos	32
3.2	Algoritmos Genéticos	35
3.2.1	Cruzamiento y Mutación	37
3.2.2	Selección	42
3.3	Algoritmos Genéticos Multiobjetivo (AGMO)	44
3.3.1	Frontera de Pareto	45
3.3.2	Algoritmo Genético con Ordenamiento No Dominado II o Non-dominated Sorting Genetic Algorithm II (NSGA-II) .	46
3.4	Consideraciones Finales	48
4	Sistemas Fuzzy Genéticos	49
4.1	Teoría de Conjuntos Fuzzy	50
4.2	Funciones de Pertenencia para Conjuntos Fuzzy	52
4.3	Conceptos Relacionados a los Conjuntos Fuzzy	55
4.3.1	Operaciones Básicas entre Conjuntos Fuzzy	57
4.4	Sistemas Fuzzy Basados en Reglas	59
4.4.1	Semántica de las Reglas Fuzzy	62
4.5	Sistemas de Clasificación Fuzzy Basados en Reglas	64
4.5.1	Método de Clasificación Fuzzy Clásico	65
4.5.2	Método de Clasificación Fuzzy General	65
4.6	Sistemas Fuzzy Genéticos	66
4.6.1	Aprendizaje Genético de Reglas Fuzzy	66
4.6.2	Optimización Genética de los Parámetro de la BC	70
4.7	Consideraciones Finales	71
5	Revisión de la Literatura	72
5.1	SFEM para generar SFBR con Equilibrio entre Precisión e Inter- pretabilidad	73
5.2	Optimización en SFBR	75

ÍNDICE GENERAL

5.2.1	Optimización de las Funciones de Pertenencia en la BD .	76
5.2.2	Optimización o Adaptación del Mecanismo de Inferencia .	82
5.3	Aprendizaje de la BC	85
5.3.1	Aprendizaje por Selección de Reglas	85
5.3.2	Aprendizaje de la Base de Reglas	87
5.3.3	Aprendizaje Simultáneo de los Componentes de la Base de Conocimiento	89
5.4	Trabajos que utilizan SFG para tratar el problema de Bases de Da- tos no Balanceadas	94
5.5	Consideraciones Finales	102
6	Método Propuesto: MO-FRBCS-IRL-ID	103
6.1	Pre-procesamiento de la Base de Datos no Balanceada	105
6.2	Aprendizaje Genético Multiobjetivo Iterativo de la Base de Reglas Fuzzy para Clasificación	105
6.2.1	Pre-definir la Base de Datos Fuzzy	105
6.2.2	Representación de las Reglas Fuzzy	106
6.2.3	Población Inicial en la Generación de Reglas Fuzzy	108
6.2.4	Función de Aptitud del Objetivo de Precisión en la Gene- ración de Reglas Fuzzy	108
6.2.5	Función de Aptitud del Objetivo de Interpretabilidad en el Aprendizaje de Reglas Fuzzy	110
6.2.6	Operadores Genéticos en la Generación de Reglas Fuzzy .	111
6.2.7	Condición de Parada y Selección de la Mejor Regla Fuzzy	113
6.2.8	Marcación de Ejemplos en el Aprendizaje de Reglas Fuzzy	114
6.3	Optimización Genética Lateral Multiobjetivo de la Base de Datos Fuzzy	114
6.3.1	Representación de la Base de Datos Fuzzy	114
6.3.2	Generación de la Población Inicial en la Optimización de la Base de Datos Fuzzy	115
6.3.3	Función de Aptitud del Objetivo de Precisión en la Optimi- zación de la Base de Datos Fuzzy	115

ÍNDICE GENERAL

6.3.4	Función de Aptitud del Objetivo de Interpretabilidad en la Optimización de la Base de Datos Fuzzy	116
6.3.5	Operadores Genéticos en la Optimización de la Base de Datos Fuzzy	116
6.3.6	Condición de Parada y Selección de la Base de Datos Fuzzy	118
6.4	Consideraciones Finales	119
7	Experimentos	120
7.1	Plan Experimental	120
7.2	Experimentación	122
7.3	Consideraciones Finales	130
8	Conclusiones	132
8.1	Trabajos Futuros	133
A	Resultados Obtenidos con el Método Propuesto	135
	Referencias	159

Índice de figuras

2.1	a) Pequeño grupo disjunto b) Superposición de ejemplos.	7
2.2	Clasificación de técnicas para resolver el problema de no balancea- miento en bases de datos - Adaptado [Ramychitra and Manikan- dan, 2014]	8
2.3	Replicación de ejemplos de la clase minoritaria - [Longadge et al., 2013].	9
2.4	Generación de un ejemplo sintético utilizando la técnica SMOTE.	12
2.5	Procedimiento de la técnica Borderline-SMOTE - [Batista et al., 2004].	13
2.6	Tipos de ejemplos de la clase minoritaria según el método Borderline- SMOTE - Adaptado [He and Garcia, 2009].	14
2.7	(a) Distribución original de la base de datos. (b) Borde de la clase minoritaria. (c) Ejemplos sintéticos generados - [Han et al., 2005].	15
2.8	Ejemplo de sobremuestro del método ADASYN - [Cao et al., 2013].	15
2.9	Procedimiento del método ADASYN - [He et al., 2008].	16
2.10	Eliminación aleatoria de ejemplos de la clase mayoritaria - [Lon- gadge et al., 2013].	20
2.11	Procedimiento para seleccionar ejemplos Tomek Link - Adapta- do [de Oliveira, 2009].	21
2.12	Ejemplo de la técnica Tomek Link para submuestreo - Adapta- do [de Oliveira, 2009].	22
2.13	Procedimiento de la técnica OSS - Adaptado [Kubat and Matwin, 1997].	22

ÍNDICE DE FIGURAS

2.14	Asignación de un clase a un ejemplo usando la técnica k-NN con $k = 3$	23
2.15	Asignación de un clase a un ejemplo usando la técnica k-NN con $k = 1$	24
2.16	Procedimiento de la técnica NCL para el submuestreo de clase mayoritaria - Adaptado [Laurikkala, 2001].	25
2.17	Procedimiento de la técnica SBC para el submuestreo de clase mayoritaria - Adaptado [Laurikkala, 2001].	26
2.18	Técnica SMOTE+TL: (a) Base de datos no balanceada original; (b) Sobre muestro usando SMOTE; (c) Identificación de ejemplos TL; (d) Eliminación de ejemplos del borde y ejemplos considerados como ruido - Adaptado [Batista et al., 2004].	27
2.19	Ejemplo de la medida de evaluación AUC.	29
3.1	Esquema general de un AE - Adaptado [Eiben and Smith, 2003]. .	33
3.2	Esquema general de un AE en forma de diagrama de flujo - [Eiben and Smith, 2003].	34
3.3	Cruzamiento de un punto - Adaptado [Luke, 2013].	37
3.4	Cruzamiento de dos puntos - Adaptado [Luke, 2013].	38
3.5	Cruzamiento uniforme - Adaptado [Luke, 2013].	39
3.6	Cruzamiento binario simulado $u = 0,2$; $n = 1,0$; $\beta = 0,63$	40
3.7	Mutación polinomial $r = 0,7$; $P_m = 0,5$; $n_m = 1,0$	42
3.8	Vector de acuerdo a la aptitud de cada individuo - Adaptado [Luke, 2013].	43
3.9	Ejemplo de objetivos contradictorios: Tiempo e Inseguridad . . .	45
3.10	Procedimiento de la técnica NSGA-II - [Deb et al., 2002b]. . . .	47
3.11	Distancia de multitud en la técnica NSGA-II - [Deb et al., 2002b].	48
4.1	Ejemplo de conjuntos clásicos.	51
4.2	Ejemplo de conjuntos fuzzy.	52
4.3	Función de pertenencia triangular.	52
4.4	Función de pertenencia trapezoidal.	53
4.5	Función de pertenencia logística.	54
4.6	Función de pertenencia gaussiana.	54

ÍNDICE DE FIGURAS

4.7	Función de pertenencia unitaria.	55
4.8	Soporte y núcleo de un conjunto normal fuzzy F	56
4.9	Conjuntos fuzzy F y G	57
4.10	Intersección entre conjuntos fuzzy.	58
4.11	Unión entre conjuntos fuzzy.	58
4.12	Complemento de un conjunto fuzzy.	59
4.13	Taxonomía de los SFGBR - [Herrera, 2008]	67
4.14	Proceso Genético de Aprendizaje de Reglas fuzzy - [Herrera, 2008]	67
4.15	Enfoque Pittsburgh.	68
4.16	Enfoque Michigan.	69
4.17	Enfoque iterativo o IRL.	69
4.18	Enfoque cooperativo-competitivo.	70
5.1	Taxonomía de dos niveles de los SFEM.- [Fazzolari et al., 2013b].	73
5.2	Curva de equilibrio entre precisión y complejidad en SFBR - [Ishibuchi, 2007]	75
5.3	Intervalos de variación para conjuntos fuzzy triangulares - [Gacto et al., 2010]	79
5.4	Codificación utilizada en el AEMO con N reglas - [Ishibuchi, 2007]	83
5.5	Parte C_{BR} del cromosoma - [Antonelli et al., 2013]	93
5.6	Parte C_{BD} del cromosoma - [Antonelli et al., 2013]	93
6.1	Método propuesto de generación genética multiobjetivo de bases de conocimiento fuzzy para clasificación en bases de datos no balanceadas usando el enfoque iterativo.	104
6.2	Etapas de aprendizaje genético multiobjetivo iterativo de la BR fuzzy.	106
6.3	Ejemplo de base de datos fuzzy uniformemente distribuida.	107
6.4	Representación de las Reglas Fuzzy.	107
6.5	Ejemplo de Cruzamiento SBX en la Generación de Reglas Fuzzy. $u = 0,3 - \eta = 1 - \beta = 0,7746$	112
6.6	Ejemplo de Mutación Polinomial en la Generación de Reglas Fuzzy. $u = 0,8 - \eta = 1$	113
6.7	Representación de una BD fuzzy.	115

ÍNDICE DE FIGURAS

6.8	Ejemplo de Cruzamiento SBX en la Optimización de la BD Fuzzy.	
	$u = 0,7 - \eta = 1 - \beta = 1,2909$	118
6.9	Ejemplo de Mutación Polinomial en la Optimización de la BD	
	Fuzzy. $u = 0,3 - \eta = 1$	118
6.10	Ejemplo de posible BD fuzzy optimizada.	119

Índice de tablas

2.1	Generación de ejemplos sintéticos utilizando la técnica SMOTE - Adaptado [Chawla et al., 2002].	12
3.1	Metáfora básica de la computación evolutiva que relaciona la evolución natural y la solución de problemas - Adaptado [Eiben and Smith, 2003].	32
7.1	Bases de datos no balanceadas utilizadas en los experimentos. . .	121
7.2	Parámetros utilizados por el método MO-FRBCS-IRL-ID en los experimentos.	122
7.3	Valores medios obtenidos por el método propuesto.	123
7.4	Comparación de medias del AUC para los datos de prueba.	125
7.5	Resultados de la prueba de rangos con signo de Wilcoxon para la precisión (medias del AUC). Nivel de significación 0.05. Intervalo de Confianza 95 %.	126
7.6	Comparación de medias de las cantidades de reglas fuzzy para los datos de prueba.	127
7.7	Resultados de la prueba de rangos con signo de Wilcoxon para la interpretabilidad (medias de la cantidad de reglas fuzzy). Nivel de significación 0.05. Intervalo de Confianza 95 %.	128
7.8	Comparación de medias de las cantidades de condiciones para los datos de prueba.	129

ÍNDICE DE TABLAS

7.9	Resultados de la prueba de rangos con signo de Wilcoxon para la interpretabilidad (medias de la cantidad de condiciones). Nivel de significación 0.05. Intervalo de Confianza 95 %.	130
A.1	Valores medios obtenidos por el método propuesto para la base de datos no balanceada abalone19.	137
A.2	Valores medios obtenidos por el método propuesto para la base de datos no balanceada abalone9-18.	138
A.3	Valores medios obtenidos por el método propuesto para la base de datos no balanceada ecoli-0-1-3-7_vs_2-6.	139
A.4	Valores medios obtenidos por el método propuesto para la base de datos no balanceada ecoli4.	140
A.5	Valores medios obtenidos por el método propuesto para la base de datos no balanceada glass-0-1-6_vs_2.	141
A.6	Valores medios obtenidos por el método propuesto para la base de datos no balanceada glass-0-1-6_vs_5.	142
A.7	Valores medios obtenidos por el método propuesto para la base de datos no balanceada glass2.	143
A.8	Valores medios obtenidos por el método propuesto para la base de datos no balanceada glass4.	144
A.9	Valores medios obtenidos por el método propuesto para la base de datos no balanceada glass5.	145
A.10	Valores medios obtenidos por el método propuesto para la base de datos no balanceada page-blocks-1-3_vs_4.	146
A.11	Valores medios obtenidos por el método propuesto para la base de datos no balanceada shuttle-c0-vs-c4.	147
A.12	Valores medios obtenidos por el método propuesto para la base de datos no balanceada shuttle-c2-vs-c4.	148
A.13	Valores medios obtenidos por el método propuesto para la base de datos no balanceada vowel0.	149
A.14	Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-0-5-6-7-9_vs_4.	150

ÍNDICE DE TABLAS

A.15 Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-1-2-8-9_vs_7.	151
A.16 Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-1-4-5-8_vs_7.	152
A.17 Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-1_vs_7.	153
A.18 Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-2_vs_4.	154
A.19 Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-2_vs_8.	155
A.20 Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast4.	156
A.21 Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast5.	157
A.22 Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast6.	158

Introducción

El aprendizaje de máquina es el área de estudio basada en el desarrollo de métodos para la extracción de patrones a partir de datos, con el objetivo de tomar decisiones inteligentes en base a esos patrones. Clasificación es uno de los principales problemas relevantes y actuales que se intenta solucionar mediante el aprendizaje de máquina. Inteligencia computacional es el área de estudio que se enfoca en problemas que solo pueden ser resueltos por humanos y animales, o problemas que requieran inteligencia; los principales temas de interés tratados en el área de inteligencia computacional son las redes neuronales, sistemas fuzzy y computación evolutiva. Según esas definiciones, este trabajo se desarrolla en el ámbito de dos áreas, aprendizaje de máquina e inteligencia computacional, al utilizar sistemas fuzzy y algoritmos evolutivos para resolver problemas de clasificación.

Este trabajo se enfoca en la generación automática de sistemas fuzzy mediante el uso de algoritmos evolutivos, específicamente la generación automática de Sistemas Fuzzy Basados en Reglas (SFBR) mediante el uso de Algoritmos Genéticos Multiobjetivo (AGMO) en bases de datos no balanceadas.

Los SFBR son una extensión de los Sistemas Basados en Reglas (SBR) que utilizan la inferencia lógica para llegar a una solución. Dentro de las principales diferencias entre los SFBR y los SBR, podemos considerar que en los SFBR utilizan

1. INTRODUCCIÓN

conjuntos fuzzy en vez de conjuntos clásicos utilizados por los SBR, teniendo como consecuencia que el mecanismo de inferencia entre ambos sea diferente. Uno de los principales usos de los SFBR es aproximar un sistema real mediante un modelo fuzzy, el cual puede ser usado para explicar, simular o predecir el comportamiento del sistema original, con alta precisión e interpretabilidad.

Los AGMO, extensión de los Algoritmos Genéticos (AG), fueron propuestos por Fonseca y Fleming en 1993 [Fonseca and Fleming, 1993], y se basan en los conceptos de dominación y diversidad de las soluciones. El resultado de un AGMO es un conjunto de soluciones no dominadas en todos los objetivos considerados (intenta encontrar la frontera de Pareto [Pareto, 1896a]). Dentro de ese conjunto de soluciones es posible escoger una solución adecuada para una tarea específica [Nuovo and Catania, 2009]. Existen diferentes técnicas de optimización genética multiobjetivo, por ejemplo, NSGA [Srinivas and Deb, 1994] y NSGA-II [Deb et al., 2002a].

Las bases de datos no balanceadas son comunes y un problema de clasificación en el mundo real, por ejemplo, el reconocimiento de rostros, sensores remotos, previsión de niveles de ozono, diagnóstico médico, entre otros. La característica principal de estas bases de datos no balanceadas es que la cantidad de ejemplos de cada clase es diferente a la cantidad de ejemplos de otra u otras clases, y esa diferencia es considerada como significativa. Podemos considerar una diferencia significativa cuando esa diferencia es mayor a 9:1, es decir, la cantidad de ejemplos de una clase es por lo menos nueve veces mayor que la cantidad de ejemplos de otra clase.

Con el fin de enriquecer la investigación en todos los puntos mencionados, nosotros proponemos y desarrollamos un nuevo método que se enfoca en la utilización del enfoque iterativo [Abadeh and Habibi, 2007] [Gonzalez and Perez, 1999] en la generación de sistemas fuzzy de clasificación en bases de datos no balanceadas basado en el pre procesamiento de datos, la generación de reglas fuzzy, y la optimización de conjuntos fuzzy, enfocados en buscar un equilibrio entre la precisión e interpretabilidad mediante el uso de un AGMO.

1.1 Justificativa y Relevancia

Los SFBR utilizados para clasificación son definidos como Sistemas de Clasificación Fuzzy Basados en Reglas (SCFBR). Cuando los AG, dada su habilidad para manejar grandes espacios de búsqueda y encontrar soluciones cercanas a las óptimas, son utilizados para generar y/o optimizar alguno o algunos de los componentes de los SFBR nos encontramos en la línea de investigación llamada Sistemas Fuzzy Genéticos Basados en Reglas (SFGBR). Los SFGBR se considera un área madura en el estado del arte [Fazzolari et al., 2013b].

La flexibilidad de los SFBR hace que sean aplicables en diferentes tipos de problemas. Dentro de otros, los problemas con múltiples objetivos contradictorios son de particular interés para los investigadores, dado que son problemas comunes y de los cuales una solución óptima en todos los objetivos es buscada. En ese tipo de problemas, mejorar un objetivo conlleva a que uno u otros objetivos empeoren, por lo tanto, generalmente no existe una solución que mejore simultáneamente todos los objetivos.

Este problema puede ser abordado mediante el uso de Algoritmos Evolutivos Multiobjetivo (AEMO) para el diseño del SF. La optimización multiobjetivo lleva a obtener un conjunto de modelos fuzzy con diferentes valores de precisión e interpretabilidad (si son considerados esos dos objetivos) en vez de un solo modelo. Después, uno de los modelo más adecuado es seleccionado, de forma manual o automática, dependiendo de los requerimientos. Ese enfoque híbrido es conocido como Sistemas Fuzzy Evolutivos Multiobjetivo (SFEM).

Dentro de las categorías por el tipo de objetivo en los SFEM, tenemos el diseño de SFBR que buscan un equilibrio entre precisión e interpretabilidad los cuales son considerados objetivos contradictorios [Casillas and Carse, 2009], es decir, entre mayor precisión la interpretabilidad disminuye, e viceversa. La precisión generalmente es determinada por la tasa de clasificación del SFBR, mientras que la interpretabilidad aún no tiene una medida definida y de consenso universal. Hoy en día, los investigadores continúan analizando diferentes medidas de interpretabilidad, pero ellos concuerdan que se necesitan considerar dos medidas, una medida basada en la complejidad (por ejemplo, número de reglas, número de anteceades,

1. INTRODUCCIÓN

etc.), y una medida basada en la semántica (por ejemplo, distinguibilidad de los conjuntos fuzzy).

En la literatura, podemos encontrar diferentes trabajos que encajan en la categoría de los SFEM, algunos de los cuales son mencionados en el Capítulo 5. Dentro de esta categoría se considera como tendencia actual su aplicación en la clasificación de bases de datos no balanceadas, por ello, aún no se han investigado y analizado todos los métodos posibles en dicha categoría. Por ello, este trabajo tiene la justificativa de proponer un nuevo método para la generación automática de SCFBR para bases de datos no balanceadas mediante el modelo iterativo y un AEMO, específicamente un AGMO, para generar modelos fuzzy con diferentes valores de equilibrio entre precisión e interpretabilidad.

La relevancia de este proyecto es dar a conocer el nuevo método propuesto. Dicho método podrá ser usado en diferentes problemas del mundo real, tendrá aporte académico y ayudará a investigaciones futuras en la misma área.

1.2 Objetivos

El objetivo principal de este trabajo es proponer un nuevo método para la generación de clasificadores basados en reglas fuzzy en bases de datos no balanceadas mediante el uso del enfoque iterativo y un AGMO que considere los objetivos contradictorios de precisión e interpretabilidad.

Los objetivos específicos de este trabajo son:

- Contribuir para la creación de métodos robustos para la construcción y optimización de clasificadores fuzzy en bases de datos no balanceadas.
- Analizar el comportamiento de técnicas de pre procesamiento de ejemplos en bases de datos no balanceadas dentro del método propuesto para la generación de clasificadores fuzzy basados en reglas.
- Realizar un análisis experimental del método propuesto y validar la calidad de los clasificadores fuzzy basados en reglas para bases de no balanceadas obtenidos en base a los objetivos de precisión e interpretabilidad.

1.3 Organización del trabajo

Este trabajo está organizado según la siguiente lista:

- El Capítulo 2 trata el problema principal a resolver en este trabajo, es decir, las bases de datos no balanceadas. Dentro de este capítulo serán detallados los principales métodos de pre procesamiento o balanceamiento de datos.
- El Capítulo 3 trata de los principales conceptos de la computación evolutiva y los AEMO, específicamente, los AGMO.
- El Capítulo 4 aborda los principales conceptos de los SF, los SFBR y los Sistemas Fuzzy Genéticos (SFG).
- En el Capítulo 5 nosotros realizaremos una descripción de los principales métodos propuestos encontrados en la literatura para la generación genética de sistemas fuzzy (también usando AEMO) enfocados para resolver el problema de clasificación en bases de datos no balanceadas.
- En el Capítulo 6 explicaremos a detalle el método propuesto para la generación genética multiobjetivo de reglas fuzzy en bases de datos no balanceadas utilizando el enfoque iterativo, que se basa en tres fases: (i) el pre procesamiento o balanceamiento de datos; (ii) la generación genética multiobjetivo de reglas fuzzy basada en el enfoque iterativo; (iii) y la optimización genética multiobjetivo de bases de los conjuntos fuzzy.
- En el Capítulo 7 presentaremos los experimentos y resultados obtenidos del método propuesto, así como su comparación con tres métodos encontrados en la literatura.
- Por último, en el Capítulo 8 presentaremos las conclusiones de nuestro trabajo y los posibles trabajos futuros.

Bases de Datos No Balanceadas

Bases de datos no balanceadas pueden ser encontradas en el mundo real, por ejemplo, reconocimiento de rostros [Liu and Chen, 2007], sensores remotos [Williams et al., 2009], previsión de los niveles de ozono [Tsai et al., 2009], diagnóstico médico [Ganji et al., 2010], entre otros. En este trabajo nos centraremos en tratar el problema de clasificación en bases de datos no balanceadas binaria o de dos clases, que considera la clase positiva o minoritaria y la clase negativa o mayoritaria. La principal característica de este tipo de bases de datos es que la cantidad de ejemplos o muestras de cada clase son diferentes, y dicha diferencia es significativa. Podemos considerar una diferencia significativa cuando esa diferencia es igual o mayor a 9:1, es decir, la cantidad de ejemplos de la clase mayoritaria es por lo menos nueve veces mayor que la cantidad de ejemplos de la clases minoritaria. La razón entre el cantidad de ejemplos de la mayoritaria y la cantidad de ejemplos de la clase minoritaria es definida como razón de desbalanceamiento (o imbalanced ratio (IR)).

La diferencia significativa es un problema en la generación automática de clasificadores, porque generalmente los algoritmos tienen la dificultad de detectar los ejemplos de la clase positiva o minoritaria (clase con menor cantidad de ejemplos)

durante el proceso de entrenamiento [Lopez et al., 2012]. Por lo general, la clase minoritaria es de mayor interés, por ejemplo, en el diagnóstico médico de un tipo de cáncer, si consideramos la presencia de cáncer en un paciente como clase positiva, y la ausencia de cáncer en el paciente (paciente sano) como clase negativa, cuando se diagnostica ausencia de cáncer en un paciente con cáncer, es decir, el paciente es actualmente positivo pero es clasificado como negativo ("falso negativo") ello es un problema más serio, peligroso y caro, porque el paciente podría perder su vida por el retardo en que el paciente tenga un diagnóstico correcto y tratamiento, que diagnosticar cáncer en un paciente sano o sin cáncer ("falso positivo") [Ganganwar, 2012].

Además de la diferencia significativa, los métodos de generación automática de clasificadores en bases de datos no balanceadas se enfrentan a otros problema, dentro de ellos los dos siguientes: los pequeños grupos disjuntos y la superposición de ejemplos. Los pequeños grupos disjuntos, los cuales son grupos que no tienen una gran cantidad de características comunes con el grupo mayor de ejemplos de la misma clase. La superposición de ejemplos es la similitud de características de ejemplos de una clase con ejemplos de la otra. Estas dos características son ilustradas en la Figura 2.1 [Fernandez et al., 2008].

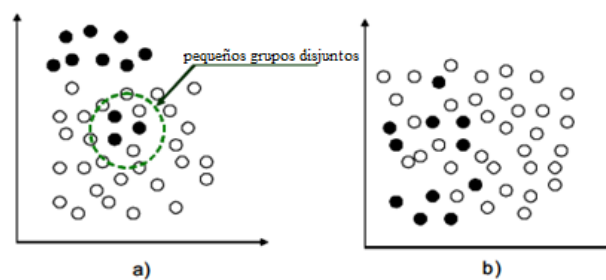


Figura 2.1: a) Pequeño grupo disjunto b) Superposición de ejemplos.

Para resolver los problemas mencionados anteriormente existen diferentes técnicas que pueden ser clasificadas en cinco niveles: Data Level, Algorithmic Level, Cost Sensitive Level, Feature Selection Level y Ensemble Level [Ramyaachitra and Manikandan, 2014]. La Figura 2.2 muestra esta clasificación:

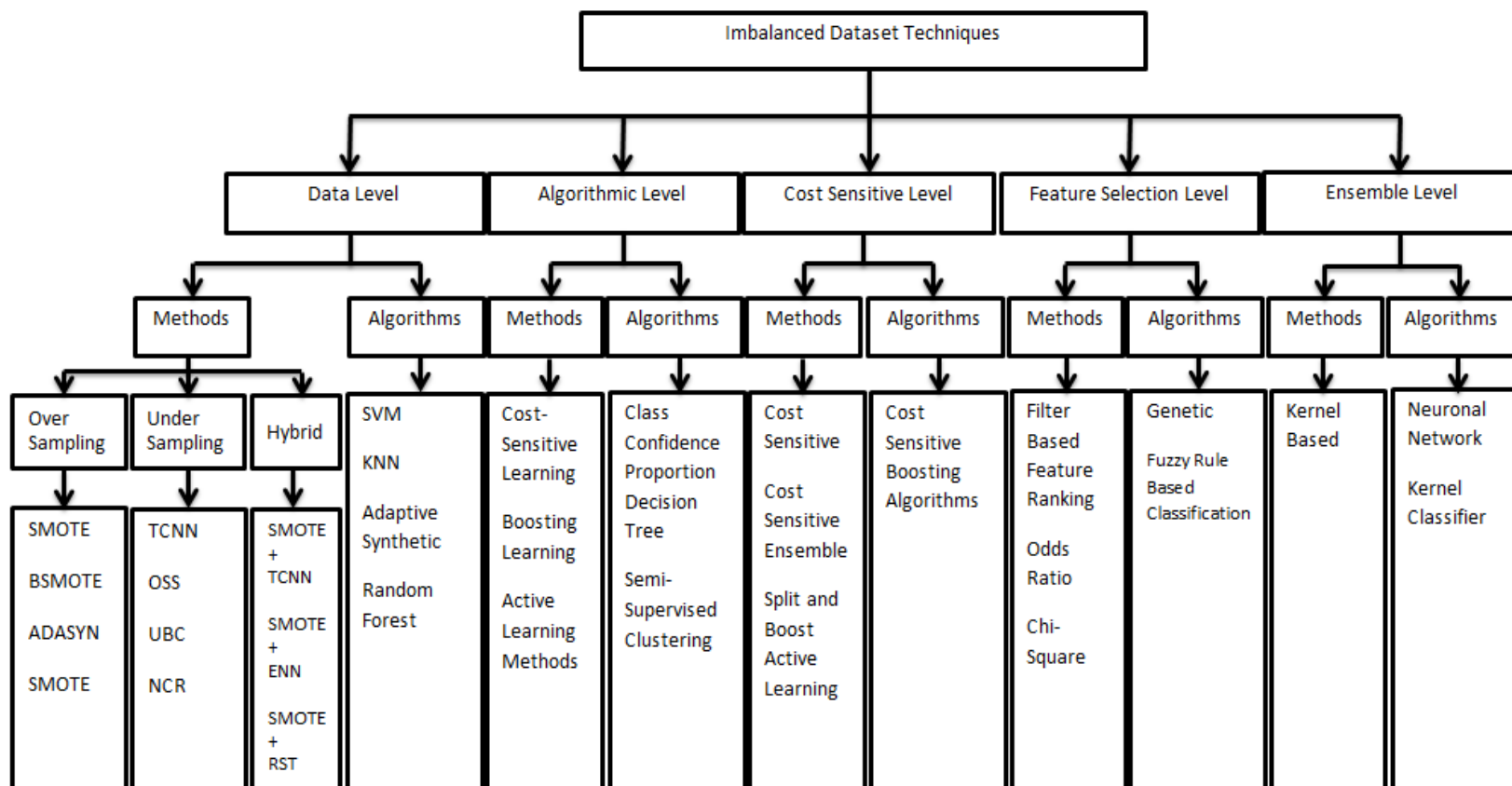


Figura 2.2: Clasificación de técnicas para resolver el problema de no balanceamiento en bases de datos - Adaptado [Ramayachitra and Manikandan, 2014]

2.1 Métodos a Nivel de Datos: Oversampling

Últimos estudios, entre ellos [Ramyaachitra and Manikandan, 2014], [Longadge et al., 2013] y [Ganganwar, 2012], consideran que el pre-procesamiento a nivel de datos provee una mejor solución que otros métodos para el problema de bases de datos no balanceados mencionados anteriormente, pero también sugieren que aplicar dos o tres técnicas, es decir, un híbrido entre dos o tres niveles, podrían dar mejores resultados. Por ello, en este trabajo utilizamos los métodos más conocidos de oversampling, undersampling e híbridos, en el nivel de datos, los cuales son detallados en las próximas secciones y subsección en este Capítulo; y el algoritmo de clasificación basado en reglas fuzzy en el nivel de selección de características, explicado en el Capítulo 6, específicamente en el objetivo de interpretabilidad de la generación de reglas fuzzy.

En caso se desee conocer más sobre las otras categorías, métodos y/o algoritmos, mostradas en la Figura 2.2, un resumen puede ser encontrado en [Ramyaachitra and Manikandan, 2014].

2.1 Métodos a Nivel de Datos: Oversampling

Para ayudar en el proceso de aprendizaje de un clasificador en bases de datos no balanceadas, uno de los métodos más utilizados en el pre-procesamiento de datos es el sobremuestreo o oversampling, el cual consiste en equilibrar la distribución de clases a través de adicionar o aumentar ejemplos de la clase positiva (o clase minoritaria). Una técnica nativa de sobremuestreo es repetir los ejemplos positivos existentes tantas veces como sea necesario para equilibrar las clases, como muestra la Figura 2.3 [Longadge et al., 2013]:

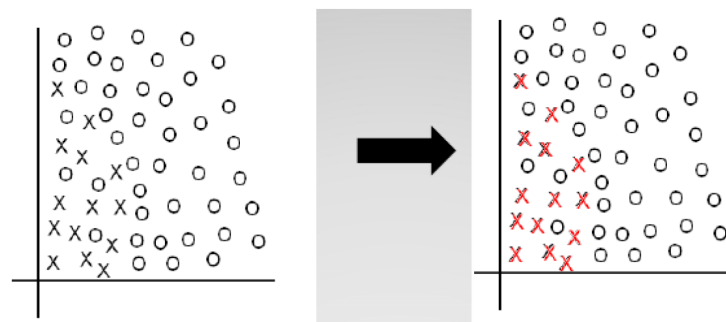


Figura 2.3: Replicación de ejemplos de la clase minoritaria - [Longadge et al., 2013].

2. BASES DE DATOS NO BALANCEADAS

Esta técnica no adiciona información adicional en el proceso de generación o aprendizaje del clasificador por lo cual generalmente no es utilizada. Las soluciones de sobremuestreo más utilizadas a menudo introducen nuevos ejemplos de la clase minoritaria, llamados de ejemplos sintéticos, para generar variación en la información que apoyen al proceso de aprendizaje del clasificador [Cao et al., 2013]. A continuación son listadas y explicadas brevemente algunas de las técnicas más conocidas de sobremuestreo que son utilizadas en método propuesto en este trabajo.

2.1.1 Synthetic Minority Oversampling TEchnique

La técnica SMOTE (Synthetic Minority Oversampling Technique) fue propuesta en [Chawla et al., 2002] y se enfoca en sobremuestrear la clase minoritaria o creación de ejemplos sintéticos, los cuales son creados mediante la realización de ciertas operaciones basadas en los valores de los datos reales.

La clase minoritaria se sobremuestra introduciendo ejemplos sintéticos a lo largo de los segmentos de recta que unen uno o todos los k vecinos más próximos pertenecientes también a la clase minoritaria. Dependiendo de la cantidad de ejemplos sintéticos requeridos, los k vecinos más cercanos son elegidos al azar, por ejemplo, si la cantidad de sobremuestreo necesaria es 200 %, solo dos vecinos de los k vecinos más cercanos son elegidos y un ejemplo se genera en la dirección de los mismos.

Resumidamente, los ejemplos sintéticos se generan de la siguiente manera: seleccionar un ejemplo real, calcular la diferencia entre ese ejemplo y los demás ejemplos, seleccionar los k vecinos reales más cercanos, para cada k vecino cercano encontrar el segmento de recta que los une al ejemplo real, y por último, un punto al azar del segmento encontrado es el nuevo ejemplo sintético. La selección de un punto al azar a lo largo del segmento de recta entre dos características específicas obliga a que la región de decisión de la clase minoritaria sea más general.

2.1 Métodos a Nivel de Datos: Oversampling

El Algoritmo 1 muestra el pseudocódigo del algoritmo de la técnica SMOTE.

Data: Number of minority class samples T ; Amount of SMOTE $N\%$;
Number of nearestneighbors k

Result: $(N/100) * T$ synthetic minority class samples

```
1 if  $N < 100$  then
2   |   then Randomize the  $T$  minority class samples
3   |    $T = (N/100) * T$ 
4   |    $N = 100$ 
5 end
6  $N = (int)(N/100)$ 
7  $k =$  Number of nearest neighbors
8  $numattrs =$  Number of attributes
9  $Sample[][]$  : array for original minority class samples
10  $newindex$  : keeps a count of number of synthetic samples generated,
    initialized to 0
11  $Synthetic[][]$  : array for synthetic samples
12 for  $i \leftarrow 1$  to  $T$  do
13   |   Compute  $k$  nearest neighbors for  $i$ , and save the indices in the  $nnarray$ 
    |    $Populate(N, i, nnarray)$ 
14 end
15  $Populate(N, i, nnarray)$ 
16 while  $N \neq 0$  do
17   |   Choose a random number between 1 and  $k$ , call it  $nn$ . This step chooses
    |   one of the  $k$  nearest neighbors of  $i$ 
18   |   for  $attr \leftarrow 1$  to  $numattrs$  do
19   |   |   Compute:  $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$ 
20   |   |   Compute:  $gap = randomnumberbetween0and1$ 
21   |   |    $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$ 
22   |   end
23   |    $newindex++$ 
24   |    $N = N - 1$ 
25 end
26 return
```

Algoritmo 1: Pseudocódigo del algoritmo de la técnica SMOTE - [Chawla et al., 2002].

2. BASES DE DATOS NO BALANCEADAS

La Tabla 2.1 muestra como generar un ejemplo sintético utilizando SMOTE.

Tabla 2.1: Generación de ejemplos sintéticos utilizando la técnica SMOTE - Adaptado [Chawla et al., 2002].

Considerar (6,4) como un ejemplo real y (4,3) uno de los vecinos más cercanos, también un ejemplo real.

(6,4) es un ejemplo real del cual k vecinos más cercanos son identificados.

(3,4) es un ejemplo real y es uno de los k vecinos más cercanos de (6,4).

Considerar:

$$f1_1=6 \quad f2_1=4 \quad f2_1-f1_1=-2$$

$$f1_2=4 \quad f2_2=3 \quad f2_2-f1_2=-1$$

El nuevo ejemplo, o ejemplo sintético, es generado como:

$$(f1', f2') = (6, 4) + \text{rand}(0-1) * (-2, -1)$$

$\text{rand}(0-1)$ genera un número aleatorio entre 0 y 1

La Figura 2.4 muestra gráficamente la generación de un ejemplo sintético utilizando la técnica SMOTE considerando los valores mostrados en la Tabla 2.1, es decir, $\text{rand}(0 - 1) = 0,5$ y $(f1', f2') = (5, 3,5)$.

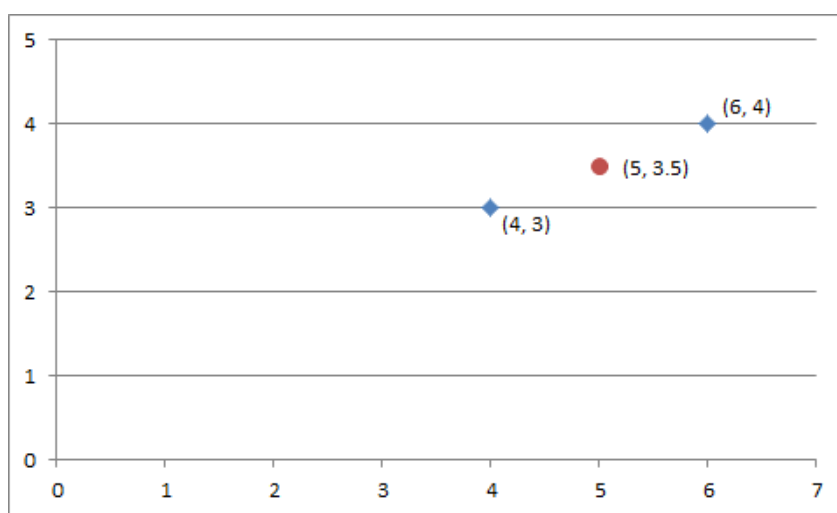


Figura 2.4: Generación de un ejemplo sintético utilizando la técnica SMOTE.

2.1.2 Borderline-Synthetic Minority Oversampling TEchnique

La técnica Borderline-Synthetic Minority Oversampling TEchnique (Borderline-SMOTE) [Han et al., 2005], como su nombre lo indica está basada en la técnica SMOTE, pero a diferencia de SMOTE, esta técnica solo genera ejemplos sintéticos a partir de ejemplos considerados como ejemplos de borde de la clase minoritaria.

Para encontrar ejemplos de borde de la clase minoritaria, primero es definido T como el conjunto de entrenamiento, P como la clase minoritaria y N la clase mayoritaria. Siendo, $P = p_1, p_2, \dots, p_{pnum}$, $N = n_1, n_2, \dots, n_{nnum}$, donde $pnum$ y $nnum$ son el número de ejemplos de la clase minoritaria y el número de ejemplos de la clase mayoritaria, respectivamente. Considerando esos valores, la Figura 2.5 muestra los pasos de la técnica Borderline-SMOTE.

Step 1. For every $p_i (i = 1, 2, \dots, pnum)$ in the minority class P , we calculate its m nearest neighbors from the whole training set T . The number of majority examples among the m nearest neighbors is denoted by $m' (0 \leq m' \leq m)$.

Step 2. If $m' = m$, i.e. all the m nearest neighbors of p_i are majority examples, p_i is considered to be noise and is not operated in the following steps. If $m/2 \leq m' < m$, namely the number of p_i 's majority nearest neighbors is larger than the number of its minority ones, p_i is considered to be easily misclassified and put into a set DANGER. If $0 \leq m' < m/2$, p_i is safe and needs not to participate in the follows steps.

Step 3. The examples in DANGER are the borderline data of the minority class P , and we can see that $DANGER \subseteq P$. We set

$$DANGER = \{p'_1, p'_2, \dots, p'_{dnum}\}, \quad 0 \leq dnum \leq pnum$$

For each example in DANGER, we calculate its k nearest neighbors from P .

Step 4. In this step, we generate $s \times dnum$ synthetic positive examples from the data in DANGER, where s is an integer between 1 and k . For each p'_i , we randomly select s nearest neighbors from its k nearest neighbors in P . Firstly, we calculate the differences, $dif_j (j = 1, 2, \dots, s)$ between p'_i and its s nearest neighbors from P , then multiply dif_j by a random number $r_j (j = 1, 2, \dots, s)$ between 0 and 1, finally, s new synthetic minority examples are generated between p'_i and its nearest neighbors:

$$synthetic_j = p'_i + r_j \times dif_j, \quad j = 1, 2, \dots, s$$

Figura 2.5: Procedimiento de la técnica Borderline-SMOTE - [Batista et al., 2004].

2. BASES DE DATOS NO BALANCEADAS

La Figura 2.6 muestra de forma gráfica como definir los ejemplos DANGER, SAFE y NOISE de la clase minoritaria, según la técnica Borderline-SMOTE considerando un valor k igual a 6. Los ejemplos sintéticos son generados a partir de los ejemplos DANGER.

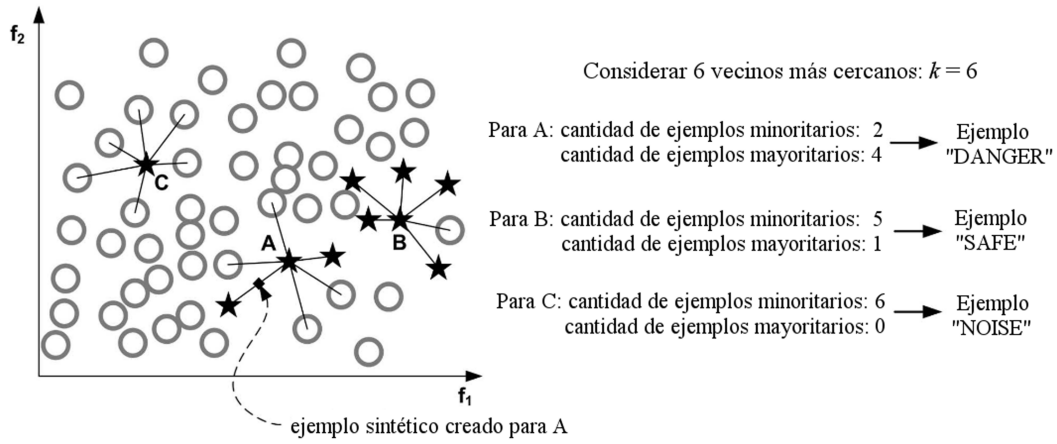


Figura 2.6: Tipos de ejemplos de la clase minoritaria según el método Borderline-SMOTE - Adaptado [He and Garcia, 2009].

Por último, para entender de forma gráfica la técnica Borderline-SMOTE, la Figura 2.7 en la parte (a) muestra la distribución original de la base de datos no balanceada, los puntos circulares representan la clase mayoritaria y los puntos en forma de cruz representan la clase minoritaria. La parte (b) representan los ejemplos de la clase minoritaria que son seleccionados para generar los ejemplos sintéticos, es decir, los ejemplos DANGER. La parte (c) muestra los ejemplos sintéticos generados según el método SMOTE, a partir de los ejemplos DANGER, representados por puntos cuadrados.

2.1.3 ADaptive SYNthetic sampling

La técnica ADaptive SYNthetic sampling (ADASYN) [He et al., 2008], se basa en la idea de generar ejemplos sintéticos, de forma adaptativa, de la clase minoritaria según la distribución de los mismos, es decir, más ejemplos sintéticos se generan para los ejemplos de la clase minoritaria que sean más difíciles de aprender, en

2.1 Métodos a Nivel de Datos: Oversampling

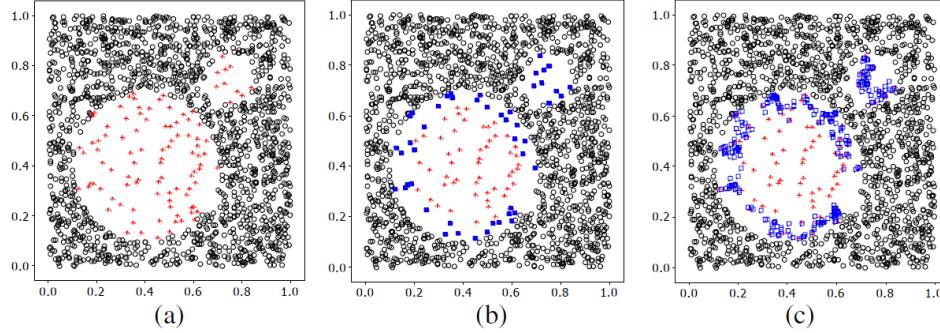


Figura 2.7: (a) Distribución original de la base de datos. (b) Borde de la clase minoritaria. (c) Ejemplos sintéticos generados - [Han et al., 2005].

comparación con los ejemplos de la clase minoritaria que son menos difíciles de aprender.

La técnica ADASYN no solo puede reducir el sesgo de aprendizaje dado por el desequilibrio de la distribución de ejemplos original, sino también puede cambiar de forma adaptativa la frontera de decisión para centrarse en los ejemplos difíciles de aprender. La Figura 2.8 muestra un ejemplo del uso de método ADASYN.

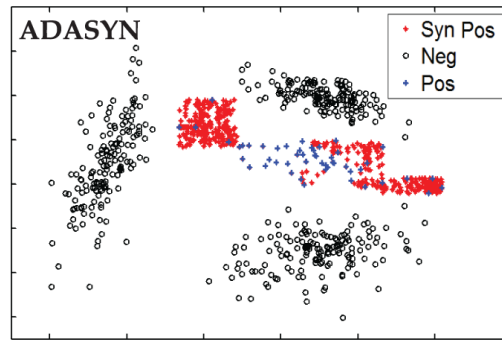


Figura 2.8: Ejemplo de sobremuestro del método ADASYN - [Cao et al., 2013].

La Figura 2.9 muestra el procedimiento del método ADASYN para un problema de clasificación de dos clases. La idea principal del algoritmo de ADASYN es utilizar la distribución de densidad $\hat{r}_i$, como criterio automático para decidir el número de ejemplos sintéticos a ser generados para cada ejemplo de la clase minoritaria, $\hat{r}_i$ indica la relación entre la cantidad de ejemplos de la clase mayoritaria que se encuentran dentro los k vecinos más cercanos de un ejemplo de la clase minoritaria.

2. BASES DE DATOS NO BALANCEADAS

Input

(1) Training data set D_{tr} with m samples $\{x_i, y_i\}$, $i = 1, \dots, m$, where x_i is an instance in the n dimensional feature space X and $y_i \in Y = \{1, -1\}$ is the class identity label associated with x_i . Define m_s and m_l as the number of minority class examples and the number of majority class examples, respectively. Therefore, $m_s \leq m_l$ and $m_s + m_l = m$.

Procedure

(1) Calculate the degree of class imbalance:

$$d = m_s/m_l \quad (1)$$

where $d \in (0, 1]$.

(2) If $d < d_{th}$ then (d_{th} is a preset threshold for the maximum tolerated degree of class imbalance ratio):

(a) Calculate the number of synthetic data examples that need to be generated for the minority class:

$$G = (m_l - m_s) \times \beta \quad (2)$$

Where $\beta \in [0, 1]$ is a parameter used to specify the desired balance level after generation of the synthetic data. $\beta = 1$ means a fully balanced data set is created after the generalization process.

(b) For each example $x_i \in \text{minorityclass}$, find K nearest neighbors based on the Euclidean distance in n dimensional space, and calculate the ratio r_i defined as:

$$r_i = \Delta_i/K, \quad i = 1, \dots, m_s \quad (3)$$

where Δ_i is the number of examples in the K nearest neighbors of x_i that belong to the majority class, therefore $r_i \in [0, 1]$;

(c) Normalize r_i according to $\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i$, so that $\hat{r}_i$ is a density distribution ($\sum_i \hat{r}_i = 1$)

(d) Calculate the number of synthetic data examples that need to be generated for each minority example x_i :

$$g_i = \hat{r}_i \times G \quad (4)$$

where G is the total number of synthetic data examples that need to be generated for the minority class as defined in Equation (2).

(e) For each minority class data example x_i , generate g_i synthetic data examples according to the following steps:

Do the **Loop** from 1 to g_i :

(i) Randomly choose one minority data example, x_{zi} , from the K nearest neighbors for data x_i .

(ii) Generate the synthetic data example:

$$s_i = x_i + (x_{zi} - x_i) \times \lambda \quad (5)$$

where $(x_{zi} - x_i)$ is the difference vector in n dimensional spaces, and λ is a random number: $\lambda \in [0, 1]$.

End **Loop**

Figura 2.9: Procedimiento del método ADASYN - [He et al., 2008].

2.1.4 Safe Level Synthetic Minority Oversampling TEchnique

La técnica Safe-Level-Synthetic Minority Oversampling TEchnique (Safe-Level-SMOTE), propuesta en [Bunkhumpornpat et al., 2009], está basada en SMOTE y define si un ejemplo positivo es seguro antes de generar ejemplos sintéticos a partir del mismo. Cada ejemplo sintético es generado con mayor cercanía a ejemplos positivos seguros y todos son generados en zonas seguras.

El nivel de seguridad o *safe level* (sl) es definido mediante la cantidad de ejemplos positivos dentro de los K vecinos más cercanos. Si el sl de un ejemplo positivo es cercano a 0, puede ser considerado como ruido y si es cercano a K es considerado seguro. La proporción de nivel de seguridad o *safe level ratio*, usado para seleccionar la posición segura de un ejemplo sintético, es definido mediante la división del sl de un ejemplo positivo entre el sl de un vecino más cercano. El Algoritmo 2 muestra el pseudocódigo de la técnica Safe-Level-SMOTE.

Donde:

p	: Es un ejemplo del conjunto original de ejemplos positivos D .
n	: Es un vecino más cercano de p .
s	: Es un ejemplo sintético que se encuentra en el conjunto de ejemplos sintéticos D' .
sl_p	: Es el sl de p .
sl_n	: Es el sl de n .
sl_ratio	: Es el <i>safe level ratio</i> .
$numattrs$	: Es el número de atributos.
dif	: Es la diferencia entre los valores de n y p en el mismo atributo id .
gap	: Es una fracción aleatoria de dif .
$p[i]$, $n[i]$ y $s[i]$	: Son los valores numéricos de los ejemplos en el atributo i^{th} .
p , n y s	: Son vectores.

2. BASES DE DATOS NO BALANCEADAS

Data: a set of all original positive instances D

Result: a set of all synthetic positive instances D'

```
1  $D' = \emptyset$ 
2 for each positive instance  $p$  in  $D$  {
3   compute  $k$  nearest neighbours for  $p$  in  $D$  and randomly select one from
   the  $k$  nearest neighbours, call it  $n$ 
4    $slp$  = the number of positive stances in  $k$  nearest neighbours for  $p$  in  $D$ 
5    $sln$  = the number of positive stances in  $k$  nearest neighbours for  $n$  in  $D$ 
6   if ( $sln \neq 0$ ) { ;  $sl$  is safe level.
7      $sl\_ratio = slp/sln$  ;  $sl\_ratio$  is safe level ratio.
8   {
9   else {
10     $sl\_ratio = \infty$ 
11  }
12  if ( $sl\_ratio = \infty$  AND  $slp = 0$ ) { ; the 1st case
13    does not generate positive synthetic instance
14  }
15  else {
16    for ( $atti = 1$  to  $numattrs$ ) { ;  $numattrs$  is the number of attributes.
17      if ( $sl\_ratio = \infty$  AND  $slp \neq 0$ ) { ; the 2nd case
18         $gap = 0$ 
19      }
20      else if ( $sl\_ratio = 1$ ) { ; the 3rd case
21        random a number between 0 and 1, call it  $gap$ 
22      }
23      else if ( $sl\_ratio > 1$ ) { ; the 4th case
24        random a number between 0 and  $1/sl\_ratio$ , call it  $gap$ 
25      }
26      else if ( $sl\_ratio < 1$ ) { ; the 5th case
27        random a number between  $1 - sl\_ratio$  and 1, call it  $gap$ 
28      }
29       $dif = n[atti] - p[atti]$ 
30       $s[atti] = p[atti] + gap\Delta dif$ 
31    }
32     $D' = D' \cup s$ 
33  }
34 }
35 return  $D'$ 
```

Algoritmo 2: Algoritmo de la técnica Safe-Level-SMOTE - Adaptado [Bunkhumpornpat et al., 2009].

2.1 Métodos a Nivel de Datos: Oversampling

Después de asignar el valor de sl a p y a n , el algoritmo calcula el sl_ratio . A partir de ese valor, cinco acciones pueden ser realizadas, las cuales son mostradas entre las líneas 12 y 28 del Algoritmo 2 que detallaremos brevemente a continuación.

El primer caso es mostrado entre las líneas 12 y 14. El sl_ratio es igual a ∞ y el sl de p es igual al 0. Ello significa que p y n son ruidos, por lo tanto, no se generará ningún ejemplo sintético porque el algoritmo considera como inseguras las regiones con ruido.

El segundo caso es mostrado entre las líneas 17 y 19. El sl_ratio es igual a ∞ y el sl de p no es igual al 0. Ello significa que n es ruido, por lo tanto, un ejemplo sintético será generado lejos del ejemplo n duplicando p , porque el algoritmo intenta evitar el ejemplo n considerado como ruido.

El tercer caso es mostrado entre las líneas 20 y 22. El sl_ratio es igual a 1. Ello significa que el sl de p y n son iguales, por lo tanto, un ejemplo sintético será generado en un punto aleatorio de la línea que une p y n , porque p es tan seguro como n .

El cuarto caso es mostrado entre las líneas 23 y 25. El sl_ratio es mayor que 1. Ello significa que el sl de p es mayor que el sl de n , por lo tanto, un ejemplo sintético será generado cerca a p , porque p es más seguro que n . El ejemplo sintético será generado en el rango de $[0, 1/sl_ratio]$.

El quinto caso es mostrado entre las líneas 26 y 28. El sl_ratio es menor que 1. Ello significa que el sl de p es menor que el sl de n , por lo tanto, un ejemplo sintético será generado cerca a n , porque n es más seguro que p . El ejemplo sintético será generado en el rango de $[1 - sl_ratio, 1]$.

Después de cada iteración de la cláusula *for* (línea 2), si el primer caso no ocurre, un ejemplo sintético s será generado según un rango específico en la línea que une p y n , e insertado a D' . Cuando el algoritmo termina, el resultado es el conjunto de ejemplos sintéticos D' . El algoritmo genera $|D| - t$ ejemplos sintéticos donde $|D|$ es el número de ejemplos positivos en D , y t es el número de ejemplos que satisfacen el primer caso.

2.2 Métodos a Nivel de Datos: Undersampling

Para realizar el proceso de aprendizaje automático en bases de datos no balanceadas, otro de los métodos más utilizados en el pre-procesamiento de datos es el Undersampling o submuestreo, el cual consiste en equilibrar la distribución de clases a través de disminuir o eliminar ejemplos de la clase mayoritaria (o clase negativa). Una técnica nativa es eliminar de forma aleatoria los ejemplos negativos existentes tantas veces como sea necesario para equilibrar las clases, como muestra la Figura 2.10:

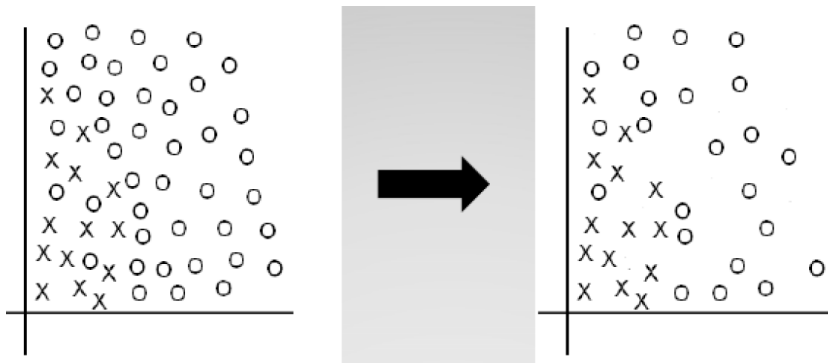


Figura 2.10: Eliminación aleatoria de ejemplos de la clase mayoritaria - [Longadge et al., 2013].

El problema con esta técnica es que se puede perder información importante. Las soluciones de submuestreo más utilizadas eliminan ejemplos negativos que son considerados redundantes o ruido [Ganganwar, 2012]. A continuación son listadas y explicadas brevemente algunas de las técnicas más conocidas de submuestreo que son utilizadas en el método propuesto en este trabajo.

2.2.1 Tomek Link

La técnica Tomek Link fue propuesta por Ivan Tomek en 1976 [Tomek, 1976]. Es una técnica de selección que elimina ejemplos de las regiones de frontera y ejemplos que probablemente son ruido. Esta técnica puede ser definida por los siguientes pasos:

1. Seleccionar dos ejemplos e_i y e_j de diferentes clases.

2. La distancia entre los dos ejemplos es definida como $dist(e_i, e_j)$.
3. El par (e_i, e_j) es llamado de Tomek Link si no existe un ejemplo e_k que sea más cercano a e_i o e_j , es decir, $dist(e_i, e_j) < dist(e_i, e_k)$ y $dist(e_i, e_j) < dist(e_j, e_k)$.

La Figura 2.11 muestra el procedimiento para seleccionar ejemplos Tomek Link.

```
Require:  $L$ : a list  
Require:  $T$ : training data set  
1. for all  $e_i$  in  $T$  do  
2.    $e_j = 1\text{-NN of } e_i, \text{ in } T$   
3.   if 1-NN of  $e_j$  in  $T$  for  $e_i$  then  
4.     if  $Class(e_i) \neq Class(e_j)$  then  
5.       insert  $(e_i, e_j)$  into  $L$   
6.     end if  
7.   end if  
8. end for  
9. return  $L$ , Tomek Links
```

Figura 2.11: Procedimiento para seleccionar ejemplos Tomek Link - Adaptado [de Oliveira, 2009].

Donde:

1 – NN : Es un vecino más cercano o *Nearest-Neighbor* (NN).

Los Tokem Links representan ejemplos de la regiones de indecisión o probable ruido. La técnica Tomek Link puede ser utilizada de dos formas. La primera cuando es utilizado como método de limpieza, en la cual los ejemplos e_i y e_j son eliminados. La segunda cuando es utilizado como método de submuestreo, en la cual los ejemplos de la clase mayoritaria son eliminados, como muestra la Figura 2.12.

2.2.2 One Sided Selection

La técnica One Sided Selection (OSS), propuesta en [Kubat and Matwin, 1997], es una técnica de submuestreo que conserva solo los ejemplos más representativos de la clase mayoritaria.

2. BASES DE DATOS NO BALANCEADAS

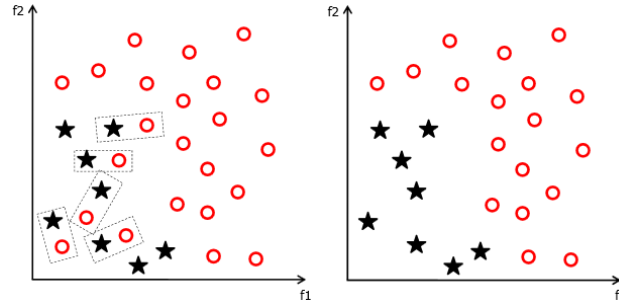


Figura 2.12: Ejemplo de la técnica Tomek Link para submuestreo - Adaptado [de Oliveira, 2009].

Para seleccionar los ejemplos más significativos, el método OSS primero escoge de forma aleatoria un ejemplo x de la clase mayoritaria. Después de ello, usa la técnica k -Nearest Neighbors (k -NN) [Silverman and Jones, 1989], que detallaremos posteriormente, con un valor de $k=1$ para clasificar el resto de ejemplos de la clase positiva. Por último, los ejemplos de la clase mayoritaria que sean correctamente clasificados son removidos porque son considerados redundantes. Así, después del proceso de submuestreo, la clase mayoritaria solo contiene los ejemplos que fueron clasificados de forma incorrecta. Finalmente, la técnica OSS usa Tomek Link para remover los ejemplos de la clase mayoritaria que se encuentran en el borde o considerados como ruido. La Figura 2.13 muestra el procedimiento de la técnica OSS.

- 1 Considerar S como el conjunto de entrenamiento original.
- 2 Inicialmente, C contiene todos los ejemplos positivos en S y un ejemplo negativo es seleccionado aleatoriamente.
- 3 Clasificar S con 1-NN usando los ejemplos en C , y comparar las etiquetas asignadas con las originales. Después, mover todos los ejemplos clasificados de forma incorrecta a C que ahora es consistente con S .
- 4 Remover de C todos los ejemplos negativos usando Tomek Link. Ello remueve los ejemplos negativos que son considerados ejemplos borde y/o ruido. Todos los ejemplos positivos se conservan. El resultado es el conjunto T .

Figura 2.13: Procedimiento de la técnica OSS - Adaptado [Kubat and Matwin, 1997].

Como mencionamos anteriormente, a continuación detallaremos de forma resu-

mida la técnica k-NN. Dicha técnica, usada como clasificador, asigna a un ejemplo la clase más común entre sus k vecinos más cercanos, por ejemplo, en la Figura 2.14 el ejemplo A , usando el algoritmo k-NN con $k = 3$, es asignado a la clase negativa (círculos blancos), porque de los 3 vecinos más cercanos 2 (B y C) son de la clase negativa y solo 1 (D) de la clase positiva (círculos negros).

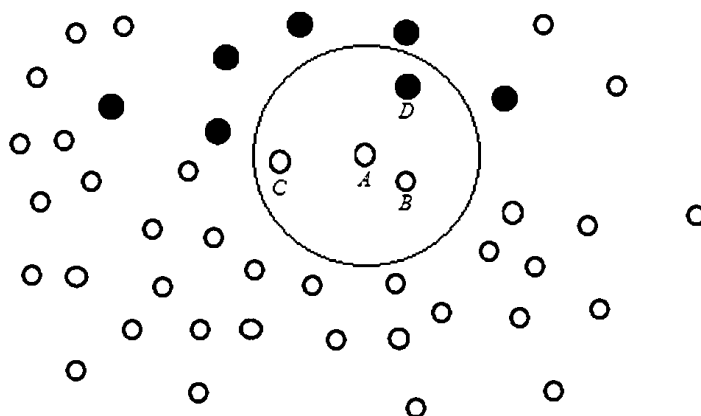


Figura 2.14: Asignación de un clase a un ejemplo usando la técnica k-NN con $k = 3$.

La técnica OSS utiliza la técnica k-NN con $k = 1$, la Figura 2.15 muestra un ejemplo de la asignación de una clase con ese valor de k . En ese caso, si el ejemplo x , seleccionado aleatoriamente, fuera el ejemplo A en la técnica OSS, el ejemplo B sería considerado como redundante y sería eliminado del conjunto de datos de entrenamiento.

2.2.3 Neighborhood CLeaning rule

La técnica Neighborhood CLeaning rule (NCL) fue propuesto en [Laurikkala, 2001] y utiliza el principio de la técnica OSS, pero se enfoca más en la limpieza de los datos que en la reducción de los datos.

El autor de la técnica NCL consideró dos puntos importantes para definir dicha técnica. En primer lugar, la calidad de los resultados de la clasificación no depende necesariamente de la cantidad de ejemplos de la clase, por lo tanto, se debe tener en cuenta, además de la distribución de las clases, otras características de los datos,

2. BASES DE DATOS NO BALANCEADAS

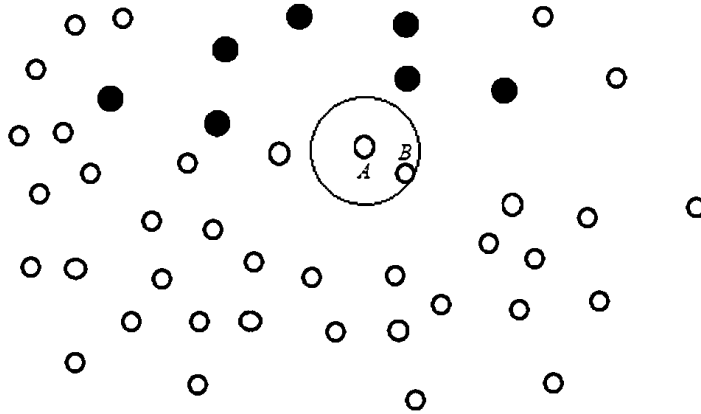


Figura 2.15: Asignación de un clase a un ejemplo usando la técnica k-NN con $k = 1$.

como el ruido, que puede obstaculizar la clasificación. En segundo lugar, los estudios realizados sobre las técnicas de submuestreo, han demostrado que es difícil mantener la precisión en la clasificación original, mientras que los datos se reduce, por ello, mientras se mejora de la identificación de la clase minoritaria, el método debe ser capaz de clasificar las otras clases con una precisión aceptable.

La Figura 2.16 muestra el procedimiento de la técnica NCL para el submuestreo de la clase mayoritaria considerando dos clases.

Donde:

- e_i : Es un ejemplo, de la clase mayoritaria o minoritaria, que pertenece a T .
- $3 - NN$: Son los tres vecinos más cercanos o *Nearest-Neighbor* (NN).
- VNN : Es un vector que contiene los 3 ejemplos e_j más cercanos al ejemplo e_i .

2.2.4 underSampling Based on Clustering

La técnica underSampling Based on Clustering (SBC) fue propuesta en [Laurikala, 2001]. Esta técnica agrupa todos los ejemplos de entrenamiento en algunos grupos, para luego seleccionar un adecuado número de ejemplos negativos a ser

2.2 Métodos a Nivel de Datos: Undersampling

Require: L : a list of examples
Require: T : training data set

1. **for all** e_i in T **do**
2. $VNN = 3\text{-NN}$ of e_i in T
3. **if** $e_i \in \text{majorityclass}$ **then**
4. **if all** e_j in $VNN \in \text{minorityclass}$ **then**
5. insert e_i into L
6. **end if**
7. **end if**
8. **if** $e_i \in \text{minorityclass}$ **then**
9. **if all** e_j in $VNN \in \text{majorityclass}$ **then**
10. insert all e_j in VNN into L
11. **end if**
12. **end if**
13. **end for**
14. $T = T - L$
15. **return** T

Figura 2.16: Procedimiento de la técnica NCL para el submuestreo de clase mayoritaria - Adaptado [Laurikkala, 2001].

eliminados. La selección se basa en la proporción entre la cantidad de ejemplos negativos y la cantidad de ejemplos positivos, es decir, ejemplos negativos que se encuentran en grupos que tengan una proporción mayor entre las cantidades ejemplos negativos con los positivos, tendrán preferencia en ser seleccionados que ejemplos negativos que se encuentran en grupos que tengan una proporción menor entre la cantidades de ejemplos negativos contra los positivos.

La técnica SBC sigue los pasos mostrados en la Figura 2.17, asumiendo que la base de datos no balanceada contiene ejemplos de la clase mayoritaria (MA) y ejemplos de la clase minoritaria (MI).

Donde:

- $SSize_{MA}^i$: El número de ejemplos de la clase mayoritaria en el grupo i a ser seleccionados.
- $m \times Size_{MI}$: Es la cantidad de ejemplos de la clase mayoritaria que se supone tendrá el conjunto de entrenamiento final.
- $Size_{MA}^i$: El número de ejemplos de la clase mayoritaria en el grupo i .
- $Size_{MI}^i$: El número de ejemplos de la clase minoritaria en el grupo i .

2. BASES DE DATOS NO BALANCEADAS

Step1.	Determine the ratio of $Size_{MA}$ to $Size_{MI}$ in the training dataset.
Step2.	Cluster all the samples in the dataset into some clusters.
Step3.	Determine the number of selected majority class samples in each cluster by using expression:
$SSize_{MA}^i = (m \times Size_{MI}) \times \frac{Size_{MA}^i / Size_{MI}^i}{\sum_{i=1}^K Size_{MA}^i / Size_{MI}^i}$	
and then randomly select the majority class samples in each cluster.	
Step4.	Combine the selected majority class samples and all the minority class samples to obtain the training dataset.

Figura 2.17: Procedimiento de la técnica SBC para el submuestreo de clase mayoritaria - Adaptado [Laurikkala, 2001].

2.3 Métodos Híbridos

2.3.1 Synthetic Minority Oversampling TEchnique + Tomek Links

La técnica Synthetic Minority Oversampling TEchnique + Tomek Links (SMOTE+TL) fue propuesta en [Batista et al., 2004]. Los autores sostienen que aunque el sobremuestreo de la clase minoritaria puede balancear la distribución de las clases, otros problemas aún pueden persistir el conjunto de datos, por ejemplo, algunos ejemplos de la clase mayoritaria podrían invadir el espacio de la clase minoritaria; el caso opuesto también puede suceder, cuando se introducen ejemplos artificiales de la clase minoritaria en el espacio del clase mayoritaria. Ello puede llevar a un clasificador al problema de sobre entrenamiento. Para resolver esos problemas, los autores proponen utilizar el método TL como método de limpieza de los datos (eliminar ejemplos de ambas clases) después de utilizar el método SMOTE. La aplicación de esta técnica es mostrada en la Figura 2.18.

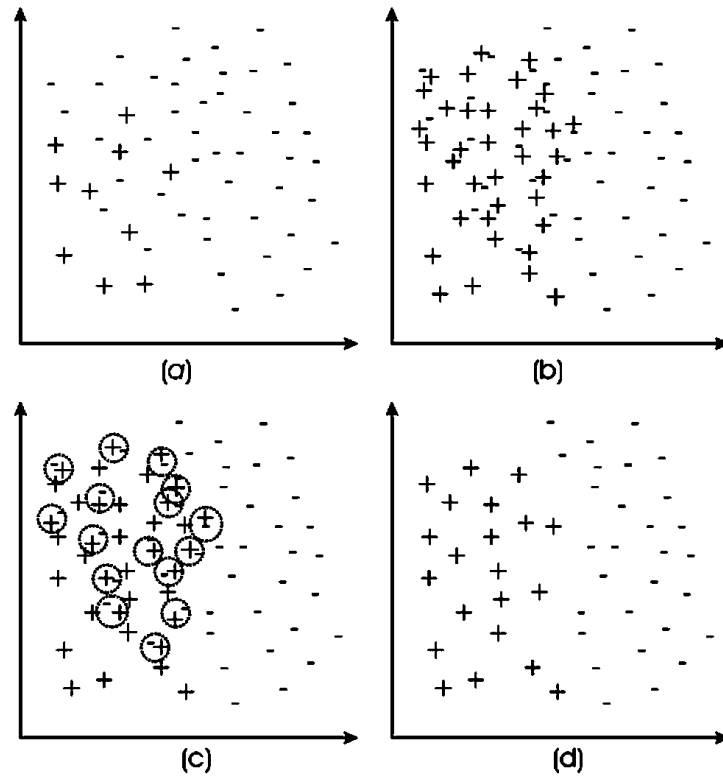


Figura 2.18: Técnica SMOTE+TL: (a) Base de datos no balanceada original; (b) Sobre muestro usando SMOTE; (c) Identificación de ejemplos TL; (d) Eliminación de ejemplos del borde y ejemplos considerados como ruido - Adaptado [Batista et al., 2004].

2.3.2 Synthetic Minority Oversampling TEchnique + Edited Nearest Neighbor

Esta técnica también fue propuesta en [Batista et al., 2004] y es similar a la técnica SMOTE+TL. La diferencia es que en vez de usar TL para limpiar ejemplos, es utilizada la técnica de Wilson llamada Edited Nearest Neighbor Rule (ENN) [Wilson, 1972].

La técnica ENN tiene el mismo principio que la técnica NCL vista en la subsección 2.2.3, pero elimina ejemplos de ambas clases. Así, cualquier ejemplo que no sea clasificado correctamente por sus tres vecinos más cercanos (3-NN) es eliminado del conjunto de datos.

2. BASES DE DATOS NO BALANCEADAS

2.4 Medidas de Evaluación

Las medidas de evaluación tienen un papel importante en la evaluación del desempeño del clasificador. Usualmente, la tasa de clasificación frecuentemente es usada para determinar la precisión de un clasificador, pero en problemas de bases de datos no balanceadas esta tasa puede estar influenciada por la cantidad de ejemplos negativos, como vimos anteriormente, por ello otras medidas son utilizadas.

Nosotros utilizamos una de las medidas más conocidas y utilizadas para comparar clasificadores en bases de datos no balanceadas, basada en la característica operativa del receptor o Receiver Operating Characteristic (ROC), llamada área bajo la curva ROC o Area Under the ROC Curve (AUC), la cual es detallada en la siguiente subsección.

2.4.1 Área Bajo la Curva ROC o Area Under the ROC Curve

En el análisis de los experimentos (Capítulo 7) utilizamos el área Bajo la Curva ROC o Area Under the ROC Curve (AUC) [Huang and Ling, 2005] para definir la precisión del clasificador propuesto y compararlo con otros clasificadores fuzzy encontrados en la literatura.

El AUC es un método estándar que resume el rendimiento de un clasificador basado en los siguientes valores:

1. Falso positivo o False Positive (FP): Un ejemplo negativo es clasificado como positivo.
2. Falso negativo o False Negative (FN): Un ejemplo positivo es clasificado como negativo.
3. Verdadero positivo o True Positive (TP): Un ejemplo positivo es clasificado como positivo.
4. Verdadero negativo o True Negative (TN): Un ejemplo negativo es clasificado como negativo.

El valor del AUC puede ser definido usando la siguiente fórmula:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \quad (2.1)$$

Donde:

TP_{rate} : Es la razón de verdaderos positivos, representa el valor del eje Y en el gráfico del AUC, y puede es calcula de la siguiente forma:

$$TP_{rate} = TP / (TP + FN).$$

FP_{rate} : Es la razón de falsos positivos, representa el valor en el eje X en el gráfico del AUC, y puede es calcula de la siguiente forma:

$$FP_{rate} = FP / (TN + FP).$$

Entre mayor sea el valor del AUC de un clasificador mejor será su precisión. Como ejemplo, vamos a calcular el AUC del un clasificador para problema simple de clientes de un entidad bancaria donde tenemos 1000 clientes que solicitaron un crédito bancario, de los cuales 300 clientes son fiables y 700 clientes no son fiables.

Para calcular el AUC definimos los valores de $TN = 100$, $FN = 100$, $FP = 200$ y $TP = 600$ que deben ser el resultado del clasificador. Después definimos los valores de $TP_{rate} = 600 / (600 + 100) = 0,86$ y $FP_{rate} = 200 / (100 + 200) = 0,67$. Por último, el valor de $AUC = (1 + TP_{rate} - FP_{rate}) / 2 = (1 + 0,86 - 0,67) / 2 = 0,60$. La Figura 2.19 muestra el ejemplo de forma gráfica.

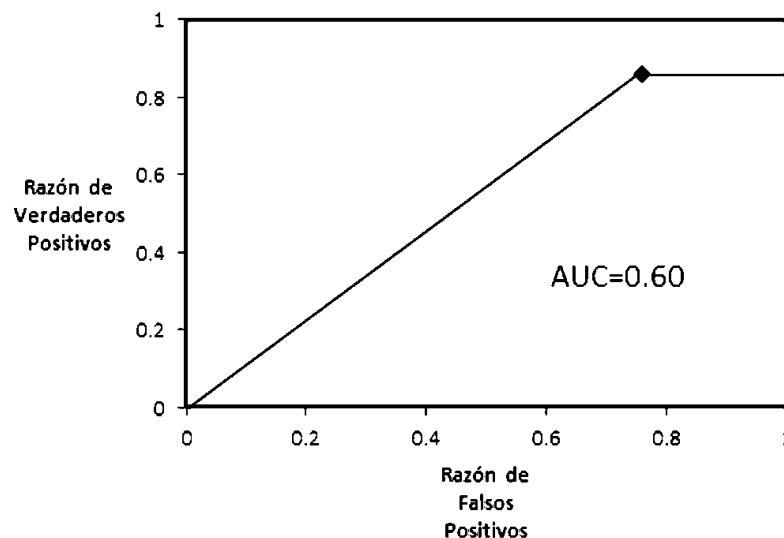


Figura 2.19: Ejemplo de la medida de evaluación AUC.

2.5 Consideraciones Finales

En este capítulo fueron mostrados conceptos básicos del problema de clasificación automática en bases de datos no balanceadas. Fueron detalladas algunas de las técnicas más utilizadas para el balanceamiento de la cantidad de ejemplos en bases de datos no balanceadas, tanto técnicas de sobremuestreo, submuestreo e híbridos. También analizamos la medida de evaluación AUC. En el siguiente capítulo destacamos los principales conceptos de la computación evolutiva y optimización multiobjetivo, temas de interés en este trabajo.

Computación Evolutiva

Computación Evolutiva (CE) es el estudio de sistemas computacionales que utilizan las ideas y están inspirados en la evolución natural y su adaptación [Yao, 1999]. CE es un área de investigación dentro de las ciencias de la computación y su principal metáfora se basa en el poder de la evolución natural para resolver problemas, basada en el estilo estocástico de solución de problemas llamado prueba y error.

Podemos considerar una evolución natural como: un determinado ambiente el cual es ocupado por una población de individuos que se esfuerzan en sobrevivir y reproducirse. La aptitud o fitness es determinada por el ambiente y define cuanto éxito tienen los individuos para lograr sus objetivos, es decir, representa sus posibilidades de supervivencia y reproducción. En el contexto del estilo estocástico prueba y error para resolver problemas, los individuos son una colección de soluciones candidatas; y la aptitud (como las soluciones resuelven el problema) determina la posibilidad que los individuos se mantengan en la población y sean base para construir futuras soluciones candidatas [Eiben and Smith, 2003]. La Tabla 3.1 muestra la metáfora básica de la computación evolutiva que relaciona la evolución natural y la solución de problemas.

Para resolver un problema de forma automática debemos definir un algoritmo, en este trabajo utilizamos un algoritmo evolutivo cuyo concepto es definido en la

3. COMPUTACIÓN EVOLUTIVA

Tabla 3.1: Metáfora básica de la computación evolutiva que relaciona la evolución natural y la solución de problemas - Adaptado [Eiben and Smith, 2003].

Evolución Natural		Solución de Problemas
Ambiente	→	Problema
Individuo	→	Solución candidata
Aptitud	→	Aptitud o Respuesta de la solución candidata

siguiente sección. Después de ello, analizaremos una de las técnicas más conocidas dentro de la CE, los algoritmos genéticos y una extensión de los mismos definida como algoritmos genéticos multiobjetivo.

3.1 Algoritmos Evolutivos

Existen diferentes variantes de los Algoritmos Evolutivos (AEs) [Eiben and Smith, 2003]. La idea fundamental de todas esas técnicas es la misma: dada una población de individuos, el ambiente influye en la selección natural (supervivencia del más apto), la cual causa un aumento de la aptitud de la población. Inicialmente, un conjunto de soluciones candidatas es generada aleatoriamente y se aplica la función de aptitud como una medida, donde el individuo con la aptitud más alta es el mejor. Basada en la aptitud, algunos de los mejores candidatos son escogidos para que sean la base de la siguiente generación aplicando cruzamiento y/o mutación entre ellos. Cruzamiento o recombinación es un operador aplicado a dos o más soluciones candidatas seleccionadas (llamadas de padres) y el resultado es una o más nuevas soluciones candidatas (llamadas de hijos). Mutación es un operador aplicado a una solución candidata y el resultado es una nueva solución candidata. Aplicar cruzamiento y mutación lleva al conjunto de nuevas soluciones candidatas (descendencia) a competir, basada en la aptitud, con las soluciones candidatas antiguas por un lugar en la nueva generación. Ese proceso puede ser repetido hasta que una solución candidata con una aptitud aceptable o suficiente (solución) sea encontrada o un conjunto de límites computacionales sean alcanzados.

En ese proceso tenemos dos tipos de operadores fundamentales que forman la base de un AE:

- Operadores de variación: Cruzamiento y mutación, crean la diversidad necesaria y facilita la creación de nuevas soluciones candidatas.
- Operadores de selección: Actúan como una fuerza para impulsar la aptitud de las soluciones candidatas.

La combinación de los operadores de variación y selección llevan a mejorar los valores de la aptitud en poblaciones consecutivas. Podemos definir este proceso evolutivo como un proceso de optimización, o al menos como un proceso de aproximación, porque los valores de aptitud de las soluciones candidatas se acercan poco a poco a valores óptimos durante el proceso.

Notamos que muchos componentes de un proceso evolutivo son estocásticos, por lo que durante la selección, los individuos más aptos tienen una posibilidad mayor de ser seleccionadas que los individuos menos aptos, pero incluso los individuos con baja aptitud tienen una oportunidad de ser cruzados, mutados e incluso de sobrevivir. Durante el cruzamiento las piezas o partes de los individuos a ser combinadas son escogidas de forma aleatoria. Similarmente para la mutación, las partes que serán mutadas y las nuevas partes, son escogidas de forma aleatoria. El esquema general de un AE es mostrado en la Figura 3.1.

```
INICIALIZAR generando una población con soluciones candidatas de forma
aleatoria;
EVALUAR cada solución candidata;
MIENTRAS (una condición de parada no es satisfecha)
    SELECCIONAR padres;
    CRUZAR pares de padres;
    MUTAR la descendencia resultante;
    EVALUAR las nuevas soluciones candidatas;
    SELECCIONAR los individuos para la siguiente generación;
```

Figura 3.1: Esquema general de un AE - Adaptado [Eiben and Smith, 2003].

La Figura 3.2 muestra el esquema general de un AE en forma de diagrama de flujo.

3. COMPUTACIÓN EVOLUTIVA

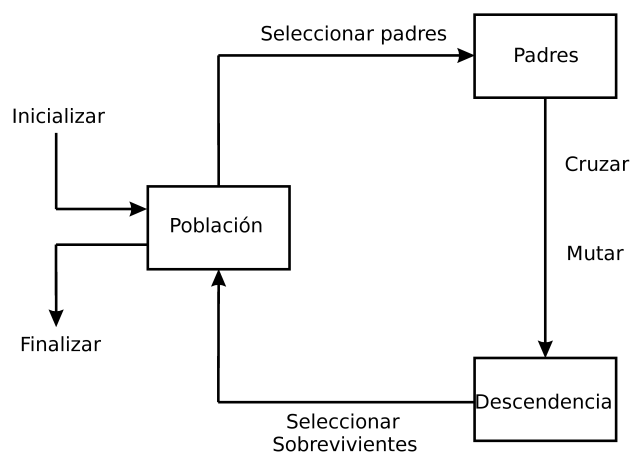


Figura 3.2: Esquema general de un AE en forma de diagrama de flujo - [Eiben and Smith, 2003].

La función de evaluación (aptitud o fitness) representan una estimación heurística de la calidad de la solución, y el proceso de búsqueda es dirigido por los operadores de variación y selección. Los AEs poseen un número de características que pueden ayudar a posicionarlos dentro de la familia de métodos de resolución de problemas prueba y error:

- Los AEs están basados en una población, es decir, poseen un colección de soluciones candidatas.
- Los AEs generalmente usan el cruzamiento para mezclar información de dos o más soluciones candidatas dentro de una nueva solución candidata.
- Los AEs son estocásticos.

Los paradigmas de la CE definen diferentes formas y nombres para cada elemento o parte de un AE, por ejemplo, la representación de un solución candidata. Típicamente, en un Algoritmo Genético (AG), una solución candidata es representada por una cadena de caracteres basada en un alfabeto finito; en las Estrategias Evolutivas (EE) una solución candidata es representada por un vector de valores reales que contiene variables objeto y variables estratégicas; en la Programación

Evolutiva (PE) una solución candidata es representada por una máquina de estados finita. En las siguientes secciones analizamos con más detalle los AGs y los AGMO, por ser de interés para este trabajo.

3.2 Algoritmos Genéticos

Los Algoritmos Genéticos (AGs) fueron propuestos por John Holland en la Universidad de Michigan en la década de 1970s [Holland, 1975]. Al igual que los AEs, los AGs iteran en base a la función de aptitud (u objetivo a minimizar o maximizar), se repite un proceso de selección y reproducción para redefinir la población. Los AGs seleccionan una pequeña cantidad de padres, de acuerdo a su aptitud, para generar una pequeña cantidad de hijos, ese proceso es repetido hasta que todos los hijos sean generados, en otras palabras, en la reproducción, los AGs comienzan con un población de hijos vacía, después dos padres son seleccionados desde la población original de acuerdo a su aptitud, ellos son cruzados (según una probabilidad), y los resultados o hijos son mutados (según una probabilidad), ello forma dos hijos, los cuales son adicionados a la población de hijos, ese proceso es repetido hasta que la población de hijos esté completa [Luke, 2013]. El Algoritmo 3 muestra el pseudocódigo de un AG.

Cada individuo o cromosoma puede ser representado por cualquier cadena de caracteres, los AGs generalmente operan sobre cadenas de tamaño fijo. Cada elemento, gen o caracter de un individuo puede tener diferentes valores, dentro de los más comunes son valores binarios (0 o 1), números enteros o números decimales. El Algoritmo 4 muestra como podemos generar un cadena de caracteres binaria representada por un vector.

En las siguientes subsecciones explicaremos los operadores de variación y de selección en los AGs.

3. COMPUTACIÓN EVOLUTIVA

```
1  $tampop \leftarrow$  tamaño de la población deseada
2  $P \leftarrow \emptyset$ 
3 for  $i \leftarrow 0$  hasta  $tampop$  do
4   |  $P \leftarrow \cup \{\text{nuevo individuo aleatorio}\}$ 
5 end
6  $Mejor \leftarrow \emptyset$ 
7 while Mejor es la solución ideal o se cumple una condición de parada do
8   | foreach  $P_i$  en  $P$  do
9     | EvaluarAptitud( $P_i$ )
10    | if  $Mejor = \emptyset$  o  $Aptitud(P_i) > Aptitud(Mejor)$  then
11      |  $Mejor \leftarrow P_i$ 
12    | end
13    |  $Q \leftarrow \emptyset$ 
14    | for  $i \leftarrow 0$  hasta  $tampop/2$  do
15      | Padre  $P_a \leftarrow$  SeleccionarConSustitución( $P$ )
16      | Padre  $P_b \leftarrow$  SeleccionarConSustitución( $P$ )
17      | Hijo  $C_a, C_b \leftarrow$  Cruzar(Copiar( $P_a$ ), Copiar( $P_b$ ))
18      |  $Q \leftarrow Q \cup \{\text{Mutar}(C_a), \text{Mutar}(C_b)\}$ 
19    | end
20  | end
21  |  $Q \leftarrow P$ 
22 end
23 return  $Mejor$ 
```

Algoritmo 3: Pseudocódigo de un AG - Adaptado [Luke, 2013].

```
1  $\vec{v} \leftarrow$  un nuevo vector  $\{v_1, v_2, \dots, v_l\}$ 
2 for  $i \leftarrow 0$  hasta  $l$  do
3   | if  $0,5 > a$  un número aleatorio seleccionado entre 0.0 y 1.0 then
4     |  $v_i \leftarrow$  verdadero
5   | else
6     |  $v_i \leftarrow$  falso
7   | end
8 end
9 return  $\vec{v}$ 
```

Algoritmo 4: Generación de una cadena de caracteres con valores binarios - Adaptado [Luke, 2013].

3.2.1 Cruzamiento y Mutación

Comenzando por el cruzamiento, el cual consiste en la mezcla o combinación de partes de dos padres (caracteres, cromosomas o vectores) para formar hijos. Como son mezclados los genes o partes dependen de la representación de los individuos. Existen diferentes formas de realizar el cruzamiento, a continuación detallamos resumidamente cuatro de las más utilizadas: cruzamiento de un punto, cruzamiento de dos puntos, cruzamiento uniforme y cruzamiento binario simulado, siendo el último aplicado en el método propuesto en este trabajo.

El cruzamiento de un punto, considerando un vector de tamaño l , define un número c entre 0 y l , e intercambia todos los índices menores a c , como muestra la Figura 3.3.

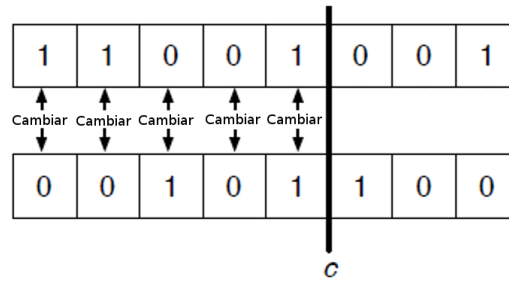


Figura 3.3: Cruzamiento de un punto - Adaptado [Luke, 2013].

El Algoritmo 5 muestra el pseudocódigo del cruzamiento de un punto.

```

1  $\vec{v} \leftarrow$  primer vector  $\{ v_1, v_2, \dots, v_l \}$  a ser cruzado
2  $\vec{w} \leftarrow$  segundo vector  $\{ w_1, w_2, \dots, w_l \}$  a ser cruzado
3  $c \leftarrow$  escoger un número entero entre 0 y  $l$  aleatoriamente
4 if  $c \neq 0$  then
5   for  $i \leftarrow 1$  hasta  $c$  do
6     | Cambiar los valores de  $v_i$  y  $w_i$ 
7   end
8 end
9 return  $\vec{v}$  y  $\vec{w}$ 

```

Algoritmo 5: Pseudocódigo del cruzamiento de un punto - Adaptado [Luke, 2013].

Si $c = 0$ el cruzamiento no se realiza (ello debe tener una probabilidad). Una de las formas de reducir esa posibilidad a 0 es definir el valor de c entre 1 y l .

3. COMPUTACIÓN EVOLUTIVA

El cruzamiento de dos puntos consiste en definir dos puntos c y d , e intercambiar los índices entre ellos. La Figura 3.4 muestra la idea general del cruzamiento de dos puntos.

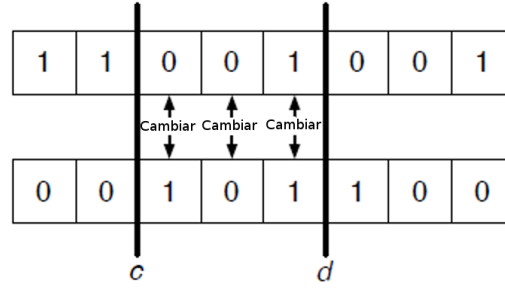


Figura 3.4: Cruzamiento de dos puntos - Adaptado [Luke, 2013].

El Algoritmo 6 muestra el pseudocódigo del cruzamiento de dos punto.

```

1  $\vec{v} \leftarrow$  primer vector  $\{ v_1, v_2, \dots, v_l \}$  a ser cruzado
2  $\vec{w} \leftarrow$  segundo vector  $\{ w_1, w_2, \dots, w_l \}$  a ser cruzado
3  $c \leftarrow$  escoger un número entero entre 1 y  $l$  aleatoriamente
4  $d \leftarrow$  escoger un número entero entre 1 y  $l$  aleatoriamente
5 if  $c > d$  then
6   | Cambiar  $c$  y  $d$ 
7 end
8 if  $c \neq d$  then
9   | for  $i \leftarrow c$  hasta  $d$  do
10    | Cambiar los valores de  $v_i$  y  $w_i$ 
11    end
12 end
13 return  $\vec{v}$  y  $\vec{w}$ 

```

Algoritmo 6: Pseudocódigo del cruzamiento de dos puntos - Adaptado [Luke, 2013].

Como en el caso del cruzamiento de un punto, cuando $c = d$ el cruzamiento no se realiza (ello debe tener una probabilidad). Una de las formas de reducir esa posibilidad a 0 es forzar que d sea diferente a c .

El cruzamiento uniforme consiste en recorrer los cromosomas y cambiar los genes de acuerdo a una probabilidad p . La Figura 3.5 muestra un ejemplo de cruzamiento uniforme.

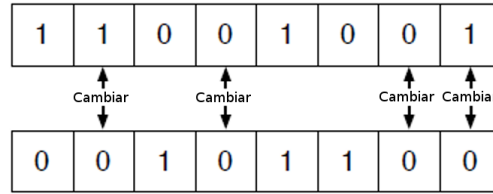


Figura 3.5: Cruzamiento uniforme - Adaptado [Luke, 2013].

El Algoritmo 7 muestra el pseudocódigo del cruce uniforme.

```

1  $p \leftarrow$  probabilidad de cambio por índice
2  $\vec{v} \leftarrow$  primer vector  $\{ v_1, v_2, \dots, v_l \}$  a ser cruzado
3  $\vec{w} \leftarrow$  segundo vector  $\{ w_1, w_2, \dots, w_l \}$  a ser cruzado
4 for  $i \leftarrow 1$  hasta  $l$  do
5   if  $p \geq$  un número aleatorio entre 0.0 y 1.0 then
6     Cambiar los valores de  $v_i$  y  $w_i$ 
7   end
8 end
9 return  $\vec{v}$  y  $\vec{w}$ 

```

Algoritmo 7: Pseudocódigo del cruce uniforme - Adaptado [Luke, 2013].

Por último, el cruce binario simulado o Simulated Binary Crossover (SBX) fue propuesto en [Deb et al., 1995] con el objetivo de simular el comportamiento del cruce de un punto para cromosomas o vectores binarios, visto anteriormente, pero para cromosomas o vectores que contienen números reales. A continuación mostramos las características del SBX:

- El valor de cada gen hijo tiene la misma distancia media al valor de los genes padre.
- Cada punto del cromosoma tiene la misma probabilidad de ser seleccionado como un punto de cruce.
- Entre menor sea el factor de propagación (β) menor será la diferencia entre los genes hijos con los genes padres.
- Es más probable que los cromosomas hijos sean parecidos a los cromosomas padre.

El cruce SBX sigue los siguientes pasos:

3. COMPUTACIÓN EVOLUTIVA

1. Seleccionar dos cromosomas padres x_1 y x_2 .
2. Generar un número aleatorio $u \in [0, 1]$.
3. Calcular β , donde n es el índice de distribución, mediante las siguiente ecuación:

$$\beta = \begin{cases} (2u)^{\frac{1}{n+1}}, & \text{if } u \leq 0,5 \\ \left(\frac{1}{2(1-u)}\right)^{\frac{1}{n+1}}, & \text{if } u > 0,5 \end{cases} \quad (3.1)$$

4. Calcular los cromosomas hijos mediante las siguientes ecuaciones:

$$x_1^{hijo} = 0,5 [(1 + \beta)x_1 + (1 - \beta)x_2] \quad (3.2)$$

$$x_2^{hijo} = 0,5 [(1 - \beta)x_1 + (1 + \beta)x_2] \quad (3.3)$$

La Figura 3.6 muestra un ejemplo de un cruzamiento binario simulado.

x_1	0.76	0.48	0.09	0.86	0.52	0.07	0.34
x_2	0.38	0.60	0.86	0.83	0.44	0.81	0.89
x_1^{hijo}	0.69	0.50	0.23	0.85	0.51	0.21	0.44
x_2^{hijo}	0.45	0.58	0.72	0.83	0.46	0.68	0.79

Figura 3.6: Cruzamiento binario simulado $u = 0,2$; $n = 1,0$; $\beta = 0,63$.

La mutación es un operador que consiste en cambiar alguno o algunos de los genes o elementos de una cadena de caracteres, cromosoma o vector que representa un individuo. Existen diferentes formas de mutación, a continuación detallaremos

dos de las más conocidas: la mutación simple y la mutación polinomial, siendo la última utilizada en el método propuesto.

La mutación simple consiste en modificar los valores de los elementos del vector que representa el individuo de acuerdo a una probabilidad. El Algoritmo 8 muestra el pseudocódigo de la mutación simple en un vector binario.

```

1  $p \leftarrow$  probabilidad de mutación de un índice
2  $\vec{v} \leftarrow$  vector binario  $\{v_1, v_2, \dots, v_l\}$  a ser mutado
3 for  $i \leftarrow 1$  hasta  $l$  do
4   if  $p \geq$  un número aleatorio entre 0.0 y 1.0 then
5      $v_i \leftarrow \neg(v_i)$ 
6   end
7 end
8 return  $\vec{v}$  y  $\vec{w}$ 

```

Algoritmo 8: Pseudocódigo de la mutación simple en un vector binario - Adaptado [Luke, 2013].

La mutación polinomial fue propuesta también en [Deb et al., 1995] y mejorada en [Deb and Tiwari, 2008]. Está basada en una función de distribución de probabilidad la cual realiza una perturbación de los genes o elementos del vector para obtener un cromosoma o vector mutado. El Algoritmo 9 muestra el pseudocódigo de la mutación polinomial.

```

1  $i \leftarrow 0$ 
2 repeat
3    $r \leftarrow U[0, 1]$ 
4   if  $r \leq P_m$  then
5      $\delta_1 \leftarrow \frac{X_i - X_i^{Lower}}{X_i^{Upper} - X_i^{Lower}}$ 
6      $\delta_2 \leftarrow \frac{X_i^{Upper} - X_i}{X_i^{Upper} - X_i^{Lower}}$ 
7      $r \leftarrow U[0, 1]$ 
8     if  $r \leq 0,5$  then
9        $\delta_q \leftarrow [(2r) + (1 - 2r) * (1 - \delta_1)^{n_m+1}]^{\frac{1}{n_m+1}} - 1$ 
10    else
11       $\delta_q \leftarrow 1 - [2(1 - r) + 2.(r - 0,5) * (1 - \delta_2)^{n_m+1}]^{\frac{1}{n_m+1}}$ 
12    end
13     $X_i \leftarrow X_i + \delta_q.(X_i^{Upper} - X_i^{Lower})$ 
14  end
15 until  $i++ = n$ 

```

Algoritmo 9: Pseudocódigo de la mutación polinomial - [Hamdan, 2012].

3. COMPUTACIÓN EVOLUTIVA

Donde:

- i : Índice del gen.
 r : Número aleatorio entre 0 y 1.
 n : Tamaño del cromosoma.
 P_m : Probabilidad de mutación de cada gen.
 X_i : Valor del gen en la posición i .
 X_i^{Lower} : Valor mínimo que puede contener X_i .
 X_i^{Upper} : Valor máximo que puede contener X_i .
 n_m : Índice de distribución.
 δ_1, δ_2 y δ_q : Valores intermedios.

La Figura 3.7 muestra un ejemplo de un cruzamiento binario simulado con límite inferior y superior para todos los genes igual a 0 y 1 respectivamente.

X	0.45	0.58	0.72	0.83	0.46	0.68
X ^{mutado}	0.55	0.58	0.72	0.19	0.46	0.35

Figura 3.7: Mutación polinomial $r = 0,7$; $P_m = 0,5$; $n_m = 1,0$.

3.2.2 Selección

Uno de los métodos de selección más utilizados en los AGs es la selección por ruleta, en el cual la selección se basa en la proporción de aptitud de cada solución o individuo, es decir, si un individuo tiene una aptitud alta, será seleccionado con mayor frecuencia. Una de las formas de realizar ello es definir un vector de acuerdo a la aptitud de cada individuo como muestra la Figura 3.8, donde s es la suma de todas las aptitudes.

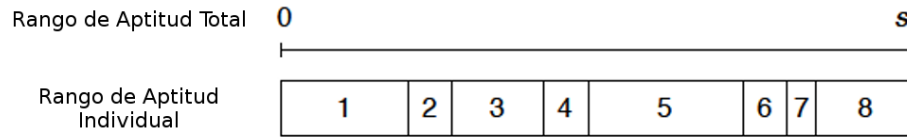


Figura 3.8: Vector de acuerdo a la aptitud de cada individuo - Adaptado [Luke, 2013].

Seguidamente, se define un número entre 0 y s que se encuentra dentro del rango de un individuo, el cual es seleccionado.

El Algoritmo 10 muestra el pseudocódigo del método de selección por ruleta.

```

1 # Para cada generación
2  $\vec{p} \leftarrow$  población copiada dentro de un vector de individuos  $(p_1, p_2, \dots, p_l)$ 
3  $\vec{f} \leftarrow$  aptitud de los individuos en  $\vec{p}$  y con el mismo orden que  $\vec{p}$ 
4 if todos en  $\vec{f}$  tienen el valor de 0.0 then
5   | Asignar a todos los valores en  $\vec{f}$  el valor de 1.0
6 end
7 for  $i \leftarrow 2$  hasta  $l$  do
8   |  $\# f_i = s$ 
9   |  $f_i \leftarrow f_i + f_{i-1}$ 
10 end
11 # Para cada selección
12  $n \leftarrow$  número aleatorio entre 0 y  $f_l$ 
13 for  $i \leftarrow 2$  hasta  $l$  do
14   | if  $f_{i-1} < n \leq f_i$  then
15     | return  $p_i$ 
16   | end
17 end
18 return  $p_1$ 

```

Algoritmo 10: Pseudocódigo del método de selección por ruleta - Adaptado [Luke, 2013].

Otro de los métodos de selección más utilizados en los AGs es la selección por torneo, el cual retorna el individuo con la aptitud más alta de un conjunto de t individuos seleccionados de forma aleatoria de la población. Dentro de principales ventajas de este método, es que no necesita un pre-procesamiento y el tamaño de t es fácil de modificar. El Algoritmo 11 muestra el pseudocódigo del método de selección por torneo. Este método de selección es utilizado en el método propuesto

3. COMPUTACIÓN EVOLUTIVA

en este trabajo.

```
1  $P \leftarrow$  población
2  $t \leftarrow$  tamaño del torneo,  $t \geq 1$ 
3  $Mejor \leftarrow$  individuo seleccionado de forma aleatorio de  $P$ 
4 for  $i \leftarrow 2$  hasta  $t$  do
5   |  $Siguiente \leftarrow$  individuo seleccionado de forma aleatorio de  $P$ 
6   | if  $Aptitud(Siguiente) > Aptitud(Mejor)$  then
7   |   |  $Mejor \leftarrow Siguiente$ 
8   | end
9 end
10 return  $Mejor$ 
```

Algoritmo 11: Pseudocódigo del método de selección por torneo - Adaptado [Luke, 2013].

3.3 Algoritmos Genéticos Multiobjetivo (AGMO)

En el mundo real, muchos problemas de optimización envuelven más de un objetivo a minimizar o maximizar. Esos objetivos en muchos casos son contradictorios, es decir, si consideramos dos objetivos, mientras uno aumenta el otro disminuye, y viceversa. En ese caso, una solución no podría satisfacer ambos objetivos, y la solución óptima para un objetivo, no necesariamente sería la solución óptima para el otro objetivo. Un ejemplo de objetivos contradictorios es mostrado en la Figura 3.9, donde se desea minimizar el tiempo de viaje en un carro (en términos de la velocidad) y minimizar la inseguridad de los pasajeros dentro del carro, esos objetivos son contradictorios, porque cuando reducimos el tiempo (aumentamos la velocidad) la inseguridad aumenta (la seguridad baja), y viceversa.

Otro ejemplo se puede encontrar en la fabricación de un producto, donde se quiere minimizar el costo y maximizar la calidad. La solución para ese tipo de problemas, generalmente es un conjunto de soluciones que tengan el mejor equilibrio entre los objetivos. Ese conjunto de soluciones es denominada frontera de Pareto, la cual es detallada en la siguiente subsección.

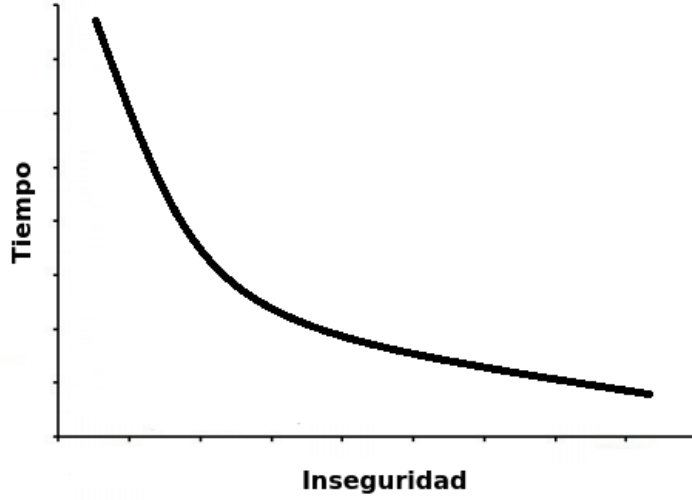


Figura 3.9: Ejemplo de objetivos contradictorios: Tiempo e Inseguridad

3.3.1 Frontera de Pareto

La frontera de Pareto fue definida por Vilfredo Pareto en el siglo XIX [Pareto, 1896b] y significa el origen del estudio en la optimización multiobjetivo. Dicha frontera está basada en que soluciones dominan otras soluciones. Si consideramos n funciones objetivos $f_1, f_2, \dots, f_n$, una solución x domina a una solución x' , en el caso de minimización, si las siguientes ecuaciones son satisfechas:

$$f_i(x) \leq f_i(x') \quad \forall i = 1, 2, \dots, n \quad (3.4)$$

$$f_i(x) < f_i(x') \quad \exists i = 1, 2, \dots, n \quad (3.5)$$

En otras palabras, una solución x domina a una solución x' , en el caso de minimización, si las funciones objetivo de la solución x son menores o iguales (por lo menos menor en una función objetivo) al valor de las funciones objetivo de la solución x' .

En un conjunto de soluciones X , una solución x pertenece a la frontera de Pareto si no es dominada por ninguna otra solución en X , es decir, no existe una solución x' que pertenece a X que sea mejor que x en todas las funciones objetivo. Las soluciones en la frontera de Pareto también se denominan soluciones

3. COMPUTACIÓN EVOLUTIVA

superiores, eficientes o no dominadas, y forman el conjunto de soluciones para un problema multiobjetivo.

Para encontrar las soluciones en la frontera de Pareto en un problema multiobjetivo se pueden utilizar varios métodos y técnicas, dentro de ellos los Algoritmos Evolutivos Multiobjetivo (AEMO) [Coello Coello et al., 2007]. Existen varios AEs que pueden ser utilizados para esa tarea, como los AGs, ellos son definidos como Algoritmos Genéticos Multiobjetivo (AGMO), que son una extensión de los AGs que por lo general utilizan una sola función objetivo. Dentro de las técnicas más utilizadas y bien sucedidas en los AGMO, tenemos el Algoritmo Genético con Ordenamiento no Dominado II o Nondominated Sorting Genetic Algorithm II (NSGA-II). Esta técnica es detallada en la siguiente subsección por ser de interés en este trabajo.

3.3.2 Algoritmo Genético con Ordenamiento No Dominado II o Non-dominated Sorting Genetic Algorithm II (NSGA-II)

La técnica de optimización genética multiobjetivo conocida como NSGA-II fue propuesta en [Deb et al., 2002b] y está basada en el ordenamiento elitista por la no dominancia de las soluciones y la distancias entre las mismas. El procedimiento de la técnica NSGA-II se muestra en la Figura 3.10.

Este procedimiento es detallado a continuación. Primero se define una población R_t (de tamaño de $2N$, siendo N el tamaño de la población en la técnica NSGA-II) combinando la población padre P_t y la población descendiente Q_t (cada una de tamaño N). La población Q_t es generada en base a la población P_t utilizando los operadores de selección, cruzamiento y mutación de un AG. Después, la población R_t se ordena en base a las fronteras generadas por la dominancia de las soluciones. La primera frontera F_1 agrupa las soluciones no dominadas o mejores soluciones, la segunda frontera F_2 agrupa las soluciones dominadas por soluciones de la frontera F_1 , y así sucesivamente. Seguidamente, si el tamaño de F_1 es menor que N , todos los miembros de F_1 son inseridos en la nueva población P_{t+1} y la población restante de P_{t+1} se selecciona de las siguientes fronteras según el orden de cada una. Sin embargo, como se muestra en la Figura 3.10, no todas las soluciones de F_3 se pueden insertar en la población de P_{t+1} , es decir, el número de vacantes o

3.3 Algoritmos Genéticos Multiobjetivo (AGMO)

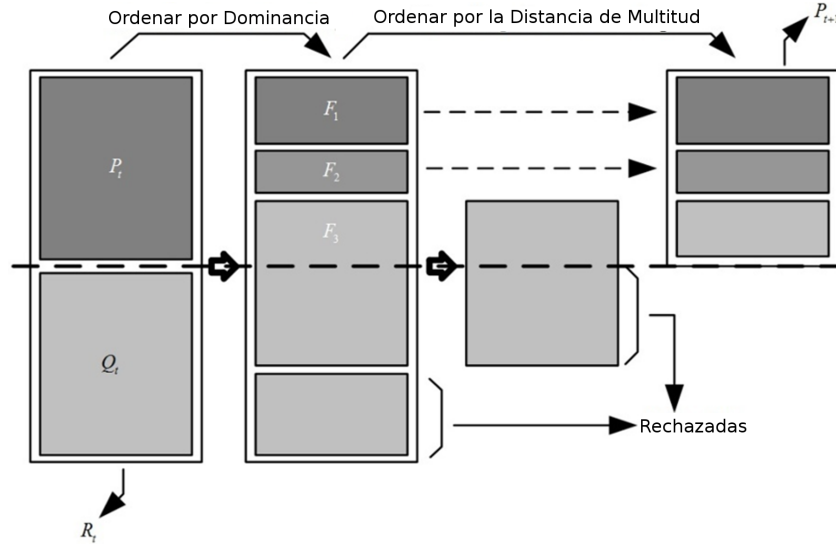


Figura 3.10: Procedimiento de la técnica NSGA-II - [Deb et al., 2002b].

espacio vacío en P_{t+1} es menor que el número de soluciones en F_3 , en ese caso, las soluciones en F_3 son ordenadas (en orden descendente) según su distancia de multitud (o crowding distance que explicaremos en breve) y las primeras soluciones que se necesitaba para completar P_{t+1} son insertadas. Por último, en la población P_{t+1} nuevamente se aplican los procesos de selección, cruzamiento y mutación para generar una nueva población Q_{t+1} , y el procedimiento es repetido hasta alcanzar un límite definido, por ejemplo, cantidad de repeticiones o generaciones.

La distancia de multitud para una solución j en una frontera F_i , definida por d_j , representa una estimativa del perímetro formado por el cuboide que tiene sus vértices en las soluciones vecinas más próximas a j que se encuentran en la misma frontera. Cuanto mayor sea la distancia de multitud, esa solución está más alejada sus soluciones vecinas, ello beneficia a las soluciones que están mejor diseminadas o dispersas en cada frontera. Las soluciones extremas para cada objetivo tienen una distancia de multitud infinita. La Figura 3.11 muestra como puede ser calculada la distancia de multitud para un problema de dos objetivos.

3. COMPUTACIÓN EVOLUTIVA

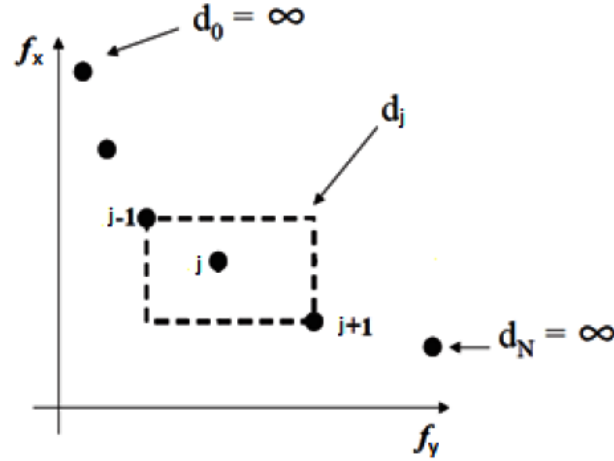


Figura 3.11: Distancia de multitud en la técnica NSGA-II - [Deb et al., 2002b].

3.4 Consideraciones Finales

En este capítulo presentamos el concepto básico de la computación evolutiva y su relación con la evolución natural. Después detallamos los AEs, dentro de ellos los AGs que son uno de los AEs más conocidos y utilizados. También explicamos los AGMO, que son una extensión de los AGs, utilizados para resolver problemas multiobjetivo, y por último la técnica NSGA-II que es una de las técnicas de optimización multiobjetivo más utilizadas, bien sucedidas y utilizada en este trabajo. En el próximo capítulo detallamos las principales características y taxonomía de los Sistemas Fuzzy Genéticos.

Sistemas Fuzzy Genéticos

Los sistemas fuzzy son sistemas basados en la teoría de conjuntos fuzzy y en la lógica fuzzy. Los sistemas fuzzy fueron propuestos por el profesor Lofti A. Zadeh en 1965 [Zadeh, 1965] para resolver problemas con información vaga, imprecisa e incompleta, abundantes en el mundo real. Para entender mejor estos tipos de información, a continuación mostramos algunos ejemplos acerca de la hora de una reunión:

- Información vaga: La reunión comienza más o menos a las 8 de la mañana.
- Información imprecisa: La reunión comienza entre las 8 y las 9 de la mañana.
- Información incompleta: No sé a que hora comienza la reunión, pero usualmente comienza a las 8 de la mañana.

Para tratar con ese tipo de información, Zadeh desarrolló una extensión de la teoría de conjuntos clásicos, definida como teoría de conjuntos fuzzy, donde la principal característica de los elementos del grupo es su grado de pertenencia, el cual es definido por un valor entre 0 y 1. Por ejemplo, en la teoría de conjuntos clásica, al tratar el estado de una puerta, generalmente tendremos dos estado: puerta abierta (representada por el valor 1) y puerta cerrada (representada por el valor

4. SISTEMAS FUZZY GENÉTICOS

0). En cambio, en la teoría de conjuntos fuzzy, podríamos manejar una cantidad infinita de estados representados con todos los valores entre 0 y 1, por ejemplo: puerta abierta (representada por el valor 1), puerta bastante abierta (representada por el valor 0.8), puerta media abierta (representada por el valor 0.5), puerta casi cerrada (representada por el valor 0.1), puerta cerrada (representada por el valor 0), entre otros.

Uno de los principales sistemas que utilizan la teoría de conjuntos fuzzy y la lógica fuzzy para inferir conclusiones, a partir información vaga, imprecisa o incompleta, son los Sistemas Fuzzy Basados en Reglas (SFBR) y dentro de ellos los Sistemas de Clasificación Fuzzy Basados en Reglas (SCFBR) que son de interés principal para este trabajo.

Cuando utilizamos los AGs, vistos en el Capítulo 3, para generar y/o adaptar partes de los SFBR o SCFBR nos encontramos dentro de la línea de investigación Sistemas Fuzzy Genéticos. En las próximas secciones son presentados los conceptos de la teoría de conjuntos fuzzy, los SFBR (dentro de ellos los SCFBR) y los sistemas fuzzy genéticos. Todos los conceptos presentados en este capítulo pueden ser encontrados en forma total en [Leon, 2005], [Nicoletti and Camargo, 2004], [Klir and Yuan, 1995], [Pedrycz and Gomide, 1998], [Herrera, 2005] y [Herrera, 2008].

4.1 Teoría de Conjuntos Fuzzy

De una forma resumida, un conjunto fuzzy es la generalización de un conjunto clásico con el objetivo de representar conjuntos cuyas fronteras no están claramente definidas [Nicoletti and Camargo, 2004].

Según la teoría de conjuntos clásica, un elemento pertenece o no pertenece a un conjunto, no existe situación intermedia, es decir, no existen miembros parciales de un conjunto. Por ejemplo, para determinar la edad de una persona, podemos considerar una persona como adolescente a partir de 12 años de edad, siguiendo la teoría de conjuntos clásica, una persona de 11 años de edad sería aún un niño, tal como muestra la Figura 4.1.

Lo mostrado en la Figura 4.1 puede que no represente en forma correcta la fase etaria de una persona en determinados contextos, por ello puede traer consecuencias no deseadas cuando queramos obtener un resultado a partir de esa información,

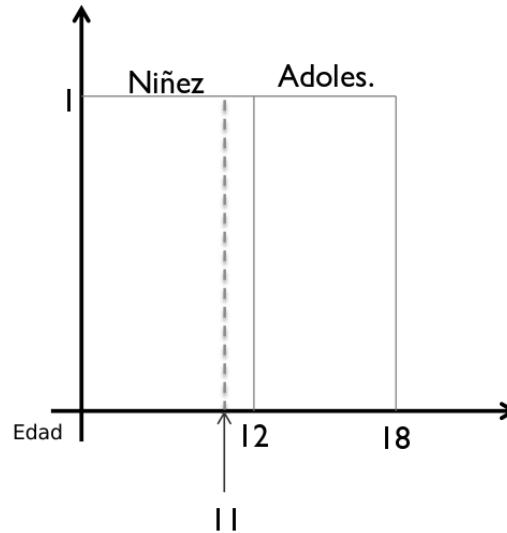


Figura 4.1: Ejemplo de conjuntos clásicos.

dado que una persona de 11 años de edad puede ser considerada con diferentes grados como niña o adolescente.

Para intentar solucionar ese problema, en la teoría de conjuntos fuzzy un elemento puede pertenecer parcialmente a uno o más conjuntos. En el contexto del ejemplo que estamos planteando sobre la edad, una persona de 11 años de edad puede pertenecer con mayor grado al conjunto adolescente, por ejemplo 0.8, y con menor grado al conjunto niñez, por ejemplo 0.2, como muestra la Figura 4.2, y con ello podemos representar mejor la realidad en este caso.

El grado de pertenencia μ_F es determinado mediante una función de pertenencia. Una función de pertenencia define un conjunto fuzzy y puede ser denotada por la siguiente ecuación.

$$\mu_F : U \rightarrow [0, 1] \quad (4.1)$$

Donde U representa el universo en discurso el cual contiene todos los posibles valores de un determinado contexto o aplicación. Existen diversas formas de funciones de pertenencia para definir conjuntos fuzzy, detallaremos las más conocidas y utilizadas en la siguiente sección.

4. SISTEMAS FUZZY GENÉTICOS

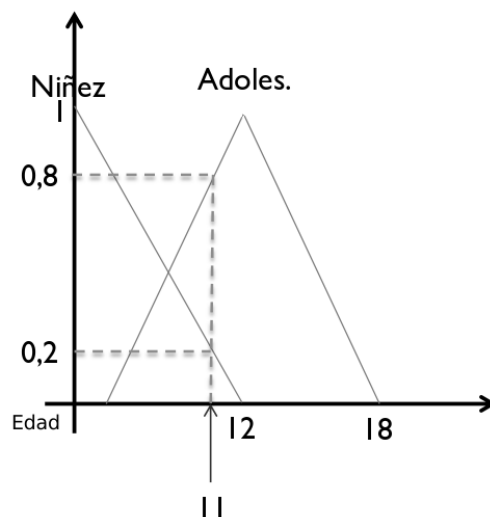


Figura 4.2: Ejemplo de conjuntos fuzzy.

4.2 Funciones de Pertenencia para Conjuntos Fuzzy

Las funciones de pertenencia para conjuntos fuzzy más conocidas y utilizadas son: triangular, trapezoidal, logística, gaussiana y conjunto unitario. Dichas funciones son las más utilizadas porque ellas tienen una simple y fácil representación, así como el cálculo del grado de pertenencia puede ser realizado directamente, como veremos a continuación:

- **Función triangular:** Definida por los parámetros (a, m, b) , tal que $a \leq m \leq b$. La función es mostrada en la Figura 4.3.

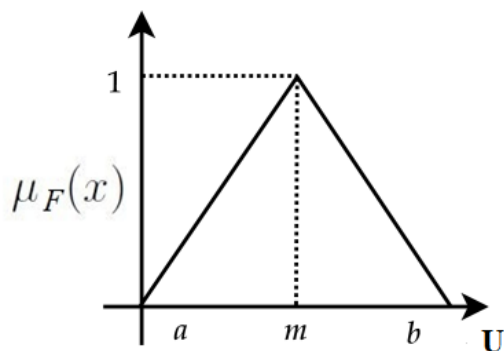


Figura 4.3: Función de pertenencia triangular.

4.2 Funciones de Pertenencia para Conjuntos Fuzzy

El grado de pertenencia de un elemento x en ese conjunto puede ser calculado usando la ecuación 4.2.

$$\mu_F(x) = \begin{cases} 0 & \text{si } x \leq a \\ \frac{x-a}{m-a} & \text{si } x \in (a, m) \\ 1 & \text{si } x = m \\ \frac{b-x}{b-m} & \text{si } x \in (m, b) \\ 0 & \text{si } x \geq b \end{cases} \quad (4.2)$$

- Función trapezoidal: Definida por los parámetros (a, m, n, b) , tal que $a \leq m < n \leq b$. La función es mostrada en la Figura 4.4.

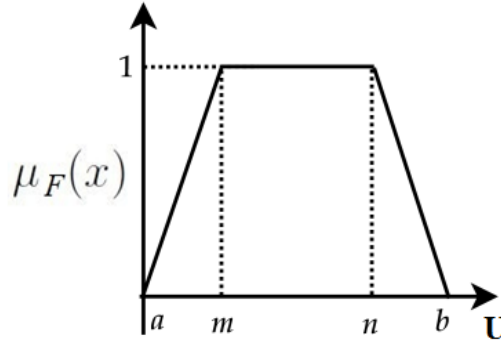


Figura 4.4: Función de pertenencia trapezoidal.

El grado de pertenencia de un elemento x en ese conjunto puede ser calculado usando la ecuación 4.3.

$$\mu_F(x) = \begin{cases} 0 & \text{si } x \leq a \\ \frac{x-a}{m-a} & \text{si } x \in (a, m) \\ 1 & \text{si } x \in [m, n] \\ \frac{b-x}{b-n} & \text{si } x \in (n, b) \\ 0 & \text{si } x \geq b \end{cases} \quad (4.3)$$

- Función logística: También llamada curva logística o curva en forma de S, es definida por los parámetros (a, m, b) tal que $m = (a + b)/2$. La función es mostrada en la Figura 4.5.

4. SISTEMAS FUZZY GENÉTICOS

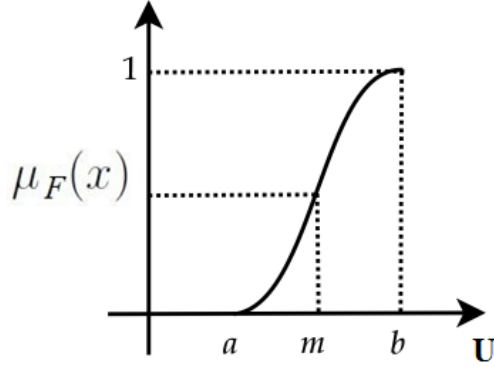


Figura 4.5: Función de pertenencia logística.

El grado de pertenencia de un elemento x en ese conjunto puede ser calculado usando la ecuación 4.4.

$$\mu_F(x) = \begin{cases} 0 & \text{si } x \leq a \\ 2\left(\frac{x-a}{b-a}\right)^2 & \text{si } x \in (a, m] \\ 1 - 2\left(\frac{x-b}{b-a}\right)^2 & \text{si } x \in (m, b] \\ 1 & \text{si } x > b \end{cases} \quad (4.4)$$

- Función gaussiana: Definida por los parámetros (m, k) tal que $k > 0$. La función es mostrada en la Figura 4.6.

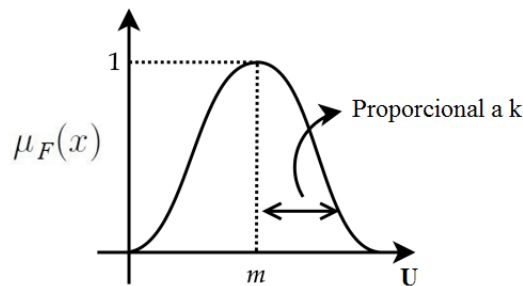


Figura 4.6: Función de pertenencia gaussiana.

El grado de pertenencia de un elemento x en ese conjunto puede ser calculado usando la ecuación 4.5.

$$\mu_F(x) = e^{-k(x-m)^2} \quad (4.5)$$

4.3 Conceptos Relacionados a los Conjuntos Fuzzy

- Conjunto Unitario: Definida por los parámetros (m, h) tal que $h \leq 1$. La función es mostrada en la Figura 4.7.

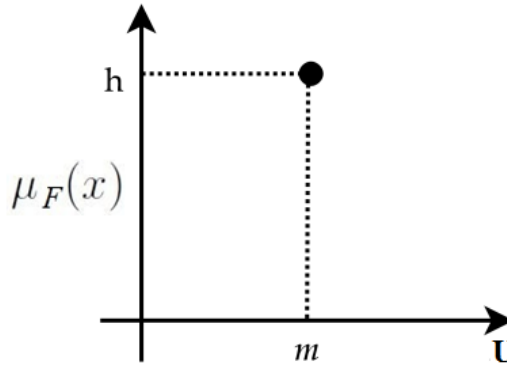


Figura 4.7: Función de pertenencia unitaria.

El grado de pertenencia de un elemento x en ese conjunto puede ser calculado usando la ecuación 4.6.

$$\mu_F(x) = \begin{cases} h & \text{si } x = m \\ 0 & \text{si } x \neq m \end{cases} \quad (4.6)$$

4.3 Conceptos Relacionados a los Conjuntos Fuzzy

A continuación explicamos los conceptos de altura, conjunto normal, soporte, núcleo, variable lingüística, término lingüístico y las operaciones básicas entre conjuntos fuzzy. Dichos conceptos son importantes para entender el funcionamiento de los SFBR.

- Altura: La altura, $h(F)$, de un conjunto fuzzy F es el mayor grado de pertenencia obtenido por cualquier elemento de F .
- Conjunto normal: Un conjunto fuzzy F es llamado de normal cuando $h(F) = 1$.
- Soporte: El soporte de un conjunto fuzzy F , en el conjunto universo U , es el conjunto clásico que contiene todos los elementos de U que tienen un grado de pertenencia mayor a cero en F . Un ejemplo es mostrado en la Figura 4.8.

4. SISTEMAS FUZZY GENÉTICOS

- Núcleo: El núcleo de un conjunto fuzzy F , en el conjunto universo U , es el conjunto clásico que contiene todos los elementos de U que tiene un grado de pertenencia igual a 1 en F (considerando un conjunto normal). Un ejemplo es mostrado en la Figura 4.8.

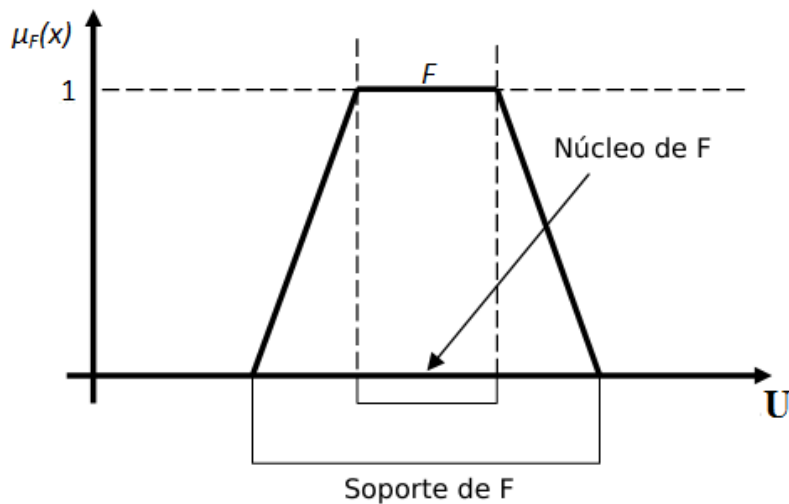


Figura 4.8: Soporte y núcleo de un conjunto normal fuzzy F .

- Variable Lingüística: Una variable lingüística es una variable cuyos valores son definidos en forma de lenguaje natural, por ejemplo: temperatura, altura, velocidad, distancia, etc. Dentro de un SFBR una variable lingüística puede definir una característica de uno o varios ejemplos y debe contener todos los valores posibles para un conjunto universo U determinado.
- Término lingüístico: Un término lingüístico es una etiqueta o un valor a una variable lingüística, por ejemplo: niñez, adolescencia, alto, bajo, rápido, etc. Dentro de un SFBR un término lingüístico debe estar asociado a un conjunto fuzzy.

En la siguiente subsección mostramos las operaciones básicas entre conjuntos fuzzy.

4.3.1 Operaciones Básicas entre Conjuntos Fuzzy

Las operaciones básicas entre conjuntos fuzzy que sirven para comprender mejor los conceptos mencionados en las próximas secciones y capítulos son: la intersección, unión y complemento, las mismas operaciones que conocemos en la teoría de conjuntos clásicos.

Cuando aplicamos esas operaciones en los conjuntos clásicos el resultado es un nuevo conjunto clásico, mientras que cuando aplicamos esas operaciones en los conjuntos fuzzy el resultado es un nuevo conjunto fuzzy.

Para mostrar el resultado de las operaciones mencionadas, consideraremos dos conjuntos fuzzy, el conjunto fuzzy triangular F y el conjunto fuzzy trapezoidal G , definidos en el universo U , mostrados en la Figura 4.9.

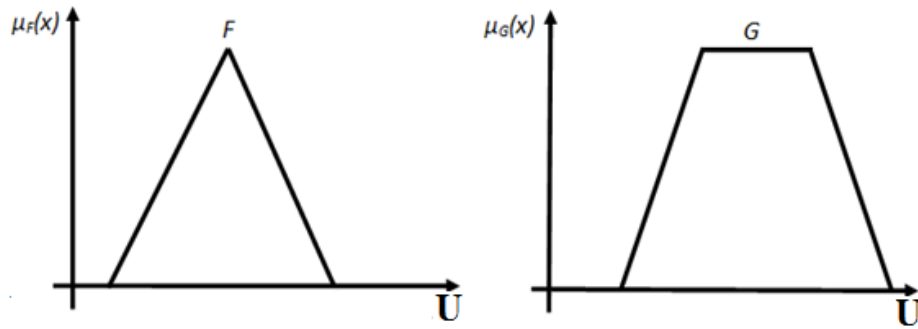


Figura 4.9: Conjuntos fuzzy F y G .

En base a esos conjuntos fuzzy, detallamos a continuación las operaciones básicas entre ellos:

- **Intersección:** La intersección entre dos conjuntos fuzzy F y G es el conjunto fuzzy E definido como $E = F \cap G$, donde para todo $x \in U$ se cumple la ecuación 4.7.

$$\mu_{F \cap G}(x) = \min[\mu_F(x), \mu_G(x)] \quad (4.7)$$

En forma gráfica la intersección es mostrada en la Figura 4.10.

4. SISTEMAS FUZZY GENÉTICOS

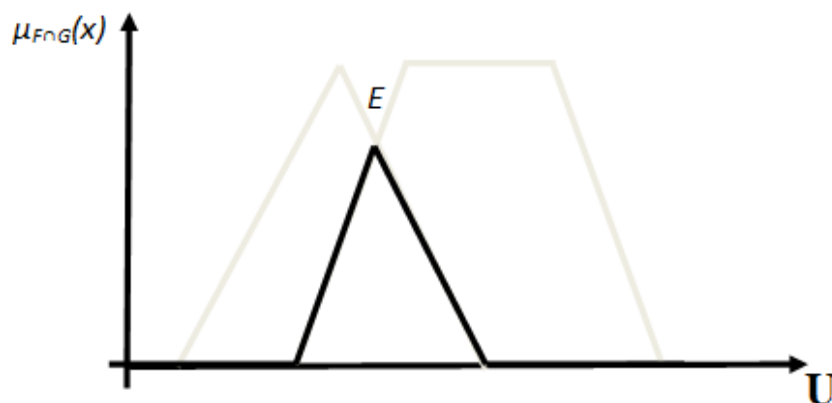


Figura 4.10: Intersección entre conjuntos fuzzy.

- Unión: La unión entre dos conjuntos fuzzy F y G es el conjunto fuzzy E definido como $E = F \cup G$, donde para todo $x \in U$ se cumple la ecuación 4.8.

$$\mu_{F \cup G}(x) = \max[\mu_F(x), \mu_G(x)] \quad (4.8)$$

En forma gráfica la unión es mostrada en la Figura 4.11.

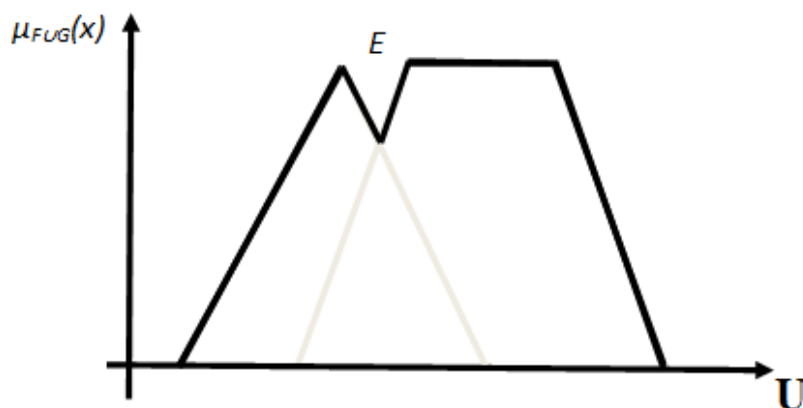


Figura 4.11: Unión entre conjuntos fuzzy.

- Complemento: El complemento de un conjunto fuzzy G es definido como $\sim G$, donde para todo $x \in U$ se cumple la ecuación 4.9.

$$\mu_{\sim G}(x) = 1 - \mu_G(x) \quad (4.9)$$

sub componen

En forma gráfica el complemento es mostrado en la Figura 4.12.

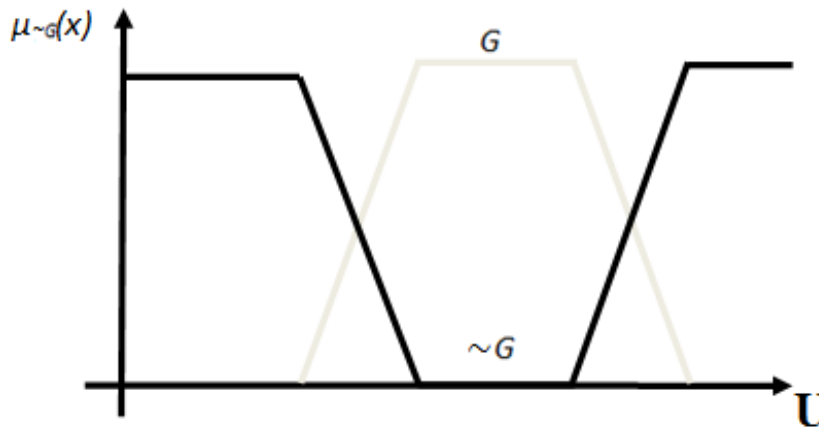


Figura 4.12: Complemento de un conjunto fuzzy.

4.4 Sistemas Fuzzy Basados en Reglas

Las reglas son un popular paradigma para representar conocimiento. Un Sistema Basado en Reglas (SBR) es un sistema donde la base de conocimiento de un dominio dado está codificado en forma de reglas [Jang and Sun, 1997]. Además de ello, las reglas tienen la ventaja de pueden ser definidas en forma lingüística con lo cual hacen que sean interpretables. En forma general, una regla puede tener la siguiente estructura:

SI (antecedente) **ENTONCES** (consecuente)

Cuando en los antecedentes y consecuentes de una regla son consideradas variables y términos lingüísticos basados en conjuntos fuzzy para representar conocimiento, nos estamos refiriendo a un SFBR; y si los consecuentes son considerados como clases, nos estamos refiriendo a Sistemas de Clasificación Fuzzy Basados en Reglas (SCFBR).

Las reglas fuzzy pueden establecer relaciones entre las variables de su antecedente, también llamadas condiciones o premisas, y las variables de sus consecuentes, también llamadas conclusiones o acciones. El antecedente de una regla fuzzy es

4. SISTEMAS FUZZY GENÉTICOS

por lo general una proposición fuzzy formada por una o un conjunto de conyunciones y/o disyunciones. Esas proposiciones pueden ser atómicas o compuestas. Por ejemplo, a continuación mostramos una regla fuzzy con una proposición atómica en el antecedente:

SI el Peso ES Alto ENTONCES el Precio ES Medio

Que puede ser representado de la siguiente forma:

SI V_1 ES T_1 ENTONCES C_1 ES TC_1

Donde:

V_1 : Variable lingüística 1 del antecedente (Peso).

T_1 : Término lingüístico 1 del antecedente (Alto).

C_1 : Variable lingüística 1 del consecuente (Precio).

TC_1 : Término lingüístico 1 del consecuente (Medio).

La siguiente regla fuzzy muestra un ejemplo con un antecedente compuesto que contine una conyunción y una diyunción:

**SI el Peso ES Alto Y el Cantidad ES Baja O el Impuesto ES Medio
ENTONCES el Precio ES Bajo**

Que puede ser representada de la siguiente forma:

SI V_1 ES T_1 Y V_2 ES T_2 O V_3 ES T_3 ENTONCES C_1 ES TC_1

Donde:

V_1 : Variable lingüística 1 del antecedente (Peso).

T_1 : Término lingüístico 1 del antecedente (Alto).

V_2 : Variables lingüística 2 del antecedente (Cantidad).

T_2 : Término lingüístico 2 del antecedente (Baja).

V_3 : Variables lingüística 3 del antecedente (Impuesto).

T_3 : Término lingüístico 3 del antecedente (Medio).

C_1 : Variable lingüística 1 del consecuente (Precio).

TC_1 : Término lingüístico 1 del consecuente (Bajo).

4.4 Sistemas Fuzzy Basados en Reglas

El consecuente de una regla fuzzy también puede ser una proposición, con una o un conjunto de conyunciones y/o disyunciones. Esas proposiciones pueden ser atómicas (como fue mostrado en los ejemplos anteriores) o compuestas. Por ejemplo, a continuación mostramos una regla fuzzy con una proposición compuesta en el consecuente:

**SI el Peso ES Alto Y el Cantidad ES Baja O el Impuesto ES Medio
ENTONCES el Precio ES Bajo Y la Ganancia ES Media**

Que puede ser representada de la siguiente forma:

SI V_1 ES T_1 Y V_2 ES T_2 O V_3 ES T_3 ENTONCES C_1 ES TC_1 Y C_2 ES TC_2

Donde:

- V_1 : Variable lingüística 1 del antecedente (Peso).
- T_1 : Término lingüístico 1 del antecedente (Alto).
- V_2 : Variables lingüística 2 del antecedente (Cantidad).
- T_2 : Término lingüístico 2 del antecedente (Baja).
- V_3 : Variables lingüística 3 del antecedente (Impuesto).
- T_3 : Término lingüístico 3 del antecedente (Medio).
- C_1 : Variable lingüística 1 del consecuente (Precio).
- TC_1 : Término lingüístico 1 del consecuente (Bajo).
- C_2 : Variable lingüística 2 del consecuente (Ganancia).
- TC_2 : Término lingüístico 2 del consecuente (Media).

Según los ejemplos mostrados, podemos definir formalmente una regla fuzzy mediante la siguiente expresión:

SI P ENTONCES Q

Las proposiciones P y Q pueden ser expresadas de la siguiente forma, siendo $n \geq 1$ y $m \geq 1$:

**SI V_1 ES T_1 Op V_2 ES T_2 Op ... Op V_n ES T_n
ENTONCES C_1 ES TC_1 Op C_2 ES TC_2 Op ... Op C_m ES TC_m**

4. SISTEMAS FUZZY GENÉTICOS

Donde:

$V_1, V_2, \dots, V_n$ y $C_1, C_2, \dots, C_m$: Variables lingüísticas dentro de cada universo.

$T_1, T_2, \dots, T_n$ y $TC_1, TC_2, \dots, TC_m$: Términos lingüísticos o conjuntos fuzzy asignados a cada variable lingüística.

Op : Operador de conyunción o disyunción.

Cada regla fuzzy puede representar una semántica o significado referente a la interpretación de las variables y términos lingüísticos en el antecedente y consecuente de la regla. Dicha semántica será analizada en la siguiente subsección.

4.4.1 Semántica de las Reglas Fuzzy

Si consideramos una regla fuzzy simple como la que vimos anteriormente:

SI V_1 ES T_1 ENTONCES C_1 ES TC_1

Esa regla puede ser expresada como una relación entre las variables lingüísticas V_1 y C_1 en función a los grados de pertenencia obtenidos en los términos lingüísticos o conjunto fuzzy normal T_1 y TC_1 , que pueden ser descritas como una función f con valores entre 0 y 1.

$$f : [0, 1]^2 \rightarrow [0, 1]$$

La selección de la función f es esencial para un SFBR dado que determina la semántica de las reglas en varios aspectos del sistema. Cuando nos referimos a la relación entre los antecedentes, en forma general, esa función puede ser derivada de dos tipos de funciones: $t - normas$ y $s - normas$. A continuación detallamos resumidamente esos tipos de funciones, así como algunas operaciones consideradas $t - normas$ o $s - normas$.

- $t - norma$ o Norma Triangular (t): Es una operación binaria $t : [0, 1]^2 \rightarrow [0, 1]$ que satisface los siguientes propiedades, para todo valor de $a, b, c \in [0, 1]$:

4.4 Sistemas Fuzzy Basados en Reglas

Conmutativa	: $t(a, b) = t(b, a)$
Asociativa	: $t(a, t(b, c)) = t(t(a, b), c)$
Condición frontera	: $t(a, 1) = a$
Monotonicidad	: Si $b \leq c$ entonces $t(a, b) \leq t(a, c)$

Entre otras, las tres operaciones mostradas a continuación son consideradas como t – *normas*:

Intersección	: $a \mathbf{t} b = \min[a, b]$
Producto Algebraico	: $a \mathbf{t} b = ab$
Diferencia Limitada	: $a \mathbf{t} b = \max[0, a + b - 1]$

- s – *norma* o Co-Norma Triangular (s): Es una operación binaria $s : [0, 1]^2 \rightarrow [0, 1]$ que satisface las siguientes propiedades, para todo valor de $a, b, c \in [0, 1]$:

Conmutativa	: $s(a, b) = s(b, a)$
Asociativa	: $s(a, s(b, c)) = s(s(a, b), c)$
Condición frontera	: $s(a, 0) = a$
Monotonicidad	: Si $b \leq c$ entonces $s(a, b) \leq s(a, c)$

Entre otras, las tres operaciones mostradas a continuación son consideradas como s – *normas*:

Unión	: $a \mathbf{s} b = \max[a, b]$
Suma Algebraica	: $a \mathbf{s} b = a + b - ab$
Suma Limitada	: $a \mathbf{s} b = \min[0, a + b]$

Por el lado del consecuente, también se pueden considerar t -normas, s -normas u otros tipos de funciones para expresar su semántica. Esas funciones no las mencionaremos por no estar en el ámbito de este trabajo, dado que utilizaremos reglas fuzzy para clasificación. Como veremos en la próxima sección el consecuente de una regla fuzzy para clasificación es un valor discreto, en este caso, una clase.

4. SISTEMAS FUZZY GENÉTICOS

4.5 Sistemas de Clasificación Fuzzy Basados en Reglas

Como fue mencionado anteriormente, clasificación es una tarea importante en diferentes áreas del mundo real. Resumiendo, el proceso de clasificación se basa en dado un conjunto de ejemplos $E = \{e_1, e_2, \dots, e_p\}$, el objetivo es asignar una clase C_j de un conjunto de clases $C = \{C_1, C_2, \dots, C_m\}$ a cada ejemplo $e_q \in E$, donde cada e_q es descrito por un conjunto de n atributos $e_q = \{a_{q1}, a_{q2}, \dots, a_{qn}\}$.

Dentro de los diferentes métodos utilizados para realizar la tarea de clasificación, encontramos los SFBR, que son denominados Sistemas de Clasificación Fuzzy Basados en Reglas (SCFBR). La principal diferencia entre un SFBR y un SCFBR es que en los SCFBR las reglas son proyectadas para resolver problemas de clasificación, por lo cual el consecuente de las reglas generalmente determinan una clase. Una regla típica de un SCFBR puede ser expresada como sigue:

$$R_i: \text{SI } V_1 \text{ ES } T_{1l_1} \text{ Y } V_2 \text{ ES } T_{2l_2} \text{ Y } \dots \text{ Y } T_n \text{ ES } T_{nl_n} \\ \text{ENTONCES Clase} = C_j$$

Donde:

R_i	: Regla fuzzy con índice i .
$V_1, V_2, \dots, V_n$	: Variables lingüísticas o atributos de cada ejemplo.
$T_{1l_1}, T_{2l_2}, \dots, T_{nl_n}$	: Términos lingüísticos o conjuntos fuzzy utilizados para representar la clase C_j .
C_j	: Es la clase de la R_i .

Un SFBR o un SCFBR tienen dos componente principales [Herrera, 2005]: la base de conocimiento y el mecanismo de inferencia, que son mencionados a continuación:

1. Base de Conocimiento (BC): Este componente incluye dos subcomponentes:
 - (a) Base de Reglas (BR): Donde se almacenan el conjunto de reglas fuzzy.
 - (b) Base de Datos (BD): Donde se almacenan las definiciones de las variables lingüísticas o atributos de los ejemplos, así como los parámetros o funciones de los términos lingüísticos o conjuntos fuzzy.

2. Mecanismos de Inferencia (MI): Es el conjunto de operaciones de inferencia fuzzy para obtener un resultado dada una determinada entrada. En un SCFBR el MI se encarga de determinar la clase de un ejemplo. Los métodos más conocidos para realizar ese proceso son el método de clasificación fuzzy clásico y el método de clasificación fuzzy general. Ambos serán detallados en las siguientes subsecciones.

4.5.1 Método de Clasificación Fuzzy Clásico

También llamado método de la regla vencedora, en el cual la clase de la regla con mayor grado de compatibilidad con el ejemplo, es la clase del ejemplo. Para determinar la clase $C_j \in C = \{C_1, C_2, \dots, C_m\}$ de un ejemplo e_q representado por los n atributos $e_q = \{a_{q1}, a_{q2}, \dots, a_{qn}\}$ se siguen los siguientes pasos:

1. Determinar el grado de compatibilidad entre el ejemplo e_q y todas las reglas en la BR. Por lo general, el grado de compatibilidad puede ser determinado por una t – norma.
2. Encontrar la regla R_i dentro de la BR con mayor grado de compatibilidad con el ejemplo e_q .
3. Atribuir al ejemplo e_q la clase C_j que se encuentra en la regla R_i .

4.5.2 Método de Clasificación Fuzzy General

En este método la clase del ejemplo es determinada por una agregación de todas las clases con la misma regla. Para determinar la clase $C_j \in C = \{C_1, C_2, \dots, C_m\}$ de un ejemplo e_q representado por los n atributos $e_q = \{a_{q1}, a_{q2}, \dots, a_{qn}\}$ se siguen los siguientes pasos:

1. Determinar el grado de compatibilidad entre el ejemplo e_q y todas las reglas en la BR. Por lo general, el grado de compatibilidad puede ser determinado por una t – norma.

4. SISTEMAS FUZZY GENÉTICOS

2. Para cada clase en C , es calculada una agregación usando los grados de compatibilidad obtenidos en el paso anterior y las reglas con la misma clase consecuente.
3. Atribuir al ejemplo e_q la clase C_j con más alta agregación obtenida en el paso anterior.

4.6 Sistemas Fuzzy Genéticos

El término de Sistemas Fuzzy Genéticos (SFG) ha sido utilizado para definir un SF combinado con un proceso de aprendizaje (o generación) y/u optimización (o adaptación) basado en AGs. Los tipos de SFG son: redes neuronales fuzzy genéticas, algoritmos de agrupamiento fuzzy genéticos y los Sistemas Fuzzy Genéticos Basados en Reglas (SFGBR), siendo el último tipo de interés para este trabajo.

En el contexto de los SFGBR, los AGs han sido utilizados para el aprendizaje y optimización de los componentes de la BC, tanto la BR y la BD, y en la adaptación del MI. La clasificación u organización de los diferentes métodos de los SFGBR se basa en que componente de un SFBR es aprendido y/u optimizado. La Figura 4.13 muestra la taxonomía de los SFGBR.

Una explicación resumida de cada uno de los métodos puede ser encontrado en [Herrera, 2008], [Herrera, 2005] y [Cordon et al., 2004]. A continuación explicaremos resumidamente dos de ellos, los cuales son de interés para este trabajo: el aprendizaje genético de reglas fuzzy (Genetic rule learning (a priori DB)) y la adaptación genética de los parámetro de la BC (Genetic tuning of KB parameters). Dentro de los SFG, existe una línea de investigación que utiliza los AGMO en vez de los AG para todos los métodos de aprendizaje y/u optimización genética que será explicada en el próximo capítulo.

4.6.1 Aprendizaje Genético de Reglas Fuzzy

Muchos de los trabajos propuestos para el generación de la BC desde información numérica está enfocada en el aprendizaje de la BR mediante un AG, usando una BD predefinida. La Figura 4.14 muestra el proceso genético de aprendizaje de reglas fuzzy.

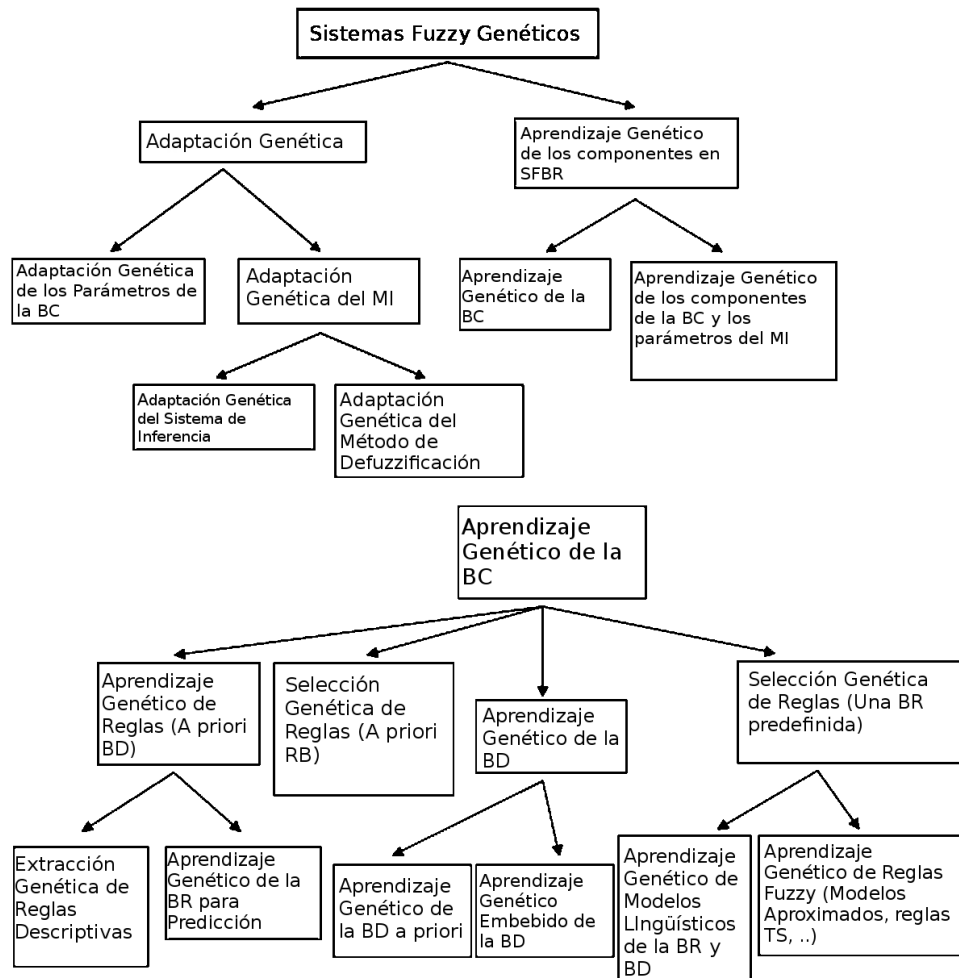


Figura 4.13: Taxonomía de los SFGBR - [Herrera, 2008]

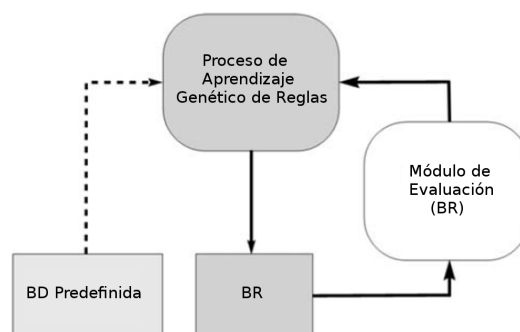


Figura 4.14: Proceso Genético de Aprendizaje de Reglas fuzzy - [Herrera, 2008]

4. SISTEMAS FUZZY GENÉTICOS

En el aprendizaje genético de reglas fuzzy, cuatro enfoques son comúnmente utilizados. El primero de ellos es el enfoque Pittsburgh donde cada cromosoma representa una BR total y la BR final es el mejor cromosoma obtenido después del proceso genético [Smith, 1980]. La Figura 4.15 muestra el enfoque Pittsburgh para el proceso genético de aprendizaje de reglas fuzzy.

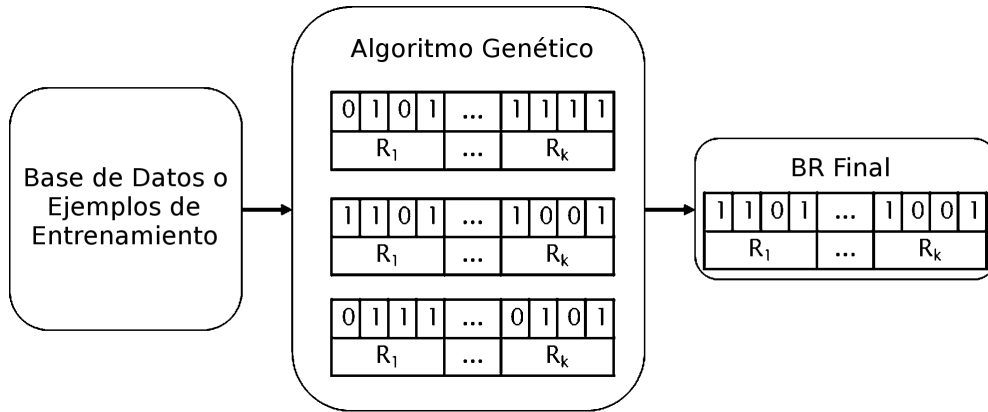


Figura 4.15: Enfoque Pittsburgh.

El segundo enfoque es el enfoque Michigan, donde cada regla fuzzy $Si \rightarrow Entoces$ es codificada en un cromosoma y la BR final es la unión de todos los cromosomas en la mejor población [Holland and Reitman, 1977]. La Figura 4.16 muestra el enfoque Michigan para el proceso genético de aprendizaje de reglas fuzzy.

El tercer enfoque es el enfoque iterativo o Iterative Rule Learning (IRL), donde cada regla fuzzy $Si \rightarrow Entoces$ es codificada en un cromosoma. La diferencia con el enfoque de Michigan es que en el enfoque IRL una nueva regla fuzzy es adaptada y adicionada a la BR final de forma iterativa en cada ejecución del AG [Gonzalez and Perez, 1999]. En este trabajo es utilizada este enfoque porque tiene las siguientes ventajas [Gonzalez and Perez, 2009]:

1. La codificación de una regla fuzzy por cromosoma no exige un cromosoma de gran tamaño, lo cual podría deteriorar el desempeño del AG.
2. Descompone el problema en subproblemas más simples que son resueltos en cada ejecución del AG, es decir, no intenta resolver todo el problema en una sola ejecución del AG.

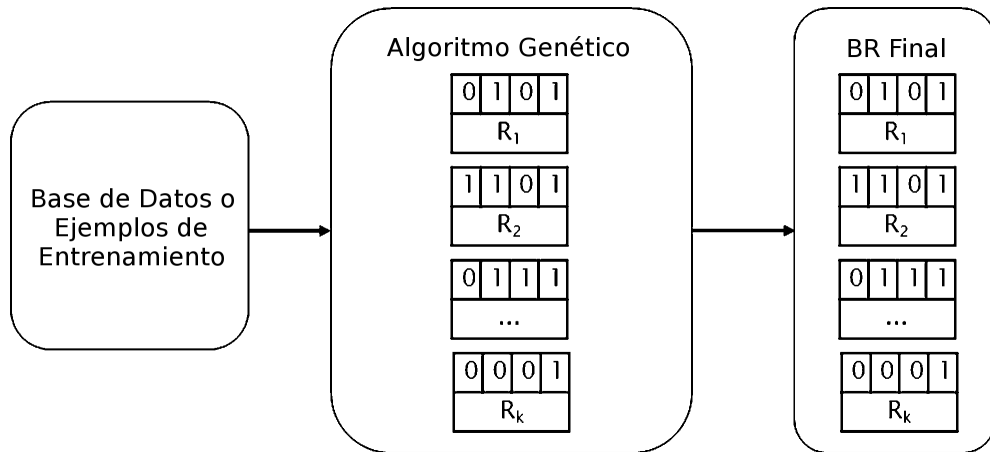


Figura 4.16: Enfoque Michigan.

La Figura 4.17 muestra el enfoque iterativo para el proceso genético de aprendizaje de reglas fuzzy.

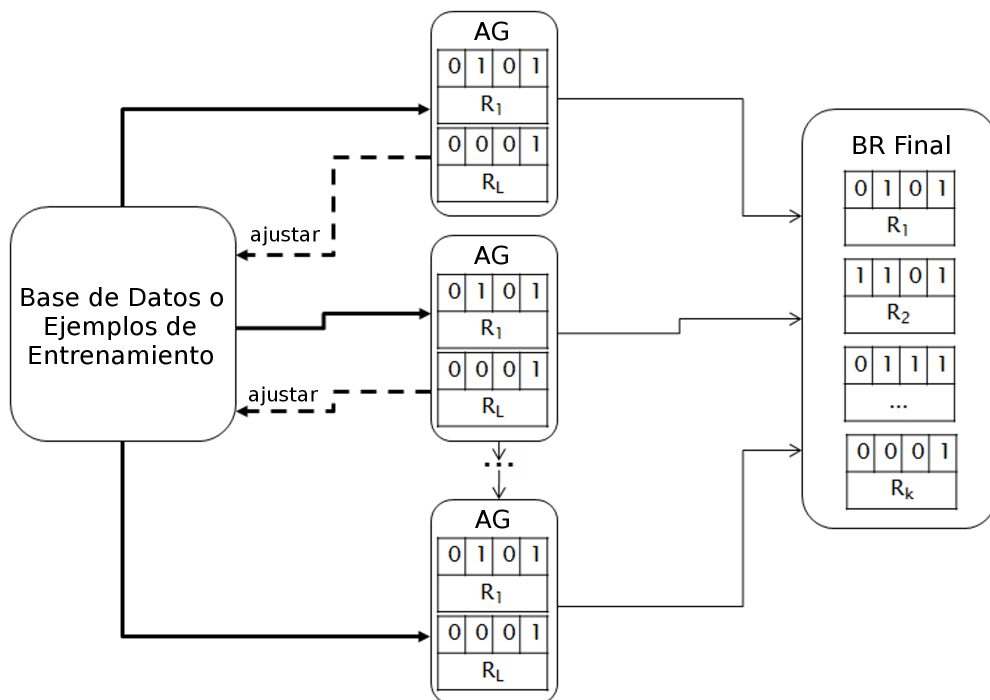


Figura 4.17: Enfoque iterativo o IRL.

El cuarto enfoque es el enfoque cooperativo-competitivo, en el cual cada cro-

4. SISTEMAS FUZZY GENÉTICOS

mosoma representa una BR total para iterar según el enfoque Pittsburgh que representa el enfoque competitivo, y cada cromosoma puede iterar según el enfoque Michigan para cooperar entre la reglas fuzzy dentro del cromosoma [Greene and Smith, 1993]. Se trata de un enfoque híbrido entre los enfoques Pittsburgh y Michigan. La Figura 4.18 muestra el enfoque cooperativo-competitivo para el proceso genético de aprendizaje de reglas fuzzy.

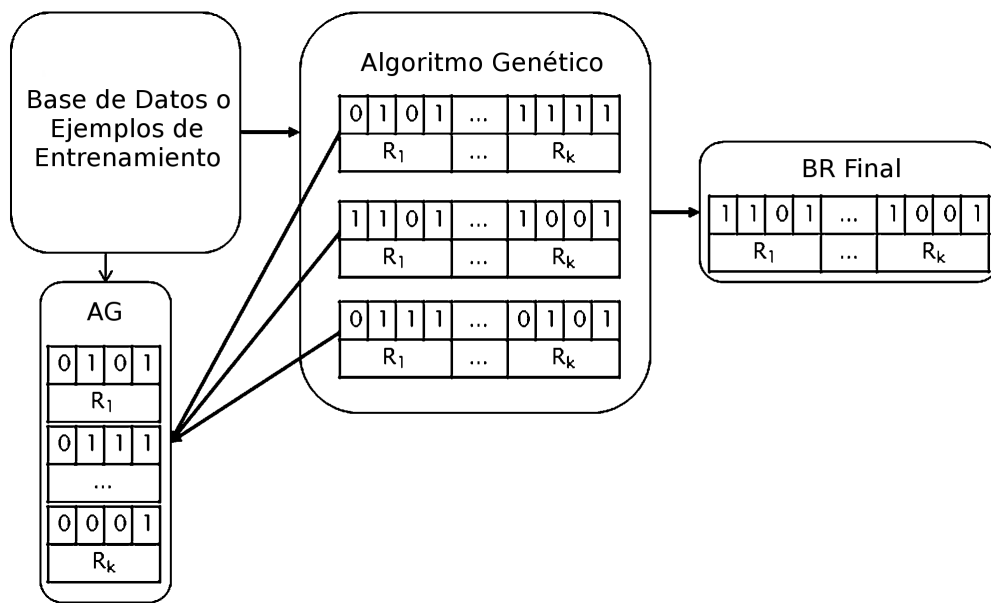


Figura 4.18: Enfoque cooperativo-competitivo.

4.6.2 Optimización Genética de los Parámetro de la BC

Este proceso de optimización o adaptación considera una BC obtenida, es decir, una BD predefinida y una BR generada, realizando un ajuste posterior de los parámetros de las funciones de pertenencia de los conjuntos fuzzy. Sin embargo, el proceso de optimización solo ajusta la forma de las funciones de pertenencia y no el número de terminos lingüísticos en cada variable lingüística, que permanece fijo desde el inicio. En [Karr, 1991] podemos encontrar la primera y clásica propuesta para esta adaptación. En este trabajo utilizaremos como base el método propuesto en [Gacto et al., 2009a] para esta tarea.

4.7 Consideraciones Finales

En este capítulo fueron abordados los conceptos básicos de los SF, teoría de conjuntos fuzzy, SFBR y SCFBR enfocados en los límites y compresión de este trabajo. Además de ello, fue analizada la línea de investigación definida como SFG que utiliza AG para generar y/o adaptar SF o SFBR. El siguiente capítulo destaca los últimos y principales trabajos que utilizan los AEMO o AGMO para generar y/u optimizar componentes de los SFBR o SCFBR, es decir, trabajos anteriores relacionados a este trabajo.

CAPÍTULO 5

Revisión de la Literatura

Como mencionamos anteriormente, la flexibilidad de los sistemas fuzzy hace que sean aplicables a una gran cantidad de áreas. La generación o adaptación de los componentes de los sistemas fuzzy puede ser considerado como un problema multiobjetivo, en los cuales cuando se mejora un objetivo lleva al empeoramiento de los otros, por lo tanto, generalmente no existe una solución única que optimice simultáneamente todos los objetivos.

Ese tipo de problemas puede ser tratado por los AEMO, vistos en el Capítulo 3, para el diseño de sistemas fuzzy. La optimización multiobjetivo lleva a obtener un conjunto de modelos fuzzy con diferentes valores de equilibrio entre todos los objetivos, en vez de considerar un solo objetivo. Después de ello, un tomador de decisiones (o un proceso automático de toma de decisiones), dependiendo de los requerimientos específicos, puede seleccionar el modelo más adecuado. Este tipo de enfoque es conocido como Sistemas Fuzzy Evolutivos Multiobjetivo (SFEM) o Multiobjective Evolutionary Fuzzy Systems (MOEFS).

Según [Fazzolari et al., 2013b], los SFEM pueden ser divididos en dos niveles, como muestra la Figura 5.1. El primer nivel se basa en la naturaleza multiobjetivo del problema a tratar, es decir, los tipos de objetivos analizados; y el segundo nivel

5.1 SFEM para generar SFBR con Equilibrio entre Precisión e Interpretabilidad

está basado en el componente del SFG que es optimizado durante el proceso evolutivo multiobjetivo. Ambos niveles determinan el tipo de espacio de búsqueda y la complejidad a ser analizada por el SFEM.

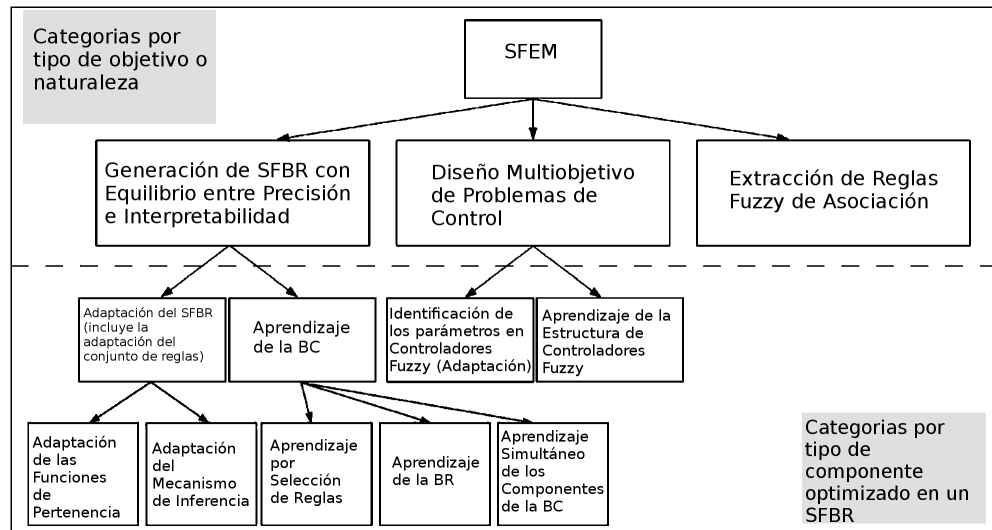


Figura 5.1: Taxonomía de dos niveles de los SFEM.- [Fazzolari et al., 2013b].

En este capítulo analizaremos el estado del arte de los SFEM que se encajan dentro del diseño de SFG, principalmente SFBR, para equilibrar los objetivos de precisión e interpretabilidad, es decir, la primera clase dentro del primer nivel; y todas las clases provenientes de ella que se encuentran en el segundo nivel, algunos de los cuales tratan el problema de bases de datos no balanceadas. Un resumen detallado de las otras clases en ambos niveles (que no serán tratadas en este trabajo) pueden ser encontrados en [Fazzolari et al., 2013b]. Seguidamente, también mencionamos algunos de los últimos y principales trabajos que utilizan SFG para tratar el problema de bases de datos no balanceadas.

5.1 SFEM para generar SFBR con Equilibrio entre Precisión e Interpretabilidad

Uno de los principales usos de los SFBR es la aproximación de sistemas reales en un modelo fuzzy, el cual puede ser utilizado para explicar, simular, o predecir el

5. REVISIÓN DE LA LITERATURA

comportamiento del sistema original, considerando que entre mayor sea la precisión, más fiable será el modelo.

Inicialmente, la interpretabilidad de un modelo fuzzy no fue considerada, porque los AEs mono objetivos permiten la optimización de solo una característica de los modelos fuzzy. El problema de mejorar la precisión mientras se mantiene o se mejora la interpretabilidad de un sistemas fuzzy fue tratada a mediados de los años 90 por Ishibuchi y su grupo [Ishibuchi et al., 1997], y la interpretabilidad o comprensión de los modelos fuzzy comienzan a ser integrados en el proceso de optimización gracias a la aplicación de los AEMO en los modelos fuzzy.

Desde ese momento, la interpretabilidad ha adquirido e incrementado su importancia en el área de investigación de SFMO, principalmente por su subjetividad. El principal problema de ello es encontrar un consenso común en la definición de interpretabilidad y medirla en los modelos fuzzy obtenidos, dado que diversos puntos deben ser tomados en cuenta para obtener un modelo fuzzy interpretable para una persona.

Sobre ese problema, varios trabajos han analizado la interpretabilidad en SFBR buscando una medida de interpretabilidad que sea aceptada por la comunidad de investigadores. Aún ese problema continua en abierto. Sin embargo, hoy en día los investigadores coinciden que dos medidas de interpretabilidad deben ser consideradas [Cordón, 2011]:

1. Medida basada en la complejidad: La cual es utilizada para disminuir la complejidad del modelo fuzzy, por ejemplo, disminuir el número de reglas o disminuir el número de antecedentes de las reglas. Siendo esta medida la más utilizada en las investigaciones y/o trabajos.
2. Medida basada en la semántica: La cual es utilizada para preservar la semántica asociada a las funciones de pertenencia de los conjuntos fuzzy, por ejemplo, la distinguibilidad y la cobertura; y de las reglas, por ejemplo, la consistencia de la reglas. Esta medida también es un problema en abierto.

En el diseño de modelos fuzzy, se sabe que existe un punto en el cual no es posible mejorar la precisión a la vez que se aumenta la interpretabilidad de un modelo fuzzy. La Figura 5.2 muestra la curva de equilibrio entre precisión e interpretabilidad en SFBR.

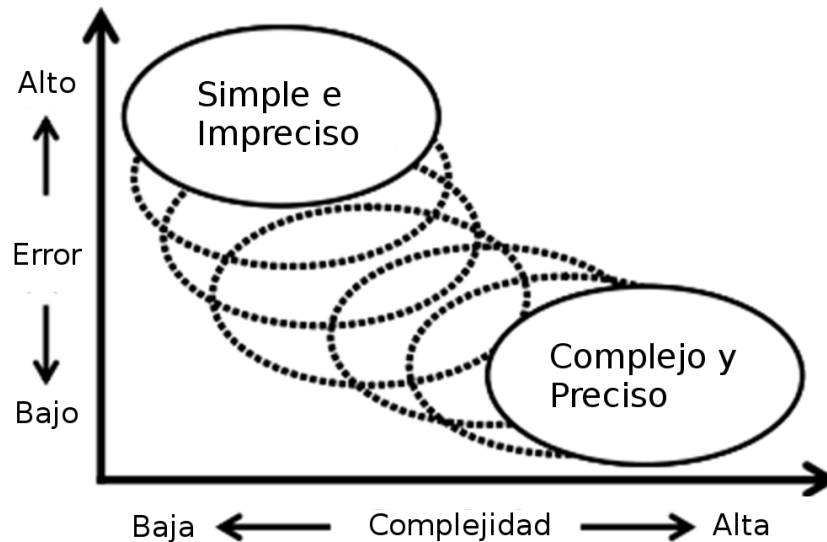


Figura 5.2: Curva de equilibrio entre precisión y complejidad en SFBR - [Ishibuchi, 2007]

A continuación, describimos un grupo de las contribuciones más relevantes y actuales en el área de SFMO que generan u optimizan SFBR encontrando valores equilibrados entre precisión e interpretabilidad, los agrupamos siguiendo el segundo nivel de la taxonomía presentada en la Figura 5.1 y son detallados resumidamente en las siguientes secciones.

5.2 Optimización en SFBR

Los AEMO pueden ser utilizados para realizar optimización evolutiva de los componentes de los SFBR. La optimización evolutiva es aplicada como un método de post-procesamiento, después de obtener la BR, para optimizar los parámetros de la BD u optimizar los parámetros del mecanismos de inferencia.

Los trabajos en este categoría pueden ser divididos en dos subcategorías: optimización de las funciones de pertenencia y optimización de los parámetros de inferencia. Además, en algunos casos el proceso de optimización puede ser combinado con un proceso de selección de reglas, para mejorar la interpretabilidad del modelo con el fin de remover reglas fuzzy innecesarias. En las siguientes subsecciones mostramos algunos de los últimos trabajos realizados en esta categoría.

5. REVISIÓN DE LA LITERATURA

5.2.1 Optimización de las Funciones de Pertenencia en la BD

Developing a bioaerosol detector using hybrid genetic fuzzy systems (2008)

El principal objetivo de este trabajo [Pulkkinen et al., 2008] es desarrollar un modelo fuzzy, el cual trabaja como un mecanismo de raciocinio en un detector de bioaerosol. La habilidad para distinguir entre aerosoles seguros y perjudiciales es uno de sus principales requisitos.

En este trabajo en vez de usar el error de clasificación como métrica de precisión son utilizadas las medidas de TP y FP, porque existe un costo de clasificación equivocada y un alto desbalanceamiento en la distribución de clases. La interpretabilidad constituye la confianza en el modelo fuzzy desarrollado y permite que el modelo fuzzy sea ajustado con facilidad. Por lo tanto, es otro requisito crucial para el modelo.

Como se ha mencionado anteriormenete, los objetivos se contradicen y por ello son aplicados los AEMO para encontrar modelos equilibrados en los dos objetivos. Clasificadores fuzzy son utilizados en este trabajo porque es un tipo de clasificador que posee reglas lingüísticas que son intuitivas para los seres humanos.

En el proceso de optimización multiobjetivo, el AEMO utiliza la transparencia de los conjuntos fuzzy como métrica de interpretabilidad; y las tasas de TP y FP como métricas de precisión. En el mismo proceso de optimización, se considera la reducción de la cantidad de las reglas y la cantidad de condiciones en las reglas, manteniendo la precisión del modelo. Los clasificadores fuzzy obtenidos son altamente comprensibles y precisos; y las reglas lingüísticas proporcionan información significativa para futuras modificaciones del detector de bioaerosol.

Adaptation and application of multi-objective evolutionary algorithms for rule reduction and parameter tuning of fuzzy rule-based systems (2009)

Este trabajo [Gacto et al., 2009b] analiza diferentes alternativas con la tarea de alcanzar un buen equilibrio entre precisión e interpretabilidad. La precisión se mejora optimizando o adaptando las funciones de pertenencia y la interpretabilidad reduciendo la complejidad del modelo. Para ello, se propone el uso de AEMO como herramienta para obtener al menos una mejor solución con respecto a un

enfoque mono objetivo clásico (es decir, una solución que domine la obtenida por ese enfoque en términos de tasa de clasificación y número de reglas).

También en este trabajo se presenta y analiza la aplicación de seis diferentes técnicas de AEMO para obtener modelos fuzzy lingüísticos precisos simples mediante la selección de reglas y la optimización de las funciones de pertenencia. Los resultados en dos escenarios diferentes muestran que el uso de conocimiento experto en el proceso de diseño del algoritmo mejora significativamente la capacidad de búsqueda de los AEMO y los mismos son capaces de mejorar ambos objetivos, obteniendo modelos más preciso y más sencillos con respecto a los algoritmos basados en un solo objetivo.

Context adaptation of fuzzy systems through a multi-objective evolutionary approach based on a novel interpretability index (2009)

Varias aplicaciones de los SFBR han levantado un gran interés entre investigadores y profesionales en el desarrollo de métodos eficaces y eficientes para generar SFBR con alta precisión y/o interpretabilidad. En este marco, la optimización o Adaptación al Contexto (AC) es una de las técnicas más prometedoras. La AC en los SFBR se basa en el siguiente supuesto: la BR modela el conocimiento universal y los significados asociados con los términos lingüísticos deben ser adaptados a fin de mejorar la precisión del SFBR. Por lo tanto, la AC es un proceso de optimización que explota la información específica del contexto con el fin de optimizar los parámetros contenidos en la BD.

La suposición anterior implica que la BR, generada por expertos, no debe modificarse durante el proceso de optimización de la BD, dado que existe una relación semántica entre el BR y la BD, esta relación obliga a algunas restricciones en la optimización de la BD.

Primero, considerando que cada término lingüístico está asociado a un conjunto fuzzy, cada conjunto fuzzy es útil y significativo solo si el término lingüístico al cual está asociado se utiliza en la BR. Por lo tanto, el número de conjuntos fuzzy en cada partición se determina directamente por el número de etiquetas lingüísticas definidas en la BR y por ello no debe variar durante la AC.

Segundo, cuando los expertos definen la BR, utilizan un orden semántico implícito de los términos lingüísticos que es importante y significativo para los seres hu-

5. REVISIÓN DE LA LITERATURA

manos. Por ejemplo, en el caso de que el término lingüístico *bajo* esté seguido del término lingüístico *alto*, y el término lingüístico *frío* preceda al término lingüístico *caliente*, ese orden no debe cambiar. Por consiguiente, para preservar la semántica del modelo fuzzy, el orden de los conjuntos fuzzy después de la AC debe ser el mismo que el original.

Resumiendo, un proceso de AC debería cumplir con los siguientes puntos:

1. La AC no debería modificar el BR;
2. La AC no debería cambiar la cantidad de términos lingüísticos definidos en la BR y, por consiguiente, el número de conjuntos fuzzy correspondientes;
3. La AC no debería afectar el orden semántico de los términos lingüísticos.

En este trabajo [Gacto et al., 2009c] se propone un método de AC para un SFBR que mejore la precisión mientras se sigan los puntos mencionados anteriormente y se mantenga la interpretabilidad de las particiones. Ese método utiliza un nuevo índice de interpretabilidad basada en relación de orden de los conjuntos fuzzy. El índice propuesto y el Error Cuadrático Medio (ECM) son utilizados como objetivos del AEMO para generar un conjunto de SFBR Pareto óptimos con diferentes valores de equilibrio entre precisión e interpretabilidad.

Integration of an index to preserve the semantic interpretability in the multiobjective evolutionary rule selection and tuning of linguistic fuzzy systems (2010)

En este trabajo [Gacto et al., 2010] se propone un índice que ayuda a preservar la interpretabilidad semántica de modelos lingüísticos fuzzy mientras se realiza la optimización de las funciones de pertenencia. El índice propuesto es la agregación de tres métricas que preservan el significado semántico original de las funciones de pertenencia tanto como sea posible mientras se realiza la optimización.

Este Índice de Interpretabilidad Semántica (IIS) está basado en la diferencia entre los conjuntos adaptados y el conjunto original. Para calcularlo primero son definidos los intervalos de variación para cada punto de cada conjunto fuzzy j . En el caso de los conjuntos fuzzy triangulares, esos intervalos son ilustrados en la

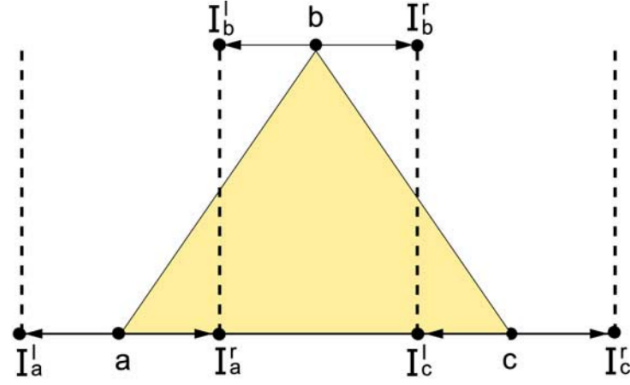


Figura 5.3: Intervalos de variación para conjuntos fuzzy triangulares - [Gacto et al., 2010]

Figura 5.3 y pueden ser calculados según las ecuaciones que se muestran a continuación.

$$[I_a^l, I_a^r] = [a - (\frac{b-a}{2}), a + (\frac{b-a}{2})] \quad (5.1)$$

$$[I_b^l, I_b^r] = [b - (\frac{b-a}{2}), b + (\frac{c-b}{2})] \quad (5.2)$$

$$[I_c^l, I_c^r] = [c - (\frac{c-b}{2}), c + (\frac{c-b}{2})] \quad (5.3)$$

Como fue mostrado en la figura anterior, cada punto puede variar entre la mitad de distancia del punto y el punto más próximo a la derecha y/o a la izquierda. A continuación detallamos las tres métricas utilizadas en el IIS:

1. Desplazamiento de los conjunto fuzzy (δ): Mide la proximidad de los puntos centrales de los conjuntos fuzzy adaptados j' y los conjuntos fuzzy originales j . El desplazamiento para un conjunto puede ser calculado por las siguientes ecuaciones:

$$\delta_j = \frac{|b_j - b'_j|}{I} \quad (5.4)$$

5. REVISIÓN DE LA LITERATURA

$$I = \frac{(I_{b_j}^r - I_{b_j}^l)}{2} \quad (5.5)$$

El máximo valor para el desplazamiento de todos los conjuntos es definido como δ^* . El cual puede ser calculado como:

$$\delta^* = \max \delta_j \quad (5.6)$$

El desplazamiento es un objetivo a maximizar, definido por la siguiente ecuación:

$$\text{Maximizar } \delta = 1 - \delta^* \quad (5.7)$$

2. Amplitud Lateral (γ): Mide la diferencia entre los parámetros, derecho e izquierdo, de los conjuntos fuzzy adaptados j' y los conjuntos fuzzy originales j . La amplitud para cada conjunto puede ser calculada por las siguientes ecuaciones:

$$\gamma_j = \frac{\min \left\{ \frac{izqS_j}{derS_j}, \frac{izqS'_j}{derS'_j} \right\}}{\max \left\{ \frac{izqS_j}{derS_j}, \frac{izqS'_j}{derS'_j} \right\}} \quad (5.8)$$

$$izqS_j = |a_j - b_j| \quad (5.9)$$

$$derS_j = |b_j - c_j| \quad (5.10)$$

$$izqS'_j = |a'_j - b'_j| \quad (5.11)$$

$$derS'_j = |b'_j - c'_j| \quad (5.12)$$

El mínimo valor para la amplitud de todos los conjuntos es definido como γ , la cual puede ser calculada como:

$$\gamma = \min \gamma_j \quad (5.13)$$

La amplitud es un objetivo a maximizar, definido por la siguiente ecuación:

$$\text{Maximizar } \gamma \quad (5.14)$$

3. Similaridad de áreas (ρ): Mide la diferencia entre las áreas de los conjuntos fuzzy adaptados j' y los conjuntos fuzzy originales j . La similaridad para cada conjunto puede ser calculada por la siguiente ecuación:

$$\rho_j = \frac{\min \{A_j, A'_j\}}{\max \{A_j, A'_j\}} \quad (5.15)$$

Donde A_j representa el área del conjunto fuzzy original y A'_j representa el área del conjunto fuzzy adaptado. El mínimo valor para la similaridad de todos los conjuntos es definido como ρ , la cual puede ser calculada como:

$$\rho = \min \rho_j \quad (5.16)$$

La similaridad es un objetivo a maximizar, definido por la siguiente ecuación:

$$\text{Maximizar } \rho \quad (5.17)$$

Las tres métricas tienen un valor entre 0 (menor interpretabilidad) y 1 (mayor interpretabilidad), las cuales son unidades en la función Geometric Mean of Three Metrics (GM3M) definida en la siguiente ecuación:

$$GM3M = \sqrt[3]{\delta \times \gamma \times \rho} \quad (5.18)$$

El IIS es un objetivo a maximizar, y es definido como:

$$\text{Maximizar } GM3M \quad (5.19)$$

Además, un mecanismo de selección de reglas es utilizado para reducir la complejidad del modelo, lo que implica otro aspecto importante de la interpretabilidad.

5. REVISIÓN DE LA LITERATURA

Para ello, se propone un AEMO aplicado a un SFBR ya definido para realizar la selección de las reglas fuzzy y la optimización de las funciones de pertenencia considerando tres objetivos: maximizar la precisión, maximizar la interpretabilidad semántica y minimizar la complejidad. Los resultados, comparados con enfoques mono objetivos para cada objetivo, confirman la eficacia del método propuesto, además de obtener un conjunto de soluciones más precisas e interpretables.

Multiobjective genetic fuzzy rule selection of single granularity-based fuzzy classification rules and its interaction with the lateral tuning of membership functions (2011)

La selección de reglas fuzzy genética multiobjetivo se basa en la generación de un conjunto de reglas fuzzy para clasificación que consideran una granularidad simple o una granularidad múltiple en las particiones fuzzy para cada variable lingüísticas, luego, se aplica un AEMO para llevar a cabo la selección de reglas fuzzy. El uso de granularidad múltiple ha sido considerada como un punto que puede llevar a un deterioro de la interpretabilidad. En este trabajo [Alcalá et al., 2011] se propone un método para definir una granularidad simple apropiada. Se realiza un estudio estadístico sobre el método propuesto y es extendido o unido a un ajuste lateral de las funciones de pertenencia. Por último, es comparado el método basado en granularidad simple con el uso de múltiple granularidad con y sin ajuste de las funciones de pertenencia. Los resultados muestran que el rendimiento de los SCFBR obtenidos mejoran mediante la definición de la granularidad y los parámetros de las funciones de pertenencia adecuados.

5.2.2 Optimización o Adaptación del Mecanismo de Inferencia

Rule base and inference system cooperative learning of Mamdani fuzzy systems with multiobjective genetic algorithms (2009)

Este trabajo [Márquez et al., 2009] presenta una modelo de aprendizaje evolutivo multiobjetivo para lograr una sinergia positiva entre el MI y la BR con el fin de obtener modelos lingüísticos fuzzy simples, compactos y precisos aprendiendo operadores de inferencia fuzzy junto a la BR. El AEMO propuesto genera un conjunto de SFBR no dominados entre interpretabilidad y precisión (cada uno de

los cuales con un alto grado de cooperación entre la BR y el MI) permitiendo que especialistas puedan seleccionar el más adecuado para la aplicación deseada.

La interpretabilidad está basada en el número de reglas y la precisión en la ECM, para ello el cromosoma utilizado en el AEMO (mostrado en la Figura 5.4) considera los métodos de conjunción, la defusificación y la BR.

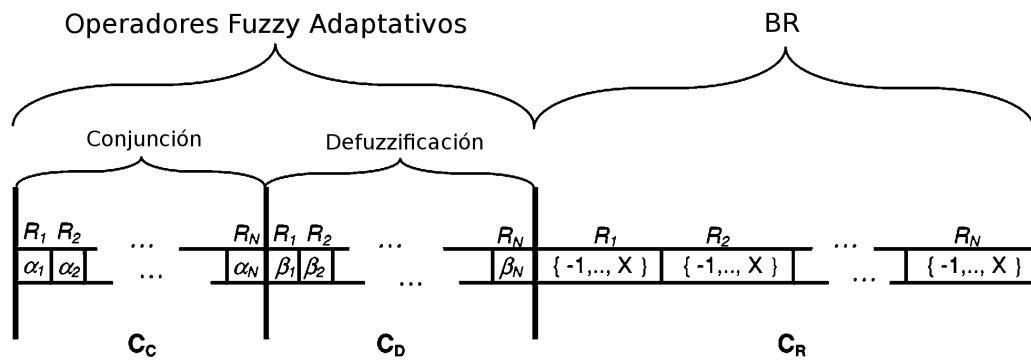


Figura 5.4: Codificación utilizada en el AEMO con N reglas - [Ishibuchi, 2007]

A multi-objective evolutionary algorithm with an interpretability improvement mechanism for linguistic fuzzy systems with adaptive defuzzification (2010)

En este trabajo [Marquez et al., 2010] se propone un AEMO con un mecanismo para mejorar la interpretabilidad, específicamente la complejidad, de los SFBR con la optimización del mecanismo de defusificación. El uso de parámetros en el operador de defusificación introduce una serie de valores o pesos asociados a cada regla, lo que mejora la precisión, pero aumenta la complejidad del SFBR y así un efecto negativo en la interpretabilidad del SFBR. Por ello, además de considerar como objetivo maximizar la precisión, en el proceso evolutivo multiobjetivo, también considera tres medidas: minimizar el número de reglas, minimizar el número de reglas con pesos asociados y el número promedio de reglas lanzadas por cada ejemplo; siendo las tres medidas unificadas para representar el objetivo de maximizar la interpretabilidad. El método propuesto se comparó con un estudio experimental que considera solo un objetivo basado en la precisión en dos problemas

5. REVISIÓN DE LA LITERATURA

reales, mostrando que muchas soluciones no dominadas en la frontera de Pareto dominan la solución encontrada con el método de un solo objetivo.

Multiobjective Genetic Optimization of Fuzzy Partitions and T-Norm Parameters in Fuzzy Classifiers (2012)

Este trabajo [Cardenas and Carmago, 2012] propone el uso de un AEMO para optimizar conjuntos fuzzy y definir la t-norma utilizada en el MI de un SCFBR. Los autores consideran un SCFBR definido, es decir, una BR y una BD definidas, donde se aplica un AEMO para optimizar las funciones de pertenencia en la BD y en forma simultánea buscar la t-norma dentro del MI que mejore la precisión del SFBR. El proceso de optimización está diseñado para manejar el equilibrio entre interpretabilidad y precisión. Los resultados muestran una mejora significativa en la precisión, manteniendo las misma interpretabilidad, de los SCFBR en la frontera de Pareto en comparación de las SCFBR donde solo se adaptó las funciones de pertenencia, sin considerar la optimización de la t-norma en el MI. La técnica de optimización multiobjetivo utilizada fue el Algoritmo Evolutivo Fuerte de Pareto 2 o Strength Pareto Evolutionary Algorithm 2 (SPEA2).

An efficient multi-objective evolutionary adaptive conjunction for high dimensional problems in linguistic fuzzy modelling (2012)

La optimización de los operadores de conjunción en el MI es una de las formas utilizadas para mejorar la precisión de los SFBR mediante la optimización del proceso de inferencia para cada regla fuzzy de la BR. Ese proceso generalmente se lleva a cabo a través de t-normas adaptativas, pero cuando se trata un problema de alta dimensión (varias variables y/o ejemplos) la optimización de los parámetros comienza a ser problemático. En este trabajo [Marquez et al., 2012] se propone un nuevo conector adaptativo de conjunción y un AEMO para realizar el proceso de aprendizaje en los problemas de alta dimensión. El método propuesto se comparó con un método de optimización de la t-norma eficiente y conocido. Los resultados obtenidos en los cinco problemas de regresión seleccionados confirman el buen desempeño del método propuesto en términos de eficiencia, simplicidad y compactación de los modelos fuzzy obtenidos.

5.3 Aprendizaje de la BC

Además de la optimización de los componentes de los SFBR, otra posibilidad es aprender toda o parte de la BC mediante los AEMO. Los trabajos en esta categoría pueden ser divididos en tres subcategorías: aprendizaje por selección de reglas, aprendizaje de la BR y aprendizaje simultáneo de los componentes de la BC.

En las siguientes subsecciones mostraremos algunos de los últimos trabajos realizados en esta categoría.

5.3.1 Aprendizaje por Selección de Reglas

Evolutionary multiobjective optimization for the design of fuzzy rule-based ensemble classifiers (2006)

En este trabajo [Ishibuchi and Nojima, 2006] se analiza la aplicación de los AEMO para buscar una gran cantidad de SCFBR no dominados a lo largo de la superficie de equilibrio entre precisión e interpretabilidad. La precisión de cada clasificador se basa en la tasa de clasificación de los conjuntos de entrenamiento y la complejidad se mide mediante el número de reglas fuzzy y el número total de condiciones antecedentes. Finalmente un unificador de clasificadores (también llamado como multclasificador) es diseñado para combinar los SCFBR no dominados. El desempeño es analizado por experimentos en seis conjuntos de datos de referencia del repositorio de aprendizaje automático de la Universidad de California en Irvine (UCI). Los autores observan que el efecto de combinar varios SCFBR depende del problema y que unificar clasificadores con alta diversidad usualmente tiene mejor desempeño.

Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning (2007)

Este trabajo [Ishibuchi and Nojima, 2007] examina el equilibrio entre precisión e interpretabilidad en SCFBR utilizando un algoritmo de aprendizaje genético multiobjetivo fuzzy. Este algoritmo es un híbrido de los enfoques Michigan y Pittsburgh basado en un AEMO. Cada regla fuzzy está representada por una cadena de enteros de tamaño fijo que representan los conjuntos fuzzy en el antecedente de la regla fuzzy. Cada SCFBR es un conjunto de reglas fuzzy, el cual se representa

5. REVISIÓN DE LA LITERATURA

como una cadena de números enteros de longitud variable. Este algoritmo maximiza simultáneamente la precisión del conjunto de reglas y minimiza su complejidad. La precisión se mide mediante la tasa de clasificación mientras que la complejidad se mide por el número de reglas fuzzy y/o el número total de antecedentes en las reglas fuzzy.

Multiobjective genetic fuzzy rule selection with fuzzy relational rules (2013)

La selección genética de reglas fuzzy ha sido utilizada con frecuencia para el diseño de SCFBR, y una cantidad de variantes han sido propuestas para esta tarea. En muchos trabajos referentes a la selección genética de reglas fuzzy, cada condición antecedente en una reglas fuzzy está definida por una variable, por ejemplo, " x_1 es pequeño" o " x_2 es grande". En este trabajo [Nojima and Ishibuchi, 2013] se analiza el uso de condiciones relacionales fuzzy con respecto a la relación entre dos variables de entrada, por ejemplo, " x_1 es aproximadamente igual a x_2 " o " x_3 es aproximadamente mayor que x_4 ". El algoritmo de selección de reglas fuzzy relacionales está basado en un AEMO.

Selection of rules by orthogonal transformations and genetic algorithms to improve the interpretability in fuzzy rule based systems (2013)

Una de las técnicas más conocidas para modelar sistemas y procesos son los modelos fuzzy. En la mayoría de los casos estos modelos fuzzy logran una alta precisión, pero muestran una baja complejidad o interpretabilidad, que es un aspecto clave de la lógica fuzzy. Existen varias técnicas y trabajos en la literatura para tratar el desafío de lograr una baja complejidad y alta interpretabilidad en los SFBR. En este trabajo [Rey et al., 2013] se propone un enfoque post-procesamiento a través de la selección genética de reglas fuzzy basada en la relevancia de cada regla fuzzy (utilizando transformaciones ortogonales) y lograr en buen equilibrio entre precisión e interpretabilidad. El objetivo principal es comprobar el significado verdadero, inconvenientes y ventajas de la selección de reglas basado en transformaciones ortogonales. SPEA2 es la técnica de optimización multiobjetivo para la selección de reglas fuzzy.

Improving a fuzzy association rule-based classification model by granularity learning based on heuristic measures over multiple granularities (2013)

Un proceso evolutivo multiobjetivo de selección de reglas fuzzy extrae un subconjunto de reglas fuzzy a partir de un conjunto inicial, mediante la aplicación de un AEMO. Dos enfoques pueden ser usados para determinar el número de términos lingüísticos (es decir, la granularidad) asociada a cada variable lingüística que aparecen en las reglas fuzzy: el primero es un enfoque de granularidad simple para todas las variables lingüísticas; y el segundo es un enfoque que considera diferentes granularidades para cada variable lingüística. El último enfoque favorece a la reducción del número de reglas fuzzy seleccionadas, pero también lleva a una posible pérdida de interpretabilidad. Para evitar ese problema, granulaciones adecuadas para cada variable lingüística pueden ser definidas mediante la aplicación de técnicas automáticas antes del proceso de generación de reglas fuzzy. En este trabajo [Fazzolari et al., 2013a] se investiga la aplicación del enfoque de granularidad simple. El objetivo es reducir la complejidad de los modelos obtenidos, tratando de mantener un buen desempeño del clasificador.

5.3.2 Aprendizaje de la Base de Reglas

A Pareto-based multi-objective evolutionary approach to the identification of Mamdani fuzzy systems (2007)

En este trabajo [Cococcioni et al., 2007] se propone un enfoque evolutivo multiobjetivo para generar un SFBR tipo Mamdani (problemas de regresión) a partir de datos numéricos. Los autores adaptan y definen una variante de la conocida técnica de optimización multiobjetivo (2+2)Pareto Archived Evolutionary Strategy ((2+2)PAES). (2+2)PAES determina una aproximación a la frontera de Pareto minimizando el ECM y la complejidad del SFBR. La complejidad es medida mediante la suma de las condiciones en los antecedentes de las reglas fuzzy en el SFBR. Por lo tanto, valores bajos de complejidad corresponden a SFBR tipo Mamdani conformados por una baja cantidad de reglas fuzzy y una baja cantidad de variables de entrada utilizadas en cada regla fuzzy. Esta versión de (2+2)PAES es analizada en tres problemas de regresión conocidos: horno de gas Box y Jenkins, las series de tiempo caóticas de Mackey-Glass y conjuntos de datos de series de tiempo

5. REVISIÓN DE LA LITERATURA

de Lorenz. Para mostrar el buen desempeño de este método los autores comparan las fronteras de Pareto producidos con los obtenidos mediante la aplicación de un enfoque heurístico basado en la descomposición SVD-QR (Singular Value Decomposition and QR algorithm) y cuatro AEMO diferentes.

Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets (2010)

Los autores utilizan un AEMO con tres objetivos para producir una aproximación a la frontera de Pareto compuesta de SCFBR para bases de datos no balanceadas con diferentes valores de equilibrio entre precisión (expresado en términos de sensibilidad y especificidad) y complejidad (definida por el número total de condiciones en los antecedentes de las reglas del clasificador). Después, se utiliza la curva ROC para seleccionar los clasificadores potencialmente óptimos dentro de la frontera de Pareto. Este método [Ducange et al., 2010a] fue probado en trece conjuntos de datos altamente no balanceados y es comparado con dos métodos evolutivos de dos objetivos, un enfoque heurístico para la generación de SCFBR y tres clasificadores conocidos. Los autores muestran por la técnica no paramétrico test del signo de rangos de Wilcoxon [Wilcoxon, 1945] [Mann and Whitney, 1947] que este método de optimización con tres objetivos tiene mejor desempeño que las otras técnicas, a excepción de un clasificador, en términos del área bajo la curva ROC. Además, todos los SCFBR generados se caracterizan por un bajo valor de complejidad. Finalmente, los autores discuten como se pueden determinar los costos de clasificación errónea y las distribuciones de las clases, utilizando ello para seleccionar el SCFBR más adecuado para un problema específico.

Multi-objective iterative genetic approach for learning fuzzy classification rules with semantic-based selection of the best rule (2013)

Este trabajo [Hinojosa Cardenas and Camargo, 2013] propone un AEMO basado en el enfoque IRL para la generación de reglas fuzzy de clasificación. El AEMO considera dos objetivos, la precisión medida por la compatibilidad de cada regla con los ejemplos de entrenamiento y la interpretabilidad medida por la cantidad de condiciones de la regla. En cada ejecución del AEMO se genera una conjunto de reglas no dominadas en ambos objetivos definidos, del cual solo una regla será seleccionada e insertada en la BR final. El criterio de selección está basado en la

interpretabilidad semántica de las reglas, es decir, una vez inserida en la BR final se determina cuantas reglas son lanzadas para clasificar la clase de la regla fuzzy. Los resultados muestran una mejora significativa en el nivel semántico de los SCFBR en comparación a una versión anterior del método.

5.3.3 Aprendizaje Simultáneo de los Componentes de la Base de Conocimiento

Learning concurrently data and rule bases of Mamdani fuzzy rule-based systems by exploiting a novel interpretability index (2011)

En este trabajo [Antonelli et al., 2011a] un AEMO se aplica para aprender un conjunto de SFBR tipo Mandani con diferentes valores de equilibrio entre precisión e interpretabilidad. La interpretabilidad en los SFBR tipo Mandani considera la integridad de las particiones fuzzy o conjuntos fuzzy para cada variable lingüística, para ello los autores presentan un nuevo índice para evaluar la interpretabilidad de los SFBR, que considera la complejidad de la BR y la integridad de la BD. También se presenta una discusión del índice en SFBR tipo Mandani mediante el aprendizaje simultáneo de la BR, la granularidad en las variables lingüísticas y los parámetros de las funciones de pertenencia en el proceso evolutivo. El método propuesto fue aplicado en seis problemas de regresión del mundo real y los resultados se han comparado con los obtenidos mediante la aplicación del mismo AEMO que considera la precisión y solo la complejidad de la BR como objetivos. Se demuestra que el método propuesto logra mejores valores de equilibrio entre interpretabilidad y precisión.

Learning knowledge bases of multi-objective evolutionary fuzzy systems by simultaneously optimizing accuracy, complexity and partition integrity (2011)

Este trabajo [Antonelli et al., 2011b] propone considerar la integridad de la particiones como un objetivo en el proceso evolutivo. Para ello, primero es discutido como la integridad de las particiones puede ser medida mediante el uso de un índice basado en la similaridad entre las particiones aprendidas durante el proceso evolutivo y las particiones iniciales interpretables definidas por un experto. Seguidamente, se presenta un AEMO de tres objetivos que aprende un conjunto de SFBR

5. REVISIÓN DE LA LITERATURA

tipo Mandani con diferentes valores de equilibrio entre complejidad, precisión e integridad de las particiones. Este proceso genera la BR y define los parámetros de los conjuntos fuzzy en la BD de forma simultánea. La precisión es medida mediante el ECM entre los valores reales y los valores previstos, la complejidad se calcula según el número total de condiciones en los antecedentes de las reglas y la integridad se mide utilizando el índice definido. Por último, se compara el método propuesto con métodos encontrados en la literatura que utilizan AEMO.

A fast and scalable multiobjective genetic fuzzy system for linguistic fuzzy modeling in high dimensional regression problems (2011)

Modelos fuzzy lingüístico en problemas de regresión de alta dimensionalidad plantea el reto de aprendizaje exponencial de reglas fuzzy cuando la cantidad de variables y/o ejemplos de entrenamiento es alta. Una forma de tratar este problema es mediante la determinación de las variables utilizadas, dividir la información lingüística y definir un conjunto reducido de reglas fuzzy, con el fin de obtener un modelo simple, pero preciso. Sin embargo, determinar estos componentes es una tarea difícil, porque además de tener un espacio de búsqueda grande, es un espacio de búsqueda complejo.

En este trabajo [Alcala et al., 2011] los autores proponen un AEMO efectivo basado en un aprendizaje genético embutido de la BC (que considera las variables utilizadas, granularidad y niveles de desplazamientos de las particiones fuzzy) que realiza un aprendizaje rápido de modelos fuzzy lingüísticos sencillos y precisos. Además, se realiza una etapa rápida de post-procesamiento para optimizar aún más las soluciones aprendidas. Este método es utilizado en diecisiete conjuntos de datos del mundo real con diferentes cantidades de variables y ejemplos. Los resultados confirman la eficacia del método propuesto no solo en términos de escalabilidad, sino en términos de sencillez y generalización de los modelos fuzzy obtenidos en comparación a tres modelos embutidos genéticos encontrados en la literatura.

Multi-objective design of highly interpretable fuzzy rule-based classifiers with semantic cointension (2011)

La evaluación de la interpretabilidad está generalmente dirigida a la evaluación de la complejidad y/o la legibilidad de los SFBR. Sin embargo, la comprensibilidad no suele ser considerada, ya que implica aspectos más cognitivos que son difíciles

de formalizar y de tratar. En este trabajo[Cannone et al., 2011] se muestra la importancia de considerar no solo la legibilidad sino también la comprensibilidad durante el proceso de diseño de SCFBR. Los autores proponen un nuevo índice para evaluar la comprensibilidad en el contexto de un marco evolutivo de tres objetivos para el diseño de SCFBR altamente interpretables, llamado de Índice de Vista Lógico (IVL) basado en la semántica. El algoritmo evolutivo propuesto realiza un aprendizaje simultáneo de la BR y BD utilizando la técnica de optimización multiobjetivo NSGA-II con los objetivos de maximizar la precisión, legibilidad y comprensibilidad de los SCFBR. El método propuesto es utilizado en un problema conocido de diagnóstico médico donde la interpretabilidad es un requisito importante, mostrando buenos resultados.

Studying the behavior of a multiobjective genetic algorithm to design fuzzy rule-based classification systems for imbalanced data-sets (2011)

En este trabajo [Villar et al., 2011] se estudia el comportamiento de un AGMO para realizar la selección de características y la granularidad en el aprendizaje de SCFBR seguida de un proceso de aprendizaje de reglas fuzzy rápido en el escenario de bases de datos no balanceadas. Los autores consideran dos medidas diferentes: una es la precisión y la otra es la complejidad del modelo, ambas medidas consideradas como objetivos a optimizar. El resultado de este método, basado en un AGMO, es un conjunto de soluciones con varios valores de equilibrio entre precisión e interpretabilidad. El análisis experimental, llevado a cabo en una amplia gama de bases de datos no balanceadas, muestra que este método es capaz de obtener un conjunto de modelos fuzzy con buen equilibrio entre precisión e interpretabilidad. Los autores también discuten del problema, aún en abierto, de como seleccionar la mejor solución dentro del conjunto de soluciones no dominadas obtenidas.

Obtaining accurate TSK Fuzzy Rule-Based Systems by Multi-Objective Evolutionary Learning in high-dimensional regression problems (2013)

Este trabajo [Gacto et al., 2013] aborda los problemas de alta dimensionalidad y regresión de gran escala en bases de datos. Para ello, se propone un método escalable de dos etapas para la obtención de modelos fuzzy precisos en problemas de regresión de alta dimensionalidad utilizando SFBR para problema de regresión tipo Takagi Sugeno Kang (TSK).

5. REVISIÓN DE LA LITERATURA

En la primera etapa, se propone un AEMO, en un proceso de aprendizaje simultáneo de los componentes de la BC. En la BD se definen variables involucradas, granularidad y un ligero desplazamiento lateral de particiones fuzzy; y la BR es definida mediante un proceso de aprendizaje inductivo. La segunda etapa es un proceso de post-procesamiento basado en un segundo AEMO para realizar selección de reglas fuzzy y optimización basada en la dispersión de las funciones de pertenencia. En ambas etapas, se incluyen un mecanismos para mejorar significativamente la precisión del modelo fuzzy y para asegurar una convergencia rápida en problemas de regresión de alta dimensión.

El método propuesto se compara con el método clásico ANFIS (Adaptive Network-based Fuzzy Inference System) y a un conocido AEMO para la obtención de sistemas de SFBR tipo TSK, en ocho conjuntos de datos con diferentes tamaños y dimensiones, obteniendo mejores resultados.

A two-stage multi-objective genetic-fuzzy mining algorithm (2013)

En este trabajo [Chen et al., 2013] se propone un método multiobjetivo de dos etapas para descubrir conocimiento lingüístico. En la primera etapa, un AGMO se utiliza para derivar un conjunto de funciones de pertenencia fuzzy no dominadas (soluciones de Pareto) con dos funciones objetivo (idoneidad y el número de variables) con un conjunto de reglas fuzzy asociadas a las funciones de pertenencia fuzzy. En la segunda etapa, una técnica de agrupación se utiliza para encontrar soluciones representativas de las soluciones de Pareto. Las soluciones representativas podrían ser empleadas para generar un conjunto de reglas fuzzy de acuerdo a los requisitos del sistema. Los experimentos en un conjunto de datos de simulación muestran la eficacia del método propuesto.

Evolutionary Fuzzy Classifiers for Imbalanced Datasets: An Experimental Comparison (2013)

En este trabajo [Antonelli et al., 2013] se realiza una comparación de tres clasificadores fuzzy evolutivos para bases de datos no balanceadas. Los dos primeros son mencionados en la siguiente subsección, un método realiza el aprendizaje de la BD con un aprendizaje de la BR embutida [VILLAR et al., 2012]; y el otro método genera una SCFBR jerárquico, donde un algoritmo basado en programación genética es utilizado para aprender la BR, después es realizado un post-procesamiento

que consiste en la selección genética de reglas fuzzy y optimización de los parámetros de los conjuntos fuzzy [López et al., 2013].

El tercer método es propuesto por los autores y es una extensión de un método evolutivo multiobjetivo para la generación de SCFBR propuesto en [Ducange et al., 2010b]. El AEMO genera concurrentemente la BR y los parámetros de los conjuntos fuzzy en la BD. El resultado es un conjunto de clasificadores con diferentes valores de equilibrio entre precisión e interpretabilidad.

En este método un cromosoma C que representa una BC completa es compuesto por dos partes: C_{BR} y C_{BD} que representan la BR y los parámetros de los conjuntos fuzzy o BD respectivamente. La Figura 5.5 muestra el vector de números enteros que codifica la parte C_{BR} del cromosoma, cada regla fuzzy R_m contiene los índices $j_{m,f}$ del antecedente, cada variable de entrada X_f y la clase consecuente C_{j_m} .

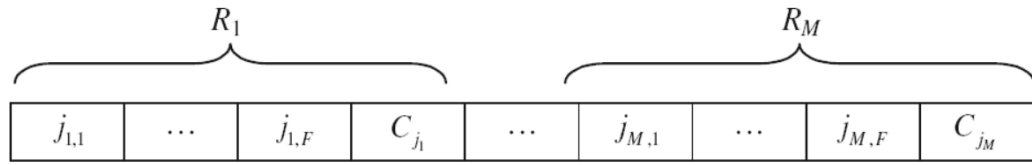


Figura 5.5: Parte C_{BR} del cromosoma - [Antonelli et al., 2013]

La Figura 5.6 muestra la representación de la parte C_{BD} la cual contiene F vectores de números reales, el f^{th} vector contiene los $[b_{f,2}, \dots, b_{f,T_f-1}]$ o puntos medios de cada conjunto fuzzy triangular para cada variable lingüística X_f .

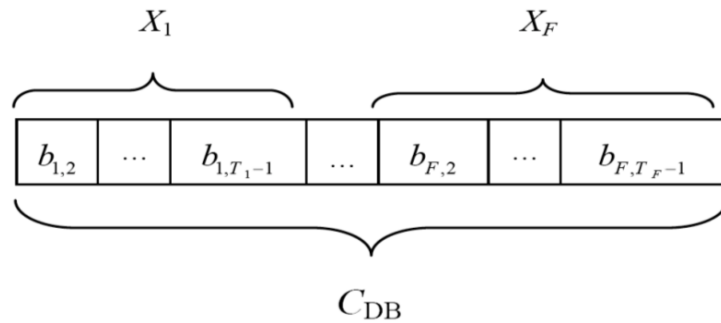


Figura 5.6: Parte C_{BD} del cromosoma - [Antonelli et al., 2013]

5. REVISIÓN DE LA LITERATURA

La técnica de optimización multiobjetivo que utilizan los autores es el (2+2) Memetic Pareto Archived Evolution Strategy ((2+2)M-PAES) intentando maximizar la precisión en base a la proporción de TP y la proporción de FP; y maximizar la interpretabilidad mediante la reducción de condiciones en las reglas fuzzy de la BR. Este método es llamado PAES-3obj y es uno de los métodos que utilizamos para comparar el método propuesto en este trabajo en el Capítulo 6.

A fast and efficient multi-objective evolutionary learning scheme for fuzzy rule-based classifiers (2014)

Durante los últimos años, los AEMO han sido utilizados para generar SFBR caracterizados por diferentes valores de equilibrio entre precisión y complejidad. En este trabajo [Antonelli et al., 2014b] se propone un enfoque basado en un AEMO para aprender de forma simultánea la BR y la BD en un SCFBR. Específicamente, la BR se genera mediante una estrategia de selección de reglas y condiciones, la cual selecciona un número reducido de reglas de un conjunto de reglas candidatas generadas heurísticamente y un número reducido de condiciones para cada regla seleccionada durante el proceso evolutivo multiobjetivo. En cuanto al aprendizaje de la BD, los parámetros de las funciones de pertenencia de cada término lingüístico utilizado en las reglas fuzzy se aprenden simultáneamente al proceso de selección. Los autores analizaron este método en veinticuatro problemas de clasificación y compararon sus resultados con los obtenidos con dos métodos basados en AEMO encontrados en la literatura. Mediante el uso de una técnica de comparación estadística no paramétrica, se demuestra que este método genera SCFBR con precisión y complejidad estadísticamente comparables, y a veces mejor que los SCFBR generados por los dos métodos mencionados anteriormente.

5.4 Trabajos que utilizan SFG para tratar el problema de Bases de Datos no Balanceadas

A first approach for cost-sensitive classification with linguistic Genetic Fuzzy Systems in imbalanced data-sets (2010)

Aquí los autores hacen énfasis en que la clasificación en dominios desequilibrados es uno de los problemas más relevantes dentro del área de aprendizaje de

5.4 Trabajos que utilizan SFG para tratar el problema de Bases de Datos no Balanceadas

máquina. Este problema ha crecido en importancia debido a que se encuentra en muchas aplicaciones reales y ocurre cuando la distribución de los ejemplos, disponibles para llevar a cabo el proceso de aprendizaje, es muy diferente entre las clases (frecuentemente en conjuntos de datos binarios). Por lo general, la clase minoritaria es de mayor interés para el problema, por lo que el costo derivado de una clasificación errónea de la clase minoritaria es mucho más alto que el costo de la clasificación errónea de la clase mayoritaria. En este trabajo [López et al., 2010] se analiza el comportamiento de un método de aprendizaje, sensible al costo, basada en un SCFBR en el escenario de bases de datos altamente no balanceadas. Los resultados experimentales muestran que este método sensible al costo, en este tipo de escenarios desbalanceados, ayudan a obtener soluciones precisas en tiempos más cortos de entrenamiento y también con una menor complejidad con respecto a otras posibilidades propuestas para la clasificación en bases de datos no balanceadas, así como el uso de pre-procesamiento para balancear la distribución de clases.

Preprocessing vague imbalanced datasets and its use in genetic fuzzy classifiers (2010)

Cuando existe una diferencia sustancial entre el número de ejemplos de la clase mayoritaria y minoritaria los sistemas de clasificación basados en el error mínimo tienden a pasar por alto los ejemplos de la clase minoritaria. Este problema se puede tratar por el pre-procesamiento del conjunto de datos o mediante la alteración de la función objetivo del clasificador. En este trabajo [Palacios et al., 2010] se analiza el pre-procesamiento del conjunto de datos, en el contexto de los SFG. Los autores analizan los diferentes mecanismos de pre-procesamiento de datos desequilibrados y muestran la necesidad de extender esos mecanismos para resolver el problema de no balanceamiento e imprecisión de la información. Además, los autores realizan una descripción de un nuevo algoritmo capaz de pre-procesar bases de datos imprecisas no balanceadas. Varios conjuntos de datos del mundo real se utilizan para evaluar esa propuesta.

On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets (2010)

Al realizar una tarea de clasificación, este trabajo indica que podemos encontrar algunas bases de datos con una distribución diferente en la cantidad de ejemplos de

5. REVISIÓN DE LA LITERATURA

las clases. Este problema se conoce como clasificación en bases de datos no balanceadas y aparece en muchas áreas de aplicación del mundo real. Por esta razón, es un tema relevante en el área de aprendizaje automático. El objetivo de este trabajo [Fernández et al., 2010] es mejorar el desempeño de un SCFBR en el contexto de bases de datos no balanceadas mediante un proceso de optimización. Específicamente, los autores realizan un proceso de optimización de conjuntos fuzzy en 2 tuplas (desplazamiento de los conjuntos fuzzy con limitaciones). Los resultados empíricos muestran que la optimización genética en 2-tuplas aumenta el desempeño del SCFBR en todas las bases de datos no balanceadas utilizadas.

Genetic fuzzy system for intrusion detection: Analysis of improving of multi-class classification accuracy using KDDCup-99 imbalance dataset (2012)

Los SFG juntan el método de raciocinio aproximado de los sistemas fuzzy con la capacidad de aprendizaje de los AE (específicamente los AG). El objetivo de este trabajo [Gaffer et al., 2012] es centrarse en un problema de clasificación dentro de la detección de intrusiones en un red, un problema que tiene una distribución de clases no balanceadas. Los autores proponen el uso de una nueva función de aptitud para tratar el problema de bases de datos no balanceadas en el post-procesamiento genético para la selección de reglas fuzzy y optimización de los parámetros de las funciones de pertenencia de la BD para un SCFBR. La eficiencia del método propuesto se ha demostrado a través de un estudio comparativo con clasificadores conocidos encontrados en la literatura. Los experimentos se realizaron con la base de datos usada en la Competición Internacional de Datos 99 o Data Mining and Knowledge Discovery Cup 99 (KDDCup-99), una base de datos no balanceada de referencia en el problema de detección de intrusiones en un red.

A Genetic Algorithm for Feature Selection and Granularity Learning in Fuzzy Rule-Based Classification Systems for Highly Imbalanced Data-Sets (2012)

En muchos problemas de clasificación, un gran número de características o variables lingüísticas puede originar BR con un gran número de reglas fuzzy, lo que puede deteriorar la interpretabilidad y posible sobre-entrenamiento (el error sobre el conjunto de entrenamiento es bajo, pero el SCFBR no presenta buen desempeño en el conjunto de test o predicción). Ese problema puede ser tratado mediante dos

5.4 Trabajos que utilizan SFG para tratar el problema de Bases de Datos no Balanceadas

enfoques: la reducción del conjunto de reglas fuzzy o la reducción de características. Tener en cuenta que en problemas de alta dimensionalidad y problemas donde un gran conjunto de ejemplos son analizados es difícil reducir el número de reglas fuzzy, además la comprensión e interpretabilidad del modelo fuzzy obtenido podría no ser tan bueno como se esperaba. Para problemas de clasificación con alta dimensionalidad, el enfoque de selección de características, que determina las características o variables lingüísticas más relevantes antes o durante el proceso de generación del SCFBR, podría ser considerado. Ello incrementa la eficiencia y eficacia en el proceso de aprendizaje y clasificación.

El número de términos lingüísticos o conjuntos fuzzy para cada variable lingüística (granularidad) es información que no ha sido considerada en la mayoría de métodos propuestos para la generación de SCFBR. Sin embargo, la granularidad puede tener una influencia significativa en el desempeño del SCFBR. El objetivo de este trabajo [VILLAR et al., 2012] es proponer un método de generación genética para mejorar la habilidad de predicción de los SCFBR para bases de datos no balanceadas con una reducción significativa de la complejidad. Esta propuesta utiliza un AG para realizar la selección de características y definir la granularidad, considerando un método clásico de aprendizaje de reglas fuzzy propuesto en [Chi et al., 1996]. Con el fin de manejar el problema de desbalanceamiento de los datos, los autores utilizan la técnica SMOTE para balancear los datos en la etapa de pre-procesamiento de los datos. Este método es llamado GA-FS-GL (Genetic Algorithm for Feature Selection and Granularity Learning) y es uno de los métodos que utilizamos para comparar el método propuesto en este trabajo en el Capítulo 6.

Designing fuzzy imbalanced classifier based on the subtractive clustering and Genetic Programming (2013)

En este trabajo [Mahdizadeh and Eftekhari, 2013] se propone una metodología para aprender un SFBR para bases de datos no balanceadas. El clasificador está basado en el MI de los SFBR tipo TSK. La BR se aprende mediante el uso de la técnica de agrupamiento por sustracción y programación genética multigen. El agrupamiento por sustracción se utiliza para producir los antecedentes de las reglas fuzzy y la programación genética multigen se utiliza para la generación de las funciones de pertenencia en los consecuentes de las reglas fuzzy. La selección

5. REVISIÓN DE LA LITERATURA

de características que se utiliza es un paso importante de pre-procesamiento para reducir la dimensión. Los experimentos se realizan en ocho conjuntos de datos KEEL (Knowledge Extraction based on Evolutionary Learning). Los resultados de la comparación revelan que el clasificador propuesto supera a otros métodos encontrados en la literatura.

Addressing covariate shift for Genetic Fuzzy Systems classifiers: A case of study with FARC-HD for imbalanced datasets (2013)

La estimación de la calidad de los modelos aprendidos en minería de datos se ha realizado tradicionalmente por medio de una técnica de partición de k divisiones. Sin embargo, la división "aleatoria" de los ejemplos puede resultar en un problema llamado de cambio de covarianza, es decir, hay una distribución de datos diferente en las divisiones de entreamiento y prueba. En la clasificación de bases de datos no balanceadas este problema es más severo. Para evitar este problema, este trabajo [Lopez et al., 2013] propone el uso de una técnica de validación específica para la división de datos, conocida como "Validación Cruzada basada en una distribución óptima, equilibrada y estratificada" (Distribution optimally balanced stratified cross-validation). Esta técnica intenta que cada región del espacio de búsqueda esté representada en cada división. Los autores muestran los beneficios de esta técnica usando un SFG, específicamente, el algoritmo FARC-HD (Fuzzy Association Rule Classification model for High-Dimensional problems).

Evolutionary fuzzy classifiers for imbalanced datasets: An experimental comparison (2013)

En este trabajo [Antonelli et al., 2013] se compara tres Clasificadores Fuzzy Evolutivos (CFE) para bases de datos no balanceadas. El primer CFE realiza el aprendizaje evolutivo simultáneo de la BR y la BD. El segundo CFE aprende un SCFBR mediante un proceso jerárquico: primero, un algoritmo de programación genética se utiliza para aprender la BR y luego un post-procesamiento, que incluye la selección genética de reglas fuzzy y la optimización de los parámetros de las funciones de pertenencia de los conjuntos fuzzy. El tercer CFE es una extensión de un método multiobjetivo de aprendizaje que los autores proponen: la BR y los parámetros de las funciones de pertenencia de los conjuntos fuzzy en un SCFBR son simultáneamente aprendidos mediante la optimización de la sensibilidad, la

5.4 Trabajos que utilizan SFG para tratar el problema de Bases de Datos no Balanceadas

especificidad y la complejidad. Mediante la realización de pruebas estadísticas no paramétricas, se muestra que, sin realizar un proceso de balanceamiento de los ejemplos de entrenamiento, el tercer CFE es mejor que los otros dos métodos en términos del área bajo la curva ROC.

A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets (2013)

En este trabajo [López et al., 2013] los autores proponen un método para obtener un SCFBR jerárquico para clasificar bases de datos no balanceadas. Un SCFBR jerárquico indica que cada regla fuzzy puede ser generada en varios niveles en vez de generar reglas fuzzy en un solo nivel, cada nivel usa un número fijo de términos lingüísticos para cada variable lingüística, por ejemplo, si las reglas fuzzy en el primer nivel tendrían dos términos lingüísticos en cada variable lingüística; en el segundo nivel cada término lingüístico del primer nivel es dividido en dos, por lo cual cada variable lingüística tendría cuatro términos lingüísticos; y así sucesivamente hasta un nivel N .

El método propuesto sigue los siguientes pasos:

1. El pre-procesamiento de los datos, basado en el método SMOTE para balancear los datos o ejemplos de entrenamiento.
2. La generación de las BC jerárquica fuzzy usando el método basado en programación genética propuesto en [Berlanga et al., 2010].
3. El post-procesamiento que incluye la selección de reglas fuzzy y la aplicación de una optimización genética para mejorar el desempeño general del SCFBR.

La combinación de esos pasos constituyen un método para resolver el problema de clasificación en bases de datos no balanceadas. Uno de los problemas de este método es que genera reglas fuzzy del tipo DNF (Disjunctive Normal Form) así como reglas fuzzy en los diferentes niveles con pesos para cada una de ellas, lo cual tienden a deteriorar la interpretabilidad del modelo fuzzy. Este método es llamado GP-COACH-H (Genetic Programming-based learning of COmpact and ACcurate

5. REVISIÓN DE LA LITERATURA

fuzzy rule-based classification systems for High-dimensional problems Hierarchical) y es uno de los métodos que utilizamos para comparar el método propuesto en este trabajo en el Capítulo 6.

An experimental study on evolutionary fuzzy classifiers designed for managing imbalanced datasets (2014)

En este trabajo [Antonelli et al., 2014a] los autores realizan un estudio experimental en un conjunto de CFE diseñados para tratar el problema de bases de datos no balanceadas. Tres de estos CFE representan métodos encontrados en la literatura. El cuarto CFE es una extensión de un método de aprendizaje multiobjetivo propuesto por los autores. Mediante el uso de pruebas no paramétricas son comparados los cuatro métodos, los resultados obtenidos muestran que el cuarto método, en términos del área bajo la curva ROC, supera a los dos primeros métodos y resulta ser estadísticamente equivalente al tercer método. Además, los SCFBR generados por el método propuesto se caracteriza por un menor número de reglas fuzzy que los otros métodos.

Cost-Sensitive Learning of Fuzzy Rules for Imbalanced Classification Problems Using FURIA (2014)

Este trabajo [Palacios et al., 2014] tiene por objetivo analizar si el aprendizaje sensible al costo es un enfoque competitivo para el aprendizaje de reglas fuzzy en problemas de clasificación en bases de datos no balanceadas. Los autores muestran que existe una matriz de costos que junto a un clasificador adecuado mejoran los resultados obtenidos por las técnicas más populares a nivel de datos. El algoritmo FURIA (Fuzzy Unordered Rule Induction Algorithm) [Hühn and Hüllermeier, 2009] se extiende para aprovechar esta definición. Un estudio numérico, en sesenta y cuatro bases de datos no balanceadas, se lleva a cabo para comparar el método propuesto basado en el algoritmo FURIA con otros algoritmos encontrados en la literatura, basados en reglas fuzzy y en otros métodos de aprendizaje de máquina.

Sistemas de Clasificación Basados en Reglas Difusas para Problemas no Balanceados. Aproximaciones y Uso de Nuevas Estrategias para Resolver Problemas Intrínsecos a los Datos no Balanceados (2014)

5.4 Trabajos que utilizan SFG para tratar el problema de Bases de Datos no Balanceadas

Este trabajo [Morales, 2014] es una tesis doctoral de la Univesidad de Granada (España) cuyo principal interés reside en el estudio de los problema de clasificación con bases de datos no balanceadas bajo la perspectiva de las características internas que presentan este tipo de problemas. Analizando todos los métodos propuestos, esta tesis desarrolla nuevos métodos de aprendizaje para SCFBR que permiten abordar las características internas. Por último se pretende extender este estudio de la clasificación en bases de datos no balanceadas al campo de big data. El estudio experimental demostró que el esquema propuesto tiene un importante impacto en el rendimiento de los clasificadores ya que son incluidos varios enfoques que representan diferentes paradigmas de aprendizaje automático.

A Proposal for Evolutionary Fuzzy Systems using Feature Weighting: Dealing with Overlapping in Imbalanced Datasets (2015)

En un contexto general de clasificación, uno de los principales inconvenientes para mejorar la precisión de los modelos es la alta superposición entre los conceptos a aprender. Este inconveniente se agrava cuando estamos tratando bases de datos no balanceadas, donde el error al identificar los ejemplos de la clase minoritaria suele estar relacionado con las áreas de frontera donde se pueden encontrar ejemplos de las clases mayoritaria y minoritaria superpuestos. Este trabajo [Alshomrani et al., 2015] propone un método de ponderación de características o variables lingüísticas que tiene como objetivo analizar la importancia de las variables del problema ponderando el grado de pertenencia dentro del proceso de inferencia. Esto se hace mediante la aplicación de un peso diferente para las variables que representan un conjunto de datos, lo que permite suavizar o destacar las áreas de frontera entre las clases mayoritaria y minoritaria. Estos pesos son aprendidos por medio de un proceso de optimización utilizando algoritmos evolutivos en un SCFBR. Los resultados experimentales, que utilizan un gran número de problemas con diferentes grados de desbalanceamiento y superposición, muestran el buen desempeño de este método.

Cost-Sensitive Linguistic Fuzzy Rule Based Classification Systems under the MapReduce Framework for Imbalanced Big Data (2015)

Clasificación en grandes volúmenes de datos se ha convertido en una tendencia en el aprendizaje de máquina. El crecimiento de datos en los últimos años ha despertado el interés en métodos de adquisición de conocimiento para analizar y prede-

5. REVISIÓN DE LA LITERATURA

cir las nuevas tendencias. La variedad y la veracidad que se relacionan a los grandes conjuntos de datos introducen un grado de incertidumbre ha manejar, además del volumen y velocidad. Estos datos por lo general también presenta el problema de bases de datos no balanceadas. Con el objetivo de tratar adecuadamente grandes volúmenes de datos no balanceados, este trabajo [López et al., 2015] propone el algoritmo Chi-FRBCS-BigDataCS, que permite aprender un SCFBR capaz manejar la incertidumbre que se introduce en los grandes volúmenes de datos sin perder de vista el aprendizaje de la clase minoritaria. Se utilizaron veinticuatro bases de datos no balanceadas de gran volumen para probar el desempeño del método propuesto, obteniendo SCFBR competitivos tanto en la precisión del modelo como en el tiempo computacional.

5.5 Consideraciones Finales

En este capítulo fueron explicados resumidamente algunos de los últimos trabajos relacionados sobre la generación genética de sistemas fuzzy. Dentro de ellos también los utilizados para resolver el problema de bases de datos no balanceadas tratando uno o múltiples objetivos.

Se detallaron los trabajos según la taxonomía de dos niveles de los SFEM, la primera categoría de acuerdo a la naturaleza de los objetivos y la segunda categoría de acuerdo a los componentes a optimizar.

En el Capítulo 6 mostramos el método propuesto en este trabajo que permite la generación genética multiobjetivo de bases de conocimiento fuzzy usando el enfoque iterativo para el aprendizaje de reglas fuzzy, y que tiene como base varios trabajos mencionados en este capítulo.

Método Propuesto: MO-FRBCS-IRL-ID

Este capítulo presenta el método propuesto de generación genética multiobjetivo de bases de conocimiento fuzzy para clasificación en bases de datos no balanceadas usando el enfoque iterativo (Multi-Objective Fuzzy Rule-Based Classification Systems based on Iterative Rule Learning approach for Imbalanced Datasets (MO-FRBCS-IRL-ID)). Este método está basado en tres etapas, como muestra la Figura 6.1. A continuación son listadas y detalladas cada una de estas etapas.

1. Pre-procesamiento o balanceamiento de la base de datos no balanceada.
2. Aprendizaje genético multiobjetivo iterativo de reglas fuzzy de clasificación sobre una base de datos balanceada.
3. Optimización genética multiobjetivo de la BD o conjuntos fuzzy.

6. MÉTODO PROPUESTO: MO-FRBCS-IRL-ID

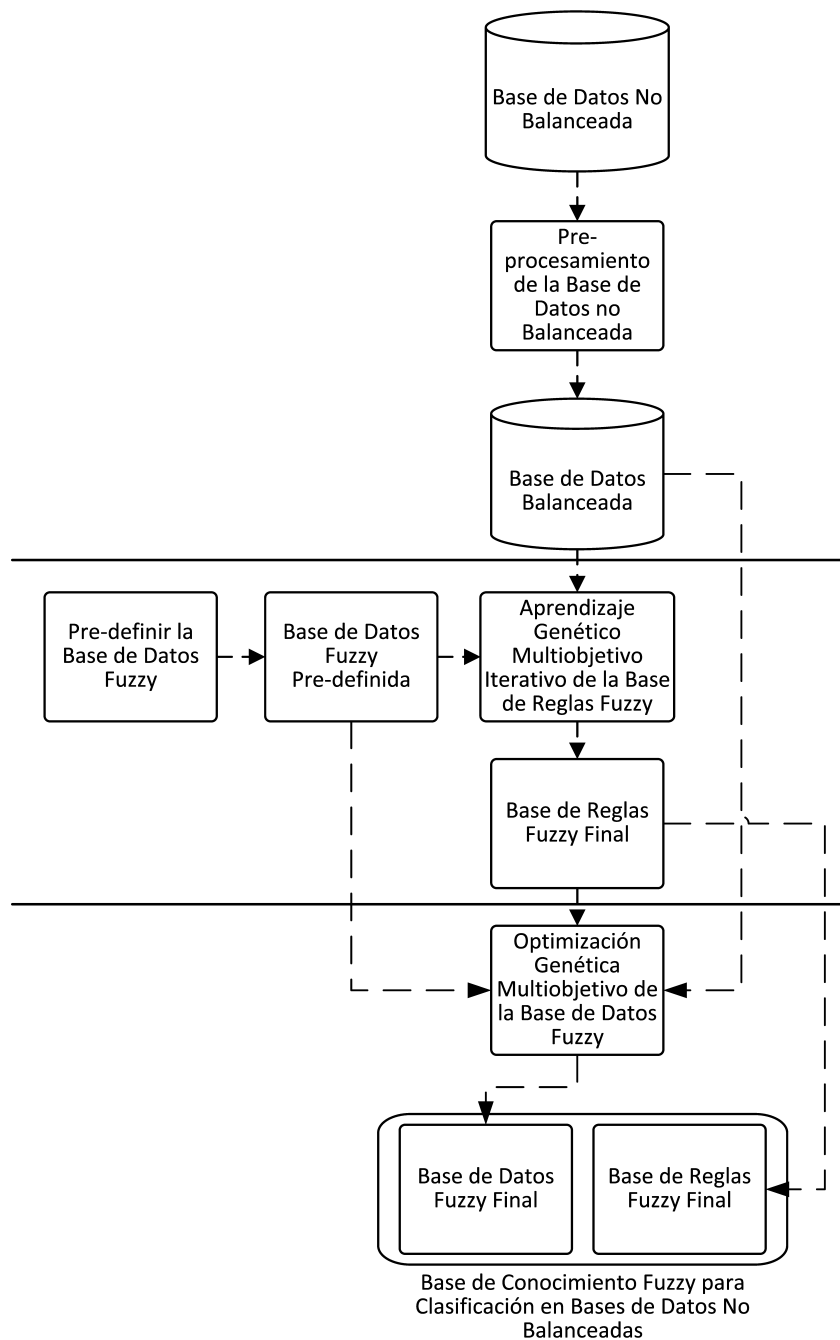


Figura 6.1: Método propuesto de generación genética multiobjetivo de bases de conocimiento fuzzy para clasificación en bases de datos no balanceadas usando el enfoque iterativo.

6.1 Pre-procesamiento de la Base de Datos no Balanceada

En esta etapa es utilizada una de las técnicas de sobremuestreo, submuestreo o híbrido (detalladas en el Capítulo 2) para balancear la cantidad de ejemplos negativos y positivos en la base de datos de entrenamiento. El resultado de esta etapa es la base de datos balanceada según la técnica seleccionada.

6.2 Aprendizaje Genético Multiobjetivo Iterativo de la Base de Reglas Fuzzy para Clasificación

Esta etapa utiliza una BD uniformemente distribuida, la técnica de optimización multiobjetivo NSGA-II (detallada en el Capítulo 3) y el enfoque IRL (detallado en el Capítulo 4), sobre la base de datos balanceada obtenida de la etapa anterior. El resultado de esta etapa es la BR final que contiene el conjunto de reglas fuzzy para clasificación. Para llegar a ese resultado, seguimos los pasos mostrados en la Figura 6.2.

A continuación detallamos el proceso de como pre-definir la BD fuzzy, los componentes utilizados en la técnica NSGA-II y el proceso de marcación de ejemplos para la generación de la BR fuzzy:

6.2.1 Pre-definir la Base de Datos Fuzzy

En este trabajo utilizamos conjuntos fuzzy triangulares detallados en el Capítulo 4. Para cada variable lingüística se define el valor mínimo (min) y el valor máximo (max) encontrados en la base de datos balanceada. Después, el espacio de valores de cada variable lingüística es dividido dependiendo la cantidad de conjuntos fuzzy o términos lingüísticos a definir (entre tres y nueve), el mismo para todas las variables lingüísticas. Cada punto de división es el punto medio de cada conjunto fuzzy (m), y punto izquierdo (a) o punto derecho (b) del conjunto fuzzy adyacente. Esta forma de definir los conjuntos fuzzy es conocida como conjuntos fuzzy uniformemente distribuidos, la cual es una manera simple y rápida de generar bases de datos

6. MÉTODO PROPUESTO: MO-FRBCS-IRL-ID

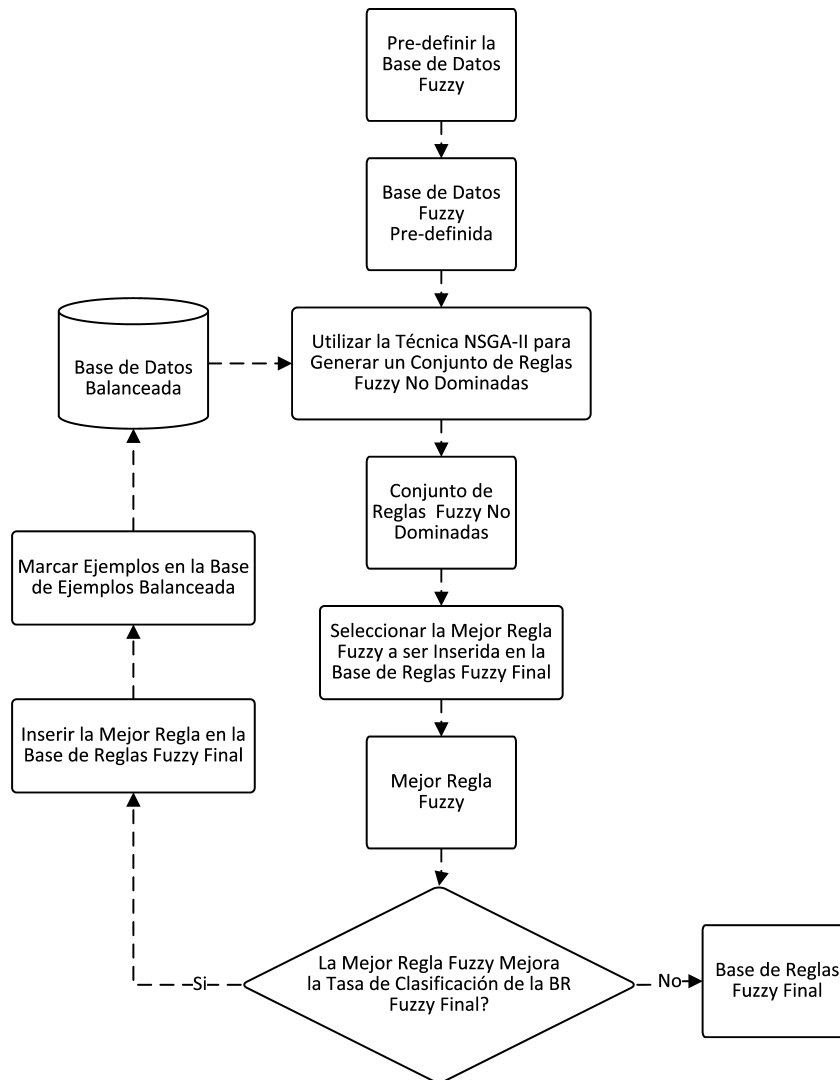


Figura 6.2: Etapa de aprendizaje genético multiobjetivo iterativo de la BR fuzzy.

fuzzy con alta interpretabilidad. Todos los conjuntos fuzzy deben estar normalizados, es decir, el grado de pertenencia del conjunto en el punto medio es uno. Un ejemplo, para cinco conjunto fuzzy, es mostrado en la Figura 6.3.

6.2.2 Representación de las Reglas Fuzzy

En la mayoría de trabajos que utilizan AG o AGMO los cromosomas o individuos, en este caso reglas fuzzy, son representados por cadenas de binarios o cadenas

6.2 Aprendizaje Genético Multiobjetivo Iterativo de la Base de Reglas Fuzzy para Clasificación

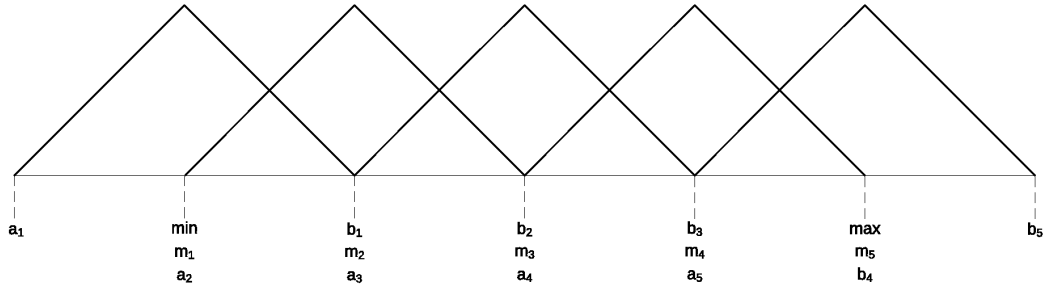


Figura 6.3: Ejemplo de base de datos fuzzy uniformemente distribuida.

de números. En el método propuesto en este trabajo, como en otros trabajos relacionados [Abadeh and Habibi, 2007] [Gonzalez and Perez, 2009] [Hinojosa and Camargo, 2011], cada regla fuzzy es representada por una cadena de números enteros.

El tamaño de cada cromosoma, o cantidad de números enteros o genes, depende de la cantidad de variables lingüísticas ordenadas que considera el problema, adicionando un gen para la clase positiva o negativa. Por ejemplo, en el caso que el problema considere cinco variables lingüísticas, el tamaño de la regla fuzzy sería de seis como muestra la Figura 6.4.

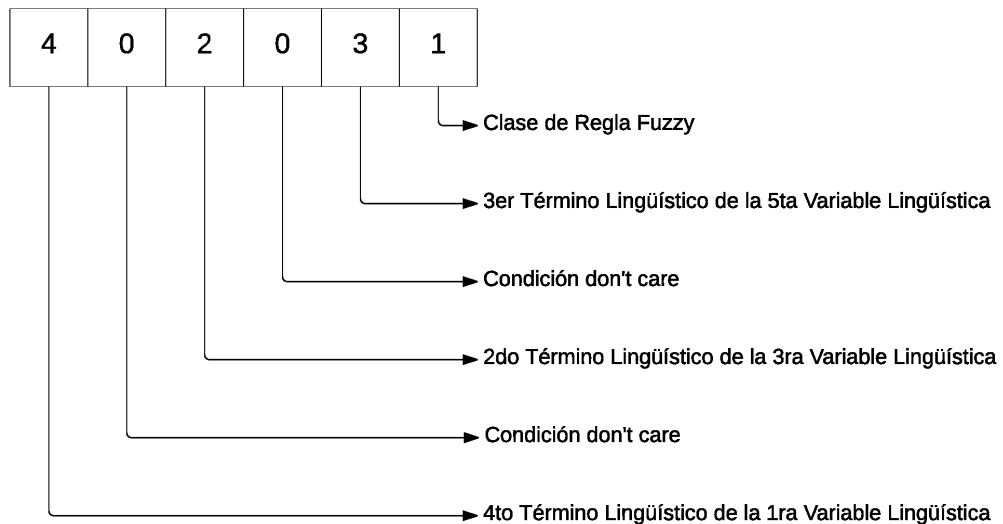


Figura 6.4: Representación de las Reglas Fuzzy.

6. MÉTODO PROPUESTO: MO-FRBCS-IRL-ID

Cada número entero o gen del cromosoma representa el índice u orden del término lingüístico dentro de una variable lingüística. El primer gen representa el término lingüístico de la primera variable lingüística, el segundo gen representa el término lingüístico de la segunda variable lingüística, y así sucesivamente. El valor máximo de cada gen depende de la cantidad de términos lingüísticos en la variable lingüística. El valor mínimo de cada gen es cero, en el caso el valor sea cero la variable lingüística no es considerada en la regla fuzzy (condición don't care). El último gen representa la clase de la regla fuzzy.

6.2.3 Población Inicial en la Generación de Reglas Fuzzy

La generación de la población inicial de reglas fuzzy es de forma aleatoria, considerando las restricciones de la cantidad de variables lingüísticas, términos lingüísticos y la cantidad de clases para el último gen.

6.2.4 Función de Aptitud del Objetivo de Precisión en la Generación de Reglas Fuzzy

Este es un tema crucial en la generación de la BR fuzzy, ya que nos indica que reglas fuzzy son seleccionadas para evolucionar y convergir para las soluciones cercanas a la(s) óptima(s).

La función de aptitud es calculada para cada regla fuzzy dentro de la población R_t en todas las iteraciones de la técnica NSGA-II. Dicha función está basada en dos criterios: 1) la integridad de la regla fuzzy, y 2) la consistencia de la regla fuzzy [Gonzalez and Perez, 1999].

En este trabajo, consideramos la integridad como la suma de los grados de compatibilidad de la regla fuzzy con todos los ejemplos de entrenamiento de la misma clase que la regla fuzzy, multiplicado por la cantidad de ejemplos no marcados de la misma clase de la regla fuzzy; menos, la suma de los grados de compatibilidad de la regla fuzzy con todos los ejemplos de entrenamiento de otra clase que la regla fuzzy, multiplicado por la cantidad de ejemplos de otra clase que la regla fuzzy. Formalmente, definimos la Integridad (I) por la siguiente ecuación:

6.2 Aprendizaje Genético Multiobjetivo Iterativo de la Base de Reglas Fuzzy para Clasificación

$$I(R_c) = \left[1 + \sum_{e=1}^n g(e_c \notin lm) \right] \times |\forall e_c \notin lm| - \left[1 + \sum_{e=1}^n g(e_{\neq c}) \right] \times |\forall e_{\neq c}| \quad (6.1)$$

Donde:

- R_c : Regla fuzzy con consecuente o clase c .
- g : Función que define el grado de compatibilidad de un ejemplo con la regla fuzzy.
- e_c : Ejemplo que pertenece a la clase c .
- $e_{\neq c}$: Ejemplo que no pertenece a la clase c .
- lm : Lista de ejemplos marcados.

La consistencia de una regla fuzzy es definida como la cantidad de nuevos ejemplos marcados con la misma clase de la regla fuzzy, entre, la cantidad de ejemplos marcados con diferente clase de la regla fuzzy. La marcación de un ejemplo se da cuando la regla fuzzy tiene un grado de compatibilidad alto (mayor a un valor gm) con el ejemplo. Formalmente, definimos la Consistencia (Co) por la siguiente ecuación:

$$Co(R_c) = \frac{1 + |\forall g(e_c \notin lm) \geq gm|}{1 + |\forall g(e_{\neq c}) \geq gm|} \quad (6.2)$$

Donde:

- R_c : Regla con consecuente o clase c .
- g : Función que define el grado de compatibilidad de un ejemplo con la regla fuzzy.
- e_c : Ejemplo que pertenece a la clase c .
- $e_{\neq c}$: Ejemplo que no pertenece a la clase c .
- lm : Lista de ejemplos marcados.
- gm : Grado o valor de compatibilidad que define si un ejemplo es marcado o no.

6. MÉTODO PROPUESTO: MO-FRBCS-IRL-ID

El objetivo de precisión es un objetivo de minimización. La función de aptitud para este objetivo (FA_{Prec}) de una regla R es formalmente definido por la siguiente ecuación:

$$FA_{Prec}(R) = \text{Minimizar} - (I \times Co) \quad (6.3)$$

6.2.5 Función de Aptitud del Objetivo de Interpretabilidad en el Aprendizaje de Reglas Fuzzy

Así como la función de aptitud del objetivo de precisión, este objetivo es crucial en la etapa de aprendizaje de la BR fuzzy.

La función de aptitud para la interpretabilidad es definida por la cantidad de condiciones del cromosoma o regla fuzzy, en otras palabras la cantidad de condiciones diferentes a condiciones don't care en la regla fuzzy o genes diferentes a cero en el cromosoma, obviamente sin considerar el último gen que representa la clase de la regla fuzzy. Este también es un objetivo a minimizar, dado que entre menor sea la cantidad de condiciones de la regla fuzzy, la misma será más interpretable. Formalmente, definimos este objetivo (FA_{Inte}) de una regla fuzzy R por la siguiente ecuación:

$$FA_{Inte}(R) = \text{Minimizar} |\forall cond \neq don't care| \quad (6.4)$$

Donde:

- R : Regla fuzzy.
- $cond$: Condición de la regla fuzzy R .
- $don't care$: Condición don't care.

En este punto es importante señalar que en caso una regla fuzzy tenga todas sus condiciones don't care, por resultado de la generación en la población inicial, cruzamiento y/o mutación; dicha regla fuzzy tendrá valores altos (o infinitos) para ambos objetivos.

6.2.6 Operadores Genéticos en la Generación de Reglas Fuzzy

A continuación mostramos los operadores genéticos utilizados en la técnica NSGA-II para la generación genética multiobjetivo de la BR fuzzy:

1. **Operador de Selección:** El operador utilizado para la selección es la selección por torneo de dos soluciones, detallado en el Capítulo 3. Antes del cruzamiento, el siguiente proceso es repetido dos veces, cada uno para seleccionar una regla fuzzy padre. Son seleccionadas dos reglas fuzzy de forma aleatoria, y comparadas según su dominancia, siendo seleccionada la regla fuzzy dominante; en caso ninguna solución sea dominante, se comparan las distancias de multitud, siendo seleccionada la regla fuzzy con mayor distancia de multitud; en caso las distancias de multitud sean iguales, una de las dos reglas fuzzy es seleccionada de forma aleatoria.
2. **Operador de Cruzamiento:** Una vez obtenidas las dos reglas fuzzy padres mediante el operador de selección, según una probabilidad $ProbCruzaBR$ es utilizado el cruzamiento SBX, detallado en el Capítulo 3, para generar dos nuevas reglas fuzzy o soluciones. En este caso consideramos las siguientes restricciones:
 - En caso el valor de algún gen o condición en una regla fuzzy hija sea un número decimal, el valor es redondeado a su entero inferior.
 - En caso el valor de algún gen o condición en una regla fuzzy hija sea menor al límite inferior, el valor es modificado al límite inferior.
 - En caso el valor de algún gen o condición en una regla fuzzy hija sea mayor al límite superior, el valor es modificado al límite superior.

Un ejemplo de cruzamiento de dos reglas fuzzy utilizado en este trabajo es mostrado en la Figura 6.5.

3. **Operador de Mutación:** Después del cruzamiento, sobre ambas reglas fuzzy hijas es aplicado o utilizado el operador de mutación según una probabilidad $ProbMutaBR$. Utilizamos la mutación polinomial, detallado en el Capítulo 3. En este caso consideramos las siguientes restricciones:

6. MÉTODO PROPUESTO: MO-FRBCS-IRL-ID

Regla Fuzzy Padre 1	4	0	2	0	3	1
Regla Fuzzy Padre 2	0	3	1	5	0	2
Pre-regla Fuzzy Hija 1	3.5	0.3	1.9	0.6	2.7	1.1
Pre-regla Fuzzy Hija 2	0.5	2.7	1.1	4.4	0.3	1.9
Regla Fuzzy Hija 1	3	0	1	0	2	1
Regla Fuzzy Hija 2	0	2	1	4	0	1

Figura 6.5: Ejemplo de Cruzamiento SBX en la Generación de Reglas Fuzzy.

$$u = 0,3 - \eta = 1 - \beta = 0,7746$$

- En caso el valor de algún gen o condición en una regla fuzzy hija sea un número decimal, el valor es redondeado a su entero inferior.
- En caso el valor de algún gen o condición en la regla fuzzy mutada sea menor al límite inferior, el valor es modificado al límite inferior.
- En caso el valor de algún gen o condición en la regla fuzzy mutada sea mayor al límite superior, el valor es modificado al límite superior.

Un ejemplo de mutación de una regla fuzzy utilizado en este trabajo es mostrado en la Figura 6.6 donde cambian los valores del segundo y último gen.

Los tres operadores son repetidos hasta completar la población Q_t según la técnica NSGA-II.

6.2 Aprendizaje Genético Multiobjetivo Iterativo de la Base de Reglas Fuzzy para Clasificación

Regla Fuzzy	3	0	1	0	2	1
Pre Regla Fuzzy Mutada	3	2.2	1	0	2	2.1
Regla Fuzzy Mutada	3	2	1	0	2	2

Figura 6.6: Ejemplo de Mutación Polinomial en la Generación de Reglas Fuzzy.

$$u = 0,8 - \eta = 1$$

6.2.7 Condición de Parada y Selección de la Mejor Regla Fuzzy

Como condición de parada utilizamos la cantidad de iteraciones para la técnica NSGA-II ($BR_CantIterNSGA - II$) a definir antes de empezar el proceso de aprendizaje de reglas fuzzy.

Una vez que la condición de parada es alcanzada, el resultado es un conjunto de reglas fuzzy no dominadas en ambos objetivos o frontera de Pareto. Ese conjunto es ordenado de mayor a menor según el objetivo de precisión y seleccionada la regla fuzzy que se encuentre en la posición central. En caso existan dos reglas fuzzy en la posición central, una de ellas es seleccionada de forma aleatoria. Dicha regla es inserida en la BR final. En caso la tasa clasificación mejore, utilizando el método de clasificación fuzzy clásico detallado en el Capítulo 4, se procede a la marcación de los ejemplos y el proceso de generación de reglas fuzzy es repetido. En caso contrario, dicha regla fuzzy no mejore la tasa de clasificación, la regla fuzzy es eliminada de la BR final y de la frontera de Pareto, y el proceso de selección de la mejor regla es repetido. Cuando todas las reglas fuzzy de la frontera de Pareto son eliminadas, ninguna regla fuzzy es adicionada a la BR final y el proceso de aprendizaje de reglas fuzzy termina.

6. MÉTODO PROPUESTO: MO-FRBCS-IRL-ID

6.2.8 Marcación de Ejemplos en el Aprendizaje de Reglas Fuzzy

En caso una regla fuzzy de la frontera de Pareto sea seleccionada e insertada en la BR final, los ejemplos, de la base de ejemplos balanceada, con un grado de compatibilidad superior a gm y de la misma clase de la regla fuzzy son marcados.

Los ejemplos marcados, como mencionado anteriormente, influyen en el cálculo de la función de aptitud del objetivo de precisión. Específicamente, en la integridad y la consistencia de la regla fuzzy los ejemplos marcados no son considerados cuando se analizan los ejemplos que tengan la misma clase de la regla fuzzy.

6.3 Optimización Genética Lateral Multiobjetivo de la Base de Datos Fuzzy

Una vez terminadas las etapas de pre-procesamiento de los ejemplos y aprendizaje de la BR fuzzy, comienza la etapa de optimización genética lateral multiobjetivo de la BD fuzzy.

Esta etapa está basada en el método propuesto en [Gacto et al., 2010], detallado en el Capítulo 5, y en la técnica de optimización multiobjetivo NSGA-II detallada en el Capítulo 3. Principalmente, esta etapa consiste en cambiar los parámetros a , m y b de los conjunto fuzzy en la BD fuzzy, dentro de los intervalos permitidos, para disminuir la tasa de error del SCFBR en la base de datos balanceada. Como en la etapa anterior, en esta etapa utilizamos el método de clasificación fuzzy clásico por lo que al cambiar los parámetros de los conjuntos fuzzy, el grado de compatibilidad de las reglas con cada ejemplo cambia y por consiguiente puede cambiar la tasa de error.

El resultado de esta etapa es la BD fuzzy optimizada que contiene el conjunto de parámetros para cada conjunto fuzzy. A continuación detallamos los componentes utilizados en la técnica NSGA-II para alcanzar dicho resultado:

6.3.1 Representación de la Base de Datos Fuzzy

Cada cromosoma en la técnica NSGA-II, definido por una cadena de números decimales, representa una BD fuzzy completa. El cromosoma es de tamaño fijo y cada

6.3 Optimización Genética Lateral Multiobjetivo de la Base de Datos Fuzzy

gen es un parámetro de algún conjunto fuzzy triangular en la BD fuzzy.

El cromosoma comienza con los parámetros a , m y b , del primer conjunto fuzzy de la primera variable lingüística del problema; continua con el segundo conjunto fuzzy de la primera variable lingüística del problema; y así sucesivamente para todos los conjuntos fuzzy de la primera variable lingüística. Después de ello, el proceso es repetido con la segunda, tercera, hasta la última variable lingüística del problema. Por ejemplo, la Figura 6.7 muestra el cromosoma que representa una BD fuzzy en el caso que el problema considere dos variables lingüística y cada variable con tres términos lingüísticos. El nombre de cada gen está representado por el nombre del parámetro (a , m o b), el índice de la variable lingüística y como subíndice el índice del término lingüístico.

-0.49	0.00	0.50	-0.03	0.38	1.00	0.49	1.00	1.50	-0.11	-0.07	0.29	-0.06	0.26	0.44	0.26	0.44	0.63
a_{1_1}	m_{1_1}	b_{1_1}	a_{1_2}	m_{1_2}	b_{1_2}	a_{1_3}	m_{1_3}	b_{1_3}	a_{2_1}	m_{2_1}	b_{2_1}	a_{2_2}	m_{2_2}	b_{2_2}	a_{2_3}	m_{2_3}	b_{2_3}

Figura 6.7: Representación de una BD fuzzy.

6.3.2 Generación de la Población Inicial en la Optimización de la Base de Datos Fuzzy

La generación de los cromosomas iniciales que representa una BD fuzzy completa es de forma aleatoria, en otras palabras, los parámetros a , m y b para cada término lingüístico en cada variable lingüística son definidos de forma aleatoria considerando los intervalos o límites según [Gacto et al., 2010] y detallado en la Sección 5.2.

La BD fuzzy utilizada para definir los intervalos para cada parámetro es la BD fuzzy uniformemente distribuida generada en la etapa de aprendizaje de la BR fuzzy. También, un cromosoma o individuo de la población inicial es la BD fuzzy uniformemente distribuida.

6.3.3 Función de Aptitud del Objetivo de Precisión en la Optimización de la Base de Datos Fuzzy

Cada cromosoma, que representa una BD fuzzy completa, es utilizada junto a la BR final en el método fuzzy de clasificación clásico para definir el porcentaje de

6. MÉTODO PROPUESTO: MO-FRBCS-IRL-ID

error o tasa de error en la base de datos balanceada. Este objetivo es un objetivo a minimizar; formalmente definimos la función de aptitud del objetivo de precisión (FA_{Prec}) mediante la siguiente ecuación:

$$FA_{Prec}(BD) = \text{Minimizar} (CantEjemErr / CantiEjem) \quad (6.5)$$

Donde:

- BD : Base de Datos fuzzy.
- $CantEjemErro$: Cantidad de ejemplos clasificados de forma incorrecta usando la BD fuzzy codificada en un cromosoma y la BR final, en la base de datos balanceada.
- $CantiEjem$: Cantidad de ejemplos en la base de datos balanceada.

6.3.4 Función de Aptitud del Objetivo de Interpretabilidad en la Optimización de la Base de Datos Fuzzy

Este objetivo es calculado en base al IIS definido como GM3M, propuesto en trabajo [Gacto et al., 2010] y detallado en la sección 5.2. Este objetivo es un objetivo a minimizar por lo cual consideramos el negativo del GM3M; formalmente definimos la función de aptitud del objetivo de interpretabilidad (FA_{Inte}) mediante la siguiente ecuación:

$$FA_{Inte}(BD) = \text{Minimizar} (-GM3M) \quad (6.6)$$

En este punto es importante indicar que la función aptitud del objetivo de interpretabilidad para la BD con conjuntos fuzzy uniformemente distribuido es uno y en base a la misma se calcula las métricas de desplazamiento, amplitud lateral y similaridad de áreas dentro del GM3M.

6.3.5 Operadores Genéticos en la Optimización de la Base de Datos Fuzzy

A continuación detallamos y explicamos que operadores genéticos son utilizados en la optimización genética multiobjetivo de la BD fuzzy, que son los mismo utili-

6.3 Optimización Genética Lateral Multiobjetivo de la Base de Datos Fuzzy

zados en el aprendizaje genético multiobjetivo de la BR fuzzy:

1. Operador de Selección: Se realiza una selección por torneo de dos soluciones. Antes del cruzamiento, el siguiente proceso es repetido dos veces, uno para cada BD fuzzy padre. Son seleccionadas dos BD fuzzy de forma aleatoria, y son comparadas según su dominancia; en caso ninguna solución domine a la otra (están en la misma frontera), se comparan las distancias de multitud seleccionando la BD fuzzy con mayor distancia de multitud; en caso las distancias de multitud sean iguales, una de las BD fuzzy es seleccionada de forma aleatoria.
2. Operador de Cruzamiento: Una vez obtenidas las BD fuzzy padres mediante el operador de selección, según una probabilidad $ProbCruzaBD$, utilizamos el cruzamiento SBX, detallado en el Capítulo 3. En este caso consideramos las siguientes restricciones:
 - En caso el valor de algún gen o parámetro de algún conjunto fuzzy sea menor al límite inferior para ese parámetro, el valor es modificado al límite inferior.
 - En caso el valor de algún gen o parámetro de algún conjunto fuzzy sea mayor al límite superior para ese parámetro, el valor es modificado al límite superior.

Un ejemplo de cruzamiento de dos reglas BD fuzzy utilizado en este trabajo es mostrado en la Figura 6.8, considerando que todos los valores de las BD fuzzy hijas están dentro de los límites permitidos para cada parámetro.

3. Operador de Mutación: Después del cruzamiento, sobre ambas BD fuzzy hijas es aplicado o utilizado el operador de mutación según una probabilidad $ProbMutaBD$. Utilizamos la mutación polinomial, detallado en el Capítulo 3. En este caso consideramos las siguientes restricciones:
 - En caso el valor de algún gen o condición de algún conjunto fuzzy después de la mutación sea menor al límite inferior, el valor es modificado al límite inferior.

6. MÉTODO PROPUESTO: MO-FRBCS-IRL-ID

BD Fuzzy Padre 1	-0.49	0.00	0.50	-0.03	0.38	1.00	0.49	1.00	1.50	-0.11	-0.07	0.29	-0.06	0.26	0.44	0.26	0.44	0.63
BD Fuzzy Padre 2	-0.30	0.25	0.47	-0.41	0.38	1.24	0.54	1.04	1.64	0.02	-0.10	0.31	-0.04	0.29	0.55	0.31	0.47	0.60
BD Fuzzy Hija 1	-0.52	-0.04	0.50	0.03	0.38	0.97	0.48	0.99	1.48	-0.13	-0.07	0.29	-0.06	0.26	0.42	0.25	0.44	0.63
BD Fuzzy Hija 2	-0.27	0.29	0.47	-0.47	0.38	1.27	0.55	1.05	1.66	0.04	-0.10	0.31	-0.04	0.29	0.57	0.32	0.47	0.60

Figura 6.8: Ejemplo de Cruzamiento SBX en la Optimización de la BD Fuzzy.

$$u = 0,7 - \eta = 1 - \beta = 1,2909$$

- En caso el valor de algún gen o condición de algún conjunto fuzzy después de la mutación sea mayor al límite superior, el valor es modificado al límite superior.

Un ejemplo de mutación de una BD fuzzy utilizado en este trabajo es mostrado en la Figura 6.9, donde cambian los valores del primer y penúltimo gen, considerando que esos valores están dentro de los límites permitidos.

BD Fuzzy	-0.52	-0.04	0.50	0.03	0.38	0.97	0.48	0.99	1.48	-0.13	-0.07	0.29	-0.06	0.26	0.42	0.25	0.44	0.63
BD Fuzzy Mutada	-0.29	-0.04	0.50	0.03	0.38	0.97	0.48	0.99	1.48	-0.13	-0.07	0.29	-0.06	0.26	0.42	0.25	0.38	0.63

Figura 6.9: Ejemplo de Mutación Polinomial en la Optimización de la BD Fuzzy.

$$u = 0,3 - \eta = 1$$

Los tres operadores son repetidos hasta completar la población Q_t según la técnica NSGA-II. Un ejemplo de una posible BD fuzzy con conjunto fuzzy optimizados es mostrado en la Figura 6.10, donde vemos que la interpretabilidad de la BD fuzzy ha disminuido pero la precisión pudo aumentar.

6.3.6 Condición de Parada y Selección de la Base de Datos Fuzzy

Como condición de parada utilizamos la cantidad de iteraciones para la técnica NSGA-II ($BDCantIterNSGA - II$) a definir antes de empezar el proceso de optimización de la BD fuzzy.

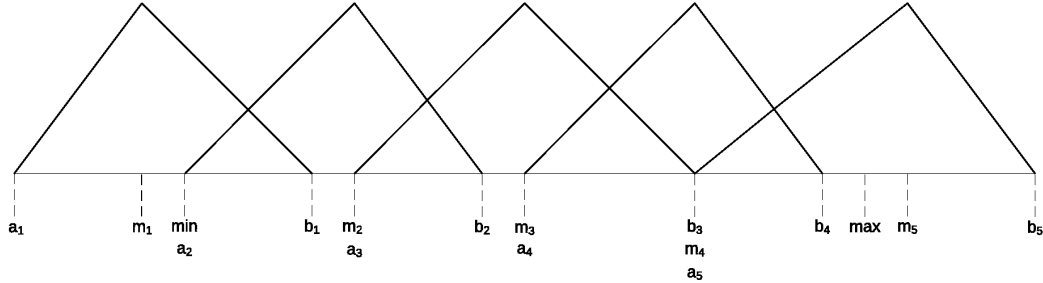


Figura 6.10: Ejemplo de posible BD fuzzy optimizada.

Una vez que la condición de parada es alcanzada, el resultado es un conjunto de BD fuzzy no dominadas en ambos objetivos o frontera de Pareto. Ese conjunto es ordenado de mayor a menor según el objetivo de precisión y seleccionada la BD fuzzy que se encuentre en la posición central. En caso existan dos BD fuzzy en la posición central, una de ellas es seleccionada de forma aleatoria. Con dicha selección el proceso de optimización de la BD fuzzy termina.

6.4 Consideraciones Finales

En este capítulo es presentado un nuevo método para la generación de SFCBR para bases de datos no balanceadas llamado MO-FRBCS-IRL-ID. El método MO-FRBCS-IRL-ID presenta algunas diferencias en relación a otros métodos para el mismo propósito, principalmente la combinación del balanceamiento o pre-procesamiento de datos con el enfoque IRL basado en un AGMO en la generación de la BR fuzzy. En el próximo capítulo es realizado un análisis experimental del método MO-FRBCS-IRL-ID y comparado con otros métodos encontrados en la literatura.

Experimentos

En este capítulo son presentados los experimentos realizados aplicando el método propuesto, llamado MO-FRBCS-IRL-ID, para resolver el problema de clasificación en bases de datos o ejemplos no balanceadas.

Con el objetivo de realizar un análisis comparativo se utilizaron los resultados de precisión e interpretabilidad obtenidos por los métodos GA-FS+GL [VILLAR et al., 2012], GP-COACH-H [López et al., 2013] y PAES-3obj [Antonelli et al., 2013], vistos en el Capítulo 5, y compararlos con el método propuesto. Los resultados de ese análisis comparativo también son discutidos en este capítulo.

7.1 Plan Experimental

Para la validación del método propuesto para la generación genética multiobjetivo de bases de conocimiento fuzzy para clasificación en bases de datos no balanceada usando el enfoque iterativo, son utilizadas bases de datos o ejemplos no balanceadas con un radio de no balanceamiento o desbalanceamiento mayor a 9 $IR > 9$. Esas bases de datos son obtenidas del repositorio de datos para aprendizaje de máquina KEEL (Knowledge Extraction based on Evolutionary Learning [Alcalá-Fdez et al., 2011]), incluyendo abalone, ecoli, glass, page-blocks, shuttle, vowel, yeast y las variantes de cada base de datos no balanceada mediante la agrupación

7.1 Plan Experimental

de las clases para considerar solo dos clases: clase positiva y clase negativa. En la Tabla 7.1 son presentadas las características de cada base de datos no balanceada, específicamente el número de ejemplos, el número de variables del problema y el IR, para cada una de las veintidós bases de datos no balanceadas utilizadas en este trabajo.

Base de Datos	# ejemplos	# variables	IR
Abalone19	4174	8	128.87
Abalone9vs18	731	8	16.68
Ecoli0137vs26	281	7	39.15
Ecoli4	336	7	13.84
Glass016vs2	192	9	10.29
Glass016vs5	184	9	19.44
Glass2	214	9	10.39
Glass4	214	9	15.47
Glass5	214	9	22.81
Page-blocks13vs2	472	10	15.85
Shuttle0vs4	1829	9	13.87
Shuttle2vs4	129	9	20.5
Vowel0	988	13	10.1
Yeast05679vs4	528	8	9.35
Yeast1289vs7	947	8	30.56
Yeast1458vs7	693	8	22.1
Yeast1vs7	459	8	13.87
Yeast2vs4	514	8	9.08
Yeast2vs8	482	8	23.1
Yeast4	1484	8	28.41
Yeast5	1484	8	32.78
Yeast6	1484	8	39.15

Tabla 7.1: Bases de datos no balanceadas utilizadas en los experimentos.

El método seleccionado para la validación de cada método es la validación cruzada con cinco particiones, altamente utilizado y recomendado para la validación

7. EXPERIMENTOS

de métodos de aprendizaje de máquina. Es importante mencionar que todos los métodos, es decir, el método propuesto y los tres métodos utilizados para la comparación, utilizan las mismas particiones (conjuntos de entrenamiento y prueba) de las bases de datos no balanceadas.

Para cada método son obtenidos los siguientes valores: la precisión definida por el AUC en los conjuntos de prueba; y la interpretabilidad definida por el número de reglas fuzzy y el número de condiciones en todas las reglas fuzzy de los SFCBR generados por cada método.

Los parámetros del método MO-FRBCS-IRL-ID fueron determinados mediante el conocimiento previo de los autores y son mostrados en la Tabla 7.2.

Parámetro	Valor
Número de individuos en la población de la generación de la BR fuzzy	100
Número de individuos en la población de la optimización de la BD fuzzy	50
Número máximo de generaciones de la generación de la BR fuzzy	1000
Número máximo de generaciones de la optimización de la BD fuzzy	500
Probabilidad de cruzamiento (<i>ProbCruzaBR</i> y <i>ProbCruzaBD</i>)	90 %
Probabilidad de mutación (<i>ProbMutaBR</i> y <i>ProbMutaBD</i>)	10 %
Grado de marcación (<i>gm</i>)	0.3

Tabla 7.2: Parámetros utilizados por el método MO-FRBCS-IRL-ID en los experimentos.

Para verificar la significancia estadísticas de las diferencias en los resultados observados por cada método, par a par, fue utilizado la prueba no paramétrica de rangos con signo de Wilcoxon [Wilcoxon, 1945] [Mann and Whitney, 1947]. Los datos utilizados son los valores medios de las AUC para cada base de ejemplos no balanceada, así como la cantidad de reglas fuzzy y condiciones. La prueba de rangos con signo de Wilcoxon es utilizado con un nivel de significación $\alpha = 0,05$.

7.2 Experimentación

Los resultados obtenidos para el método MO-FRBCS-IRL-ID son presentados en la Tabla 7.3.

7.2 Experimentación

Base de Datos	Pre-Proc.	AUC	# Regl.	# Condi.	IIS
abalone19	SMOTE	0.785 (0.106)	5.800 (1.924)	9.400 (3.647)	0.179 (0.138)
abalone9-18	Borderline_SMOTE	0.791 (0.101)	4.800 (1.095)	8.000 (2.449)	0.265 (0.121)
ecoli-0-1-3-7_vs_2-6	SMOTE_TL	0.852 (0.203)	7.000 (1.225)	10.800 (3.114)	0.128 (0.161)
ecoli4	SMOTE_ENN	0.959 (0.031)	6.800 (1.304)	12.400 (3.435)	0.371 (0.121)
glass-0-1-6_vs_2	Safe_Level_SMOTE	0.714 (0.071)	4.600 (1.140)	6.600 (2.966)	0.209 (0.219)
glass-0-1-6_vs_5	Borderline_SMOTE	0.980 (0.016)	7.200 (1.483)	12.600 (2.510)	0.188 (0.087)
glass2	Safe_Level_SMOTE	0.746 (0.061)	5.200 (1.483)	9.000 (3.742)	0.144 (0.197)
glass4	Borderline_SMOTE	0.949 (0.075)	6.800 (1.643)	13.600 (3.050)	0.064 (0.079)
glass5	SMOTE_TL	0.891 (0.100)	8.200 (3.347)	16.200 (6.058)	0.192 (0.154)
page-blocks-1-3_vs_4	Borderline_SMOTE	0.990 (0.005)	5.400 (1.140)	9.600 (1.949)	0.154 (0.134)
shuttle-c0-vs-c4	None	1.000 (0.000)	3.000 (0.000)	3.000 (0.000)	0.185 (0.265)
shuttle-c2-vs-c4	NCL	1.000 (0.000)	2.200 (0.447)	2.200 (0.447)	0.150 (0.160)
vowel0	ADASYN	0.975 (0.008)	9.600 (2.966)	21.400 (11.261)	0.153 (0.169)
yeast-0-5-6-7-9_vs_4	SMOTE	0.841 (0.032)	6.600 (1.517)	13.400 (3.362)	0.113 (0.148)
yeast-1-2-8-9_vs_7	ADASYN	0.758 (0.065)	6.600 (1.817)	15.000 (3.742)	0.163 (0.125)
yeast-1-4-5-8_vs_7	SMOTE	0.662 (0.055)	5.200 (2.387)	10.200 (5.263)	0.256 (0.164)
yeast-1_vs_7	ADASYN	0.684 (0.060)	8.600 (3.130)	17.600 (8.989)	0.228 (0.153)
yeast-2_vs_4	Safe_Level_SMOTE	0.915 (0.028)	4.400 (0.894)	8.200 (2.588)	0.050 (0.074)
yeast-2_vs_8	SMOTE_ENN	0.810 (0.105)	2.800 (0.447)	4.200 (1.643)	0.041 (0.067)
yeast4	SMOTE_TL	0.859 (0.022)	7.200 (2.588)	13.200 (5.119)	0.224 (0.205)
yeast5	Borderline_SMOTE	0.965 (0.011)	4.600 (0.548)	7.600 (1.817)	0.162 (0.099)
yeast6	Borderline_SMOTE	0.917 (0.072)	7.400 (0.548)	12.600 (2.191)	0.177 (0.147)

Tabla 7.3: Valores medios obtenidos por el método propuesto.

7. EXPERIMENTOS

Los valores de desviación estándar son mostrados entre paréntesis. Para cada base de datos no balanceada es utilizado el método MO-FRBCS-IRL-ID con diferentes técnicas de balanceamiento. La Tabla 7.3 muestra la técnica de balanceamiento donde se obtuvo mejor resultado en base a la precisión. En esos resultados apenas una base de datos, shuttle-c0-vs-c4, muestra mejores resultados cuando no se aplica ninguna técnica de balanceamiento (none) y también apenas una base de datos shuttle-c2-vs-c4 muestra mejores resultados al utilizar una técnica de submuestreo (NCL). Las demás bases de datos, muestran mejores resultados al utilizar alguna técnica de sobremuestreo o híbrido, específicamente, quince que utilizan una técnica de sobre muestreo y cinco que utilizan una técnica híbrida. Por lo cual podemos considerar que las técnicas de sobremuestreo e híbridos al poblar el área de los pequeños grupos disyuntos y disminuir la superposición de ejemplos ayudan a la generación genética multiobjetivo iterativa para generar reglas fuzzy y optimizar la BD fuzzy que describan los ejemplos de la clase positiva y de la clase negativa.

Los resultados para cada técnica de balanceamiento, así como los resultados obtenidos por la generación de reglas fuzzy, es decir, antes de la optimización de la BD fuzzy, para las veintidós bases de datos no balanceadas son mostrados en el Anexo A.

La Tabla 7.4 muestra el valor de la media del AUC para el método propuesto en este trabajo y los tres métodos utilizados para la comparación. Para cada base de datos no balanceada, el mejor resultado es resaltado en negrita. En caso de empate, los resultados son resaltados en negrita.

El método propuesto MO-FRBCS-IRL-ID obtiene mejores resultados en la mayoría de bases de datos no balanceadas, específicamente en quince de las veintidós bases de datos no balanceadas. El método PAES-3obj obtuvo mejores resultados en seis bases de datos no balanceadas. Seguida del método GP-COACH-H que obtuvo mejores resultados en dos bases de datos no balanceadas, siendo una de ellas empate con el método MO-FRBCS-IRL-ID. Por último el método GA-FS+GL no obtuvo mejor resultado en ningún caso, pero esos resultados están cerca de la media de los otros métodos, incluso obteniendo mejor media general que el método GP-COACH-H. Por consiguiente, el método propuesto MO-FRBCS-IRL-ID obtuvo una mejor media de clasificación basado en el AUC que los otros métodos.

Base de Datos	GA-FS +GL	GP-COACH -H	PAES -3obj	MO-FRBCS -IRL-ID
abalone19	0.699	0.594	0.727	0.785
abalone9-18	0.628	0.780	0.811	0.791
ecoli-0-1-3-7_vs_2-6	0.790	0.898	0.892	0.852
ecoli4	0.860	0.934	0.909	0.959
glass-0-1-6_vs_2	0.626	0.646	0.662	0.714
glass-0-1-6_vs_5	0.867	0.871	0.885	0.980
glass2	0.729	0.609	0.705	0.746
glass4	0.853	0.793	0.864	0.949
glass5	0.755	0.856	0.898	0.891
page-blocks-1-3_vs_4	0.937	0.884	0.951	0.990
shuttle-c0-vs-c4	0.999	0.996	0.996	1.000
shuttle-c2-vs-c4	0.992	1.000	0.937	1.000
vowel0	0.928	0.946	0.903	0.975
yeast-0-5-6-7-9_vs_4	0.793	0.785	0.835	0.841
yeast-1-2-8-9_vs_7	0.750	0.696	0.688	0.758
yeast-1-4-5-8_vs_7	0.641	0.588	0.645	0.662
yeast-1_vs_7	0.754	0.681	0.775	0.684
yeast-2_vs_4	0.884	0.897	0.944	0.915
yeast-2_vs_8	0.710	0.731	0.814	0.810
yeast4	0.835	0.813	0.868	0.859
yeast5	0.935	0.928	0.952	0.965
yeast6	0.870	0.880	0.888	0.917
Media	0.811	0.809	0.843	0.866

Tabla 7.4: Comparación de medias del AUC para los datos de prueba.

Para mostrar si los valores medios del AUC obtenidos con el método propuesto MO-FRBCS-IRL-ID son significativamente diferentes, se comparó el método propuesto con cada uno de los otros métodos utilizando la prueba no paramétrica de rangos con signo de Wilcoxon. La Tabla 7.5 muestran los valores de ρ , en el caso que ρ sea menor al nivel de significación $\alpha = 0,05$ se rechaza la hipótesis nula (no

7. EXPERIMENTOS

existe diferencia significativa) y se concluye que existe diferencia significativa. En el caso de la comparación entre el método MO-FRBCS-IRL-ID y GP-COACH-H el valor de N es veintiuno porque existe un valor medio del AUC igual para ambos métodos. Los resultados muestran que existe diferencia significativa entre los valores del método propuesto y los tres métodos utilizados en la comparación de acuerdo a la prueba de rangos con signo de Wilcoxon.

	GA-FS +GL	GP-COACH -H	PAES -3obj
MO-FRBCS	$N = 22$	$N = 21$	$N = 22$
-IRL-ID	$\rho = 0,000$	$\rho = 0,000$	$\rho = 0,035$

Tabla 7.5: Resultados de la prueba de rangos con signo de Wilcoxon para la precisión (medias del AUC). Nivel de significación 0.05. Intervalo de Confianza 95 %.

Se analizó la interpretabilidad del SCFBR para bases de datos no balanceadas según la cantidad de reglas fuzzy y la cantidad de condiciones en la BR. La Tabla 7.6 muestra la cantidad de reglas fuzzy obtenidas por el método GA-FS+GL y el método propuesto MO-FRBCS-IRL-ID en este trabajo. No se muestra la cantidad de reglas obtenidas por el método GP-COACH-H porque en el trabajo [López et al., 2013] los autores no proveen información acerca de esas cantidad, y por otro lado, las reglas generadas son del tipo DNF, las cuales son reglas fuzzy de baja interpretabilidad. También, se muestra la cantidad media, entre la cantidad media mínima y la cantidad media máxima de reglas fuzzy, del método PAES-3obj dado que ese método genera un conjunto de SCFBR no dominados para bases de datos no balanceadas.

Comparando las medias de las cantidad de reglas fuzzy es posible notar que el método PAES-3obj levemente obtiene una cantidad de reglas fuzzy menor que los otros métodos. El método PAES-3obj es mejor que los otros métodos en nueve de las veintidós bases de datos, así como que la media total es menor; mientras que el método GA-FS+GL es mejor que los otros métodos en siete de las veintidós bases de datos, pero muestra una media mayor, dado que en algunas bases de datos la cantidad de reglas es considerablemente mayor, por ejemplo, en el caso de Ecoli4;

Base de Datos	GA-FS +GL	PAES -3obj	MO-FRBCS -IRL-ID
Abalone19	7.400	7.785	5.800
Abalone9-18	4.400	7.350	4.800
Ecoli0137vs26	4.200	5.030	7.000
Ecoli4	58.200	4.500	6.800
Glass016vs2	3.600	6.965	4.600
Glass016vs5	4.000	3.550	7.200
Glass2	8.200	5.900	5.200
Glass4	8.400	5.015	6.800
Glass5	5.200	4.100	8.200
Page-Blocks13vs4	6.600	4.580	5.400
shuttle0vs4	14.800	2.635	3.000
shuttle2vs4	4.200	2.050	2.200
Vowel0	4.400	5.335	9.600
Yeast05679vs4	9.400	6.900	6.600
Yeast1289vs7	13.800	7.300	6.600
Yeast1458vs7	3.800	6.865	5.200
yeast1vs7	6.800	6.730	8.600
Yeast2vs4	2.200	5.700	4.400
Yeast2vs8	5.000	5.550	2.800
Yeast4	6.400	6.365	7.200
Yeast5	4.800	6.165	4.600
Yeast6	5.000	7.400	7.400
Mean	8.670	5.630	5.909

Tabla 7.6: Comparación de medias de las cantidades de reglas fuzzy para los datos de prueba.

por último el método propuesto MO-FRBCS-IRL-ID en este trabajo solo obtiene mejores resultados en seis de las veintidós bases de datos, aunque su media es similar a la del método PAES-3obj.

Para mostrar si los valores medios de la cantidad de reglas fuzzy obtenidos

7. EXPERIMENTOS

con el método propuesto MO-FRBCS-IRL-ID son significativamente diferentes, se comparó el método propuesto con cada uno de los otros métodos utilizando la prueba no paramétrica de rangos con signo de Wilcoxon. La Tabla 7.7 muestra los valores de ρ . En el caso que ρ sea menor al nivel de significación $\alpha = 0,05$ se rechaza la hipótesis nula (no existe diferencia significativa) y se concluye que existe diferencia significativa. Esos resultado muestran que no existe diferencia significativa entre los valores del método propuesto y los tres métodos utilizados en la comparación de acuerdo con la prueba de rangos con signo de Wilcoxon. El valor de N es veintiuno cuando se comparó el método PAES-3obj y el método MO-FRBCS-IRL-ID porque en una base de datos las medias de cantidad de reglas fuzzy son iguales.

	GA-FS +GL	PAES -3obj
MO-FRBCS	$N = 22$	$N = 21$
-IRL-ID	$\rho = 0,758$	$\rho = 0,677$

Tabla 7.7: Resultados de la prueba de rangos con signo de Wilcoxon para la interpretabilidad (medias de la cantidad de reglas fuzzy). Nivel de significación 0.05. Intervalo de Confianza 95 %.

La Tabla 7.8 muestra la cantidad de condiciones obtenidas por el método GA-FS+GL y el método MO-FRBCS-IRL-ID propuesto en este trabajo. No se muestra la cantidad de condiciones del método GP-COACH-H por la misma razón que no se mostraron la cantidad media de reglas fuzzy. También, se muestra la cantidad media de condiones, entre la cantidad media mínima y la cantidad media máxima de condiciones, del método PAES-3obj.

Después de realizar la comparación de las medias de las cantidad de condiciones es posible notar que el método propuesto MO-FRBCS-IRL-ID obtiene menores cantidades que los otros métodos. En dieciséis de las veintidós bases de datos no balanceadas el método propuesto obtiene mejores resultados, mientras que los métodos GA-FS+GL y PAES-obj obtienen mejores resultados en seis bases de datos no balanceadas, tres bases para cada método. La media general de cantidad de

Base de Datos	GA-FS +GL	PAES -3obj	MO-FRBCS -IRL-ID
Abalone19	14.800	37.450	9.400
Abalone9-18	12.320	33.865	8.000
Ecoli0137vs26	10.080	14.050	10.800
Ecoli4	162.960	14.265	12.400
Glass016vs2	10.800	33.580	6.600
Glass016vs5	12.000	8.770	12.600
Glass2	24.600	26.500	9.000
Glass4	26.880	18.450	13.600
Glass5	15.600	11.920	16.200
Page-Blocks13vs4	21.120	20.985	9.600
shuttle0vs4	32.560	6.350	3.000
shuttle2vs4	12.600	2.065	2.200
Vowel0	22.880	33.870	21.400
Yeast05679vs4	20.680	30.100	13.400
Yeast1289vs7	41.400	33.085	15.000
Yeast1458vs7	8.360	32.265	10.200
yeast1vs7	19.040	27.185	17.600
Yeast2vs4	6.600	23.535	8.200
Yeast2vs8	27.000	23.115	4.200
Yeast4	17.920	28.570	13.200
Yeast5	14.400	26.400	7.600
Yeast6	14.000	33.035	12.600
Mean	24.940	23.610	10.764

Tabla 7.8: Comparación de medias de las cantidades de condiciones para los datos de prueba.

condiciones es similar entre los métodos GA-FS+GL y PAES-obj, los cuales son superiores a la media general obtenida por el método propuesto.

Para mostrar si los valores medios de la cantidad de condiciones obtenidos con el método propuesto MO-FRBCS-IRL-ID son significativamente diferentes,

7. EXPERIMENTOS

se comparó el método propuesto con cada uno de los otros métodos utilizando la prueba no paramétrica de rangos con signo de Wilcoxon. La Tabla 7.9 muestran los valores de ρ . En el caso que ρ sea menor al nivel de significación $\alpha = 0,05$ se rechaza la hipótesis nula (no existe diferencia significativa) y se concluye que existe diferencia significativa. Esos resultado muestran que existe diferencia significativa entre los valores del método propuesto y los otros dos métodos utilizados en la comparación de acuerdo con la prueba de rangos con signo de Wilcoxon.

	GA-FS +GL	PAES -3obj
MO-FRBCS	$N = 22$	$N = 22$
-IRL-ID	$\rho = 0,001$	$\rho = 0,000$

Tabla 7.9: Resultados de la prueba de rangos con signo de Wilcoxon para la interpretabilidad (medias de la cantidad de condiciones). Nivel de significación 0.05. Intervalo de Confianza 95 %.

La media de valores de IIS no son comparados con los otros métodos dado que el método GA-FS+GL utiliza diferentes valores de granularidad en la BD, el método GP-COACH-H utiliza conjuntos fuzzy de 2-tuplas (donde cada término lingüístico está compuesto por un término lingüístico y un número [Herrera et al., 1999]), el método PAES-obj utiliza particiones fuertes fuzzy (los parámetros de los conjuntos fuzzy consecutivos en cada variable lingüística están relacionados), y no es un objetivo que los otros métodos utilizaron en la generación de la BC en SFCBR para bases de datos no balanceadas.

7.3 Consideraciones Finales

Este capítulo presentó un análisis experimental del método MO-FRBCS-IRL-ID propuesto en este trabajo. Los resultados obtenidos fueron comparados con tres métodos actuales que generan SFCBR para bases de datos no balanceadas. Los resultados muestran que el método MO-FRBCS-IRL-ID puede obtener mejor precisión que los otros tres métodos utilizando una técnica de sobremuestreo o híbrida

7.3 Consideraciones Finales

en el procesamiento de los datos, así como una menor cantidad de condiciones, siendo la cantidad de reglas fuzzy similar a los otros métodos.

Conclusiones

Los SCFBR han sido ampliamente utilizados para resolver el problema de clasificación dado que son modelos con alta precisión e interpretabilidad, principalmente porque la reglas fuzzy se asemejan al raciocinio humano. Los AGs, así como los AGMO, han sido utilizados para generar diferentes o todos los componentes de los SCFBR. Una nueva tendencia, utiliza los AGMO para generar SCFBR en el problema de bases de datos no balanceadas. Siguiendo esa tendencia, varios trabajos han sido propuestos en base a diferentes enfoques y diferentes objetivos. Este trabajo se une a los otros proponiendo el uso del balanceamiento de datos no balanceados junto al enfoque iterativo en la generación de reglas fuzzy de clasificación e intenta encontrar SCFBR con equilibrados entre los objetivos de precisión e interpretabilidad, para la generación de BC fuzzy (BR y BD) en la clasificación en bases de datos no balanceadas.

Este trabajo propone un método definido como MO-FRBCS-IRL-ID (Multi-Objective Fuzzy Rule-Based Classification Systems based on Iterative Rule Learning approach for Imbalanced Datasets). Este método difiere esencialmente de los demás métodos encontradas en la literatura en tres aspectos: la utilización del balanceamiento de datos junto al enfoque IRL; el enfoque IRL está basado en un AGMO que genera una conjunto de reglas fuzzy no dominadas para bases de datos

balanceadas artificialmente; y las funciones de aptitud en los objetivos de precisión utilizadas en el aprendizaje de reglas fuzzy y la optimización de los conjuntos fuzzy.

Los resultados obtenidos en veintidós bases de de datos no balanceadas, en diversos problemas de clasificación, muestran estadísticamente que el método propuesto MO-FRBCS-IRL-ID tuvo mejor desempeño al obtener mejores resultados en la media del AUC que los métodos GA-FS+GL, GP-COACH-H y PAES-3obj encontrados en la literatura.

Los resultados obtenidos también muestran que el método propuesto obtiene mejores resultados usando técnicas de sobremuestreo o híbridas para el preprocesamiento de los datos no balanceados. Ello se debe a que el balanceamiento de los datos, específicamente los métodos de sobremuestreo o híbridos, ayudan a resolver el problema de pequeños grupos disjuntos y sobre posición de ejemplos, tanto en el aprendizaje de cada regla fuzzy usando al enfoque IRL basado en un AGMO, así como la optimización de los conjuntos fuzzy basado en un AGMO.

Por último, podemos concluir que la cantidad de reglas fuzzy obtenida por el método propuesto MO-SCFRB-IRL-ID no muestra una mejora significativa comparado con los otros tres métodos. Sin embargo, la principal diferencia en la interpretabilidad de los SCFBR obtenidos, incluso mayor diferencia que la obtenida en la precisión, es el menor número de condiciones en la BR final de los SCFBR, dado que se utilizan condiciones tipo don't care en la generación de reglas fuzzy de clasificación.

8.1 Trabajos Futuros

A continuación presentamos algunas ideas de trabajos futuros, así como las mejoras que se pueden hacer al método propuesto en este trabajo definido como MO-FRBCS-IRL-ID.

Además de la técnica de optimización genética multiobjetivo NSGA-II, existen otras técnicas de optimización multiobjetivo que se pueden utilizar tanto para el aprendizaje de reglas fuzzy o para la optimización de los conjuntos fuzzy. Dentro de los más conocidos tenemos la técnica Strength Pareto Evolutionary Algorithm 2 (SPEA2) y la técnica Pareto Achieved Evolution Strategy (PAES).

8. CONCLUSIONES

Se pretende incorporar otras formas de selección de la mejor solución de la frontera de Pareto obtenida por la técnica de optimización multiobjetivo. Principalmente, en la selección de la mejor regla fuzzy en el proceso iterativo basado en el enfoque IRL, por ejemplo, la reglas fuzzy que se encuentre más cerca al punto medio de la curva generada por la frontera de Pareto. Así como, la posibilidad de escoger o combinar más de una regla fuzzy de la frontera Pareto para disminuir el tiempo de entrenamiento del método propuesto.

Sería interesante que además de considerar como características de interpretabilidad la cantidad de reglas fuzzy, la cantidad de condiciones en cada reglas fuzzy y el IIS, considerar otras características, por ejemplo, reglas fuzzy utilizadas o lanzadas (grado de compatibilidad mayor a cero con un ejemplo) al mismo tiempo para clasificar un ejemplo, o disminuir el número de reglas fuzzy contradictorias, es decir, reglas con antecedentes similares y diferentes consecuentes; que pueden ser considerados como objetivos adicionales de interpretabilidad en la generación de reglas fuzzy o en la optimización de conjuntos.

Finalmente, se pretende extender el método propuesto MO-FRBCS-IRL-ID para tratar otra nueva tendencia en la generación de SCFBR usando AGMO, la cual está relacionada al problema de clasificación en bases de datos no balanceadas. Esa nueva tendencia es el problema de clasificación en grandes bases de datos, dado que ese gran tamaño puede afectar el cálculo de las funciones de aptitud al incrementar demasiado el tiempo de procesamiento (por el gran espacio de búsqueda), pudiendo generar también modelos fuzzy más complejos al generar más reglas fuzzy y condiciones.



Resultados Obtenidos con el Método Propuesto

En este anexo mostramos todos los resultados obtenidos por el método propuesto en este trabajo definido como MO-FRBCS-IRL-ID, para las veintidós bases de datos no balanceadas utilizadas en los experimentos y comparación de resultados. Para cada una de las bases de datos no balanceadas utilizadas se muestran los siguientes resultados:

- BDNB: Base de Datos No Balanceada.
- MB: Método de Balanceamiento.
- AUC_BR: Media del AUC o área bajo la curva ROC después del aprendizaje de la BR fuzzy del SCFBR utilizando el enfoque IRL y un AGMO. La desviación estándar de las cinco particiones de la validación cruzada es mostrada entre paréntesis.
- AUC_BD: Media del AUC o área bajo la curva ROC después del aprendizaje de la BR fuzzy y la optimización de la BD fuzzy del SCFBR utilizando un AGMO. La desviación estándar de las cinco particiones de la validación cruzada es mostrada entre paréntesis.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

- # Regl.: Cantidad de reglas fuzzy en el SCFBR obtenido por el método propuesto. La desviación estándar de las cinco particiones de la validación cruzada es mostrada entre paréntesis.
- # Cond.: Cantidad de condiciones en el SCFBR obtenido por el método propuesto. La desviación estándar de las cinco particiones de la validación cruzada es mostrada entre paréntesis.
- # IIS: Índice de Interpretabilidad Semántica en el SCFBR obtenido por el método propuesto. La desviación estándar de las cinco particiones de la validación cruzada es mostrada entre paréntesis.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
abalone19	None	0.500 (0.000)	0.500 (0.000)	2.800 (1.095)	2.800 (1.095)	0.260 (0.414)
abalone19	SMOTE	0.771 (0.101)	0.785 (0.106)	5.800 (1.924)	9.400 (3.647)	0.179 (0.138)
abalone19	Borderline_- SMOTE	0.676 (0.103)	0.680 (0.105)	6.200 (1.304)	11.400 (3.847)	0.457 (0.059)
abalone19	ADASYN	0.767 (0.114)	0.770 (0.110)	5.400 (2.191)	9.400 (5.683)	0.097 (0.099)
abalone19	Safe_Level_- SMOTE	0.783 (0.106)	0.784 (0.106)	5.600 (1.817)	10.600 (4.393)	0.246 (0.156)
abalone19	TL	0.500 (0.000)	0.500 (0.000)	3.200 (1.095)	3.200 (1.095)	0.227 (0.432)
abalone19	OSS	0.500 (0.001)	0.500 (0.001)	1.800 (0.447)	2.000 (0.707)	0.632 (0.504)
abalone19	NCL	0.500 (0.000)	0.500 (0.000)	2.400 (0.894)	2.400 (0.894)	0.073 (0.102)
abalone19	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
abalone19	SMOTE_TL	0.775 (0.102)	0.776 (0.102)	6.200 (1.789)	11.400 (4.037)	0.201 (0.183)
abalone19	SMOTE_- ENN	0.773 (0.109)	0.776 (0.106)	6.400 (1.140)	12.200 (3.564)	0.347 (0.137)

Tabla A.1: Valores medios obtenidos por el método propuesto para la base de datos no balanceada abalone19.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
abalone9-18	None	0.497 (0.003)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	0.109 (0.103)
abalone9-18	SMOTE	0.730 (0.106)	0.736 (0.106)	3.000 (0.707)	4.600 (2.881)	0.357 (0.240)
abalone9-18	Borderline_- SMOTE	0.772 (0.127)	0.791 (0.101)	4.800 (1.095)	8.000 (2.449)	0.265 (0.121)
abalone9-18	ADASYN	0.736 (0.135)	0.752 (0.112)	4.400 (1.140)	7.000 (3.000)	0.257 (0.164)
abalone9-18	Safe_Level_- SMOTE	0.685 (0.074)	0.706 (0.068)	2.800 (0.447)	3.200 (1.095)	0.167 (0.150)
abalone9-18	TL	0.498 (0.003)	0.500 (0.000)	1.400 (0.548)	1.600 (0.894)	0.361 (0.397)
abalone9-18	OSS	0.658 (0.134)	0.640 (0.122)	2.200 (0.447)	2.600 (0.894)	0.091 (0.125)
abalone9-18	NCL	0.498 (0.003)	0.500 (0.000)	1.200 (0.447)	1.200 (0.447)	0.224 (0.435)
abalone9-18	SBC	0.647 (0.041)	0.716 (0.063)	2.600 (0.894)	3.600 (1.673)	0.009 (0.012)
abalone9-18	SMOTE_TL	0.714 (0.094)	0.745 (0.114)	3.600 (1.140)	5.600 (2.881)	0.098 (0.082)
abalone9-18	SMOTE_- ENN	0.724 (0.066)	0.719 (0.058)	2.800 (0.837)	4.600 (3.130)	0.222 (0.152)

Tabla A.2: Valores medios obtenidos por el método propuesto para la base de datos no balanceada abalone9-18.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
ecoli-0-1-3- 7_vs_2-6	None	0.546 (0.109)	0.648 (0.223)	4.400 (0.894)	4.600 (1.140)	0.206 (0.444)
ecoli-0-1-3- 7_vs_2-6	SMOTE	0.842 (0.196)	0.845 (0.204)	7.800 (1.483)	11.600 (2.702)	0.299 (0.051)
ecoli-0-1-3- 7_vs_2-6	Borderline_- SMOTE	0.815 (0.219)	0.832 (0.216)	6.600 (1.517)	7.400 (1.817)	0.194 (0.089)
ecoli-0-1-3- 7_vs_2-6	ADASYN	0.809 (0.197)	0.814 (0.199)	6.200 (2.683)	9.400 (4.336)	0.239 (0.152)
ecoli-0-1-3- 7_vs_2-6	Safe_Level_- SMOTE	0.820 (0.206)	0.831 (0.198)	8.000 (2.121)	12.800 (2.588)	0.244 (0.199)
ecoli-0-1-3- 7_vs_2-6	TL	0.846 (0.221)	0.846 (0.221)	5.200 (0.447)	5.800 (0.447)	0.427 (0.524)
ecoli-0-1-3- 7_vs_2-6	OSS	0.772 (0.082)	0.779 (0.067)	2.200 (0.447)	2.600 (0.548)	0.804 (0.439)
ecoli-0-1-3- 7_vs_2-6	NCL	0.796 (0.275)	0.848 (0.222)	5.400 (0.548)	6.000 (0.707)	0.223 (0.435)
ecoli-0-1-3- 7_vs_2-6	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
ecoli-0-1-3- 7_vs_2-6	SMOTE_TL	0.849 (0.201)	0.852 (0.203)	7.000 (1.225)	10.800 (3.114)	0.128 (0.161)
ecoli-0-1-3- 7_vs_2-6	SMOTE_- ENN	0.845 (0.198)	0.847 (0.199)	7.400 (1.517)	9.600 (3.050)	0.181 (0.245)

Tabla A.3: Valores medios obtenidos por el método propuesto para la base de datos no balanceada ecoli-0-1-3-7_vs_2-6.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
ecoli4	None	0.500 (0.000)	0.500 (0.000)	3.000 (0.000)	4.000 (1.000)	0.240 (0.426)
ecoli4	SMOTE	0.935 (0.071)	0.939 (0.069)	6.800 (0.837)	11.600 (2.966)	0.284 (0.139)
ecoli4	Borderline_- SMOTE	0.890 (0.056)	0.918 (0.071)	5.400 (1.517)	9.600 (3.715)	0.204 (0.250)
ecoli4	ADASYN	0.923 (0.082)	0.924 (0.078)	6.000 (1.732)	11.600 (4.561)	0.211 (0.225)
ecoli4	Safe_Level_- SMOTE	0.899 (0.060)	0.905 (0.064)	7.600 (1.140)	13.400 (2.408)	0.146 (0.227)
ecoli4	TL	0.622 (0.171)	0.622 (0.171)	3.200 (0.837)	5.000 (2.739)	0.254 (0.427)
ecoli4	OSS	0.899 (0.078)	0.884 (0.061)	2.400 (0.548)	4.600 (1.140)	0.105 (0.129)
ecoli4	NCL	0.619 (0.166)	0.619 (0.166)	4.000 (0.707)	6.200 (2.775)	0.437 (0.514)
ecoli4	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
ecoli4	SMOTE_TL	0.927 (0.076)	0.931 (0.078)	7.200 (2.280)	14.000 (3.808)	0.256 (0.154)
ecoli4	SMOTE_- ENN	0.957 (0.030)	0.959 (0.031)	6.800 (1.304)	12.400 (3.435)	0.371 (0.121)

Tabla A.4: Valores medios obtenidos por el método propuesto para la base de datos no balanceada ecoli4.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
glass-0-1-6_-vs_2	None	0.500 (0.000)	0.500 (0.000)	3.000 (0.707)	3.800 (1.304)	0.622 (0.518)
glass-0-1-6_-vs_2	SMOTE	0.575 (0.050)	0.584 (0.058)	5.400 (1.817)	9.600 (4.099)	0.172 (0.207)
glass-0-1-6_-vs_2	Borderline_-SMOTE	0.617 (0.062)	0.620 (0.055)	4.800 (0.447)	9.000 (1.414)	0.203 (0.168)
glass-0-1-6_-vs_2	ADASYN	0.650 (0.060)	0.659 (0.049)	5.000 (1.581)	9.600 (5.683)	0.132 (0.121)
glass-0-1-6_-vs_2	Safe_Level_-SMOTE	0.706 (0.068)	0.714 (0.071)	4.600 (1.140)	6.600 (2.966)	0.209 (0.219)
glass-0-1-6_-vs_2	TL	0.494 (0.008)	0.494 (0.008)	2.400 (0.548)	2.600 (0.894)	0.664 (0.470)
glass-0-1-6_-vs_2	OSS	0.449 (0.055)	0.469 (0.021)	1.600 (0.548)	1.800 (0.837)	0.602 (0.545)
glass-0-1-6_-vs_2	NCL	0.491 (0.008)	0.491 (0.008)	2.600 (0.548)	2.600 (0.548)	0.838 (0.362)
glass-0-1-6_-vs_2	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
glass-0-1-6_-vs_2	SMOTE_TL	0.633 (0.022)	0.641 (0.024)	4.600 (1.140)	8.400 (5.177)	0.153 (0.074)
glass-0-1-6_-vs_2	SMOTE_-ENN	0.694 (0.121)	0.683 (0.105)	6.800 (0.447)	12.200 (2.775)	0.096 (0.110)

Tabla A.5: Valores medios obtenidos por el método propuesto para la base de datos no balanceada glass-0-1-6_vs_2.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
glass-0-1-6_-vs_5	None	0.497 (0.006)	0.497 (0.006)	2.600 (0.894)	2.800 (1.304)	0.077 (0.106)
glass-0-1-6_-vs_5	SMOTE	0.896 (0.127)	0.946 (0.040)	7.200 (2.049)	14.200 (6.535)	0.254 (0.145)
glass-0-1-6_-vs_5	Borderline_-SMOTE	0.827 (0.233)	0.980 (0.016)	7.200 (1.483)	12.600 (2.510)	0.188 (0.087)
glass-0-1-6_-vs_5	ADASYN	0.934 (0.036)	0.934 (0.036)	5.600 (0.548)	9.400 (1.817)	0.195 (0.209)
glass-0-1-6_-vs_5	Safe_Level_-SMOTE	0.826 (0.238)	0.929 (0.023)	6.000 (1.581)	11.800 (2.387)	0.103 (0.133)
glass-0-1-6_-vs_5	TL	0.600 (0.224)	0.647 (0.218)	3.000 (1.000)	3.600 (1.817)	0.202 (0.446)
glass-0-1-6_-vs_5	OSS	0.713 (0.244)	0.780 (0.194)	3.200 (1.095)	5.800 (1.643)	0.115 (0.137)
glass-0-1-6_-vs_5	NCL	0.647 (0.226)	0.697 (0.270)	3.200 (0.837)	3.600 (1.342)	0.223 (0.436)
glass-0-1-6_-vs_5	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
glass-0-1-6_-vs_5	SMOTE_TL	0.946 (0.027)	0.951 (0.030)	7.400 (0.548)	15.600 (3.578)	0.249 (0.081)
glass-0-1-6_-vs_5	SMOTE_-ENN	0.943 (0.045)	0.943 (0.045)	5.600 (1.342)	13.000 (4.848)	0.168 (0.078)

Tabla A.6: Valores medios obtenidos por el método propuesto para la base de datos no balanceada glass-0-1-6-vs_5.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
glass2	None	0.500 (0.000)	0.500 (0.000)	2.600 (0.548)	2.600 (0.548)	0.609 (0.536)
glass2	SMOTE	0.678 (0.039)	0.690 (0.031)	4.800 (0.837)	6.800 (1.924)	0.125 (0.184)
glass2	Borderline_- SMOTE	0.693 (0.059)	0.698 (0.063)	5.400 (0.894)	8.000 (2.121)	0.145 (0.109)
glass2	ADASYN	0.673 (0.045)	0.693 (0.046)	3.600 (0.894)	5.000 (1.414)	0.062 (0.104)
glass2	Safe_Level_- SMOTE	0.731 (0.059)	0.746 (0.061)	5.200 (1.483)	9.000 (3.742)	0.144 (0.197)
glass2	TL	0.497 (0.006)	0.497 (0.006)	2.600 (0.548)	2.600 (0.548)	0.604 (0.542)
glass2	OSS	0.477 (0.019)	0.511 (0.076)	1.600 (0.894)	1.800 (1.304)	0.800 (0.447)
glass2	NCL	0.497 (0.006)	0.497 (0.006)	2.200 (0.447)	2.200 (0.447)	0.438 (0.513)
glass2	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
glass2	SMOTE_TL	0.670 (0.039)	0.688 (0.041)	3.800 (0.837)	4.800 (1.924)	0.049 (0.069)
glass2	SMOTE_- ENN	0.708 (0.036)	0.713 (0.030)	4.400 (0.548)	7.800 (2.490)	0.205 (0.182)

Tabla A.7: Valores medios obtenidos por el método propuesto para la base de datos no balanceada glass2.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
glass4	None	0.640 (0.096)	0.738 (0.148)	5.000 (0.707)	8.800 (1.789)	0.072 (0.096)
glass4	SMOTE	0.905 (0.107)	0.910 (0.103)	6.800 (1.483)	14.800 (2.280)	0.050 (0.070)
glass4	Borderline_- SMOTE	0.942 (0.070)	0.949 (0.075)	6.800 (1.643)	13.600 (3.050)	0.064 (0.079)
glass4	ADASYN	0.908 (0.103)	0.908 (0.109)	8.400 (3.209)	18.800 (7.855)	0.171 (0.065)
glass4	Safe_Level_- SMOTE	0.902 (0.075)	0.935 (0.017)	7.200 (2.490)	18.200 (8.075)	0.124 (0.087)
glass4	TL	0.744 (0.151)	0.823 (0.150)	4.800 (0.837)	10.000 (2.550)	0.080 (0.115)
glass4	OSS	0.834 (0.125)	0.832 (0.129)	3.200 (0.837)	5.800 (1.483)	0.317 (0.410)
glass4	NCL	0.897 (0.105)	0.930 (0.102)	5.600 (1.140)	11.400 (2.608)	0.421 (0.530)
glass4	SBC	0.450 (0.112)	0.450 (0.112)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
glass4	SMOTE_TL	0.913 (0.114)	0.915 (0.114)	6.000 (1.225)	12.200 (1.924)	0.140 (0.139)
glass4	SMOTE_- ENN	0.908 (0.109)	0.905 (0.109)	6.600 (1.140)	13.800 (4.025)	0.133 (0.131)

Tabla A.8: Valores medios obtenidos por el método propuesto para la base de datos no balanceada glass4.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
glass5	None	0.498 (0.005)	0.500 (0.000)	2.200 (0.447)	2.600 (1.342)	0.600 (0.547)
glass5	SMOTE	0.813 (0.218)	0.818 (0.222)	8.200 (2.168)	16.200 (6.907)	0.092 (0.150)
glass5	Borderline_- SMOTE	0.878 (0.123)	0.880 (0.125)	7.800 (1.095)	13.000 (2.345)	0.166 (0.169)
glass5	ADASYN	0.884 (0.088)	0.884 (0.088)	6.200 (1.643)	12.600 (5.505)	0.222 (0.167)
glass5	Safe_Level_- SMOTE	0.805 (0.127)	0.810 (0.131)	6.200 (2.775)	12.200 (7.190)	0.097 (0.107)
glass5	TL	0.498 (0.005)	0.598 (0.218)	2.600 (0.894)	2.800 (1.304)	0.800 (0.447)
glass5	OSS	0.685 (0.260)	0.688 (0.263)	3.400 (1.517)	6.000 (3.808)	0.606 (0.540)
glass5	NCL	0.495 (0.007)	0.548 (0.106)	3.000 (0.707)	3.400 (1.517)	0.600 (0.548)
glass5	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
glass5	SMOTE_TL	0.889 (0.106)	0.891 (0.100)	8.200 (3.347)	16.200 (6.058)	0.192 (0.154)
glass5	SMOTE_- ENN	0.832 (0.209)	0.882 (0.096)	7.600 (1.517)	15.800 (4.604)	0.135 (0.138)

Tabla A.9: Valores medios obtenidos por el método propuesto para la base de datos no balanceada glass5.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
page-blocks-1-3_vs_4	None	0.676 (0.101)	0.692 (0.087)	3.800 (0.837)	5.800 (2.387)	0.029 (0.042)
page-blocks-1-3_vs_4	SMOTE	0.908 (0.051)	0.950 (0.030)	3.600 (1.140)	7.800 (4.207)	0.260 (0.199)
page-blocks-1-3_vs_4	Borderline_-SMOTE	0.980 (0.009)	0.990 (0.005)	5.400 (1.140)	9.600 (1.949)	0.154 (0.134)
page-blocks-1-3_vs_4	ADASYN	0.950 (0.035)	0.959 (0.037)	6.400 (2.191)	11.800 (4.764)	0.347 (0.231)
page-blocks-1-3_vs_4	Safe_Level_-SMOTE	0.871 (0.048)	0.917 (0.027)	3.200 (1.643)	6.200 (1.924)	0.109 (0.052)
page-blocks-1-3_vs_4	TL	0.711 (0.145)	0.716 (0.141)	4.600 (0.548)	6.800 (1.095)	0.043 (0.036)
page-blocks-1-3_vs_4	OSS	0.746 (0.116)	0.812 (0.070)	3.000 (1.000)	4.200 (0.837)	0.128 (0.131)
page-blocks-1-3_vs_4	NCL	0.707 (0.099)	0.739 (0.107)	4.000 (0.707)	6.600 (2.702)	0.203 (0.445)
page-blocks-1-3_vs_4	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
page-blocks-1-3_vs_4	SMOTE_TL	0.945 (0.018)	0.964 (0.016)	3.200 (1.304)	6.200 (2.683)	0.137 (0.106)
page-blocks-1-3_vs_4	SMOTE_-ENN	0.896 (0.060)	0.932 (0.044)	4.000 (1.732)	8.000 (4.743)	0.158 (0.175)

Tabla A.10: Valores medios obtenidos por el método propuesto para la base de datos no balanceada page-blocks-1-3_vs_4.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
shuttle-c0-vs-c4	None	0.992 (0.011)	1.000 (0.000)	3.000 (0.000)	3.000 (0.000)	0.185 (0.265)
shuttle-c0-vs-c4	SMOTE	0.992 (0.011)	1.000 (0.000)	3.600 (0.548)	5.000 (1.225)	0.116 (0.136)
shuttle-c0-vs-c4	Borderline_-SMOTE	0.992 (0.011)	1.000 (0.000)	3.000 (0.000)	3.000 (0.000)	0.044 (0.039)
shuttle-c0-vs-c4	ADASYN	0.986 (0.018)	0.996 (0.006)	2.800 (1.095)	3.400 (1.673)	0.216 (0.178)
shuttle-c0-vs-c4	Safe_Level_-SMOTE	0.992 (0.011)	1.000 (0.000)	3.400 (0.548)	3.600 (0.548)	0.015 (0.017)
shuttle-c0-vs-c4	TL	0.992 (0.011)	1.000 (0.000)	3.000 (0.000)	3.000 (0.000)	0.099 (0.075)
shuttle-c0-vs-c4	OSS	0.959 (0.047)	0.996 (0.008)	2.800 (0.447)	3.200 (0.837)	0.464 (0.496)
shuttle-c0-vs-c4	NCL	0.992 (0.011)	1.000 (0.000)	3.400 (0.548)	3.400 (0.548)	0.232 (0.432)
shuttle-c0-vs-c4	SBC	0.496 (0.009)	0.500 (0.000)	1.600 (0.548)	1.600 (0.548)	0.606 (0.539)
shuttle-c0-vs-c4	SMOTE_TL	0.992 (0.011)	1.000 (0.000)	4.800 (1.095)	6.400 (2.074)	0.061 (0.039)
shuttle-c0-vs-c4	SMOTE_-ENN	0.992 (0.011)	1.000 (0.000)	4.600 (1.517)	6.200 (2.490)	0.369 (0.387)

Tabla A.11: Valores medios obtenidos por el método propuesto para la base de datos no balanceada shuttle-c0-vs-c4.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
shuttle-c2- vs-c4	None	0.892 (0.219)	0.900 (0.224)	2.000 (0.707)	2.000 (0.707)	0.226 (0.433)
shuttle-c2- vs-c4	SMOTE	0.892 (0.219)	0.896 (0.221)	2.800 (1.095)	3.400 (1.140)	0.229 (0.137)
shuttle-c2- vs-c4	Borderline_- SMOTE	0.842 (0.217)	0.950 (0.112)	2.600 (0.548)	2.600 (0.548)	0.129 (0.180)
shuttle-c2- vs-c4	ADASYN	0.888 (0.217)	0.996 (0.009)	3.000 (1.000)	4.600 (1.140)	0.188 (0.190)
shuttle-c2- vs-c4	Safe_Level_- SMOTE	0.818 (0.196)	0.927 (0.101)	3.000 (1.000)	4.000 (1.581)	0.207 (0.123)
shuttle-c2- vs-c4	TL	0.992 (0.011)	1.000 (0.000)	2.400 (0.894)	2.400 (0.894)	0.042 (0.050)
shuttle-c2- vs-c4	OSS	0.822 (0.200)	0.922 (0.097)	2.000 (0.000)	2.000 (0.000)	0.418 (0.532)
shuttle-c2- vs-c4	NCL	0.992 (0.011)	1.000 (0.000)	2.200 (0.447)	2.200 (0.447)	0.150 (0.160)
shuttle-c2- vs-c4	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
shuttle-c2- vs-c4	SMOTE_TL	0.892 (0.219)	0.896 (0.222)	3.200 (0.837)	3.600 (1.342)	0.191 (0.245)
shuttle-c2- vs-c4	SMOTE_- ENN	0.892 (0.219)	1.000 (0.000)	3.600 (0.894)	4.000 (1.414)	0.079 (0.095)

Tabla A.12: Valores medios obtenidos por el método propuesto para la base de datos no balanceada shuttle-c2-vs-c4.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
vowel0	None	0.868 (0.078)	0.889 (0.070)	7.000 (1.581)	11.000 (3.240)	0.076 (0.099)
vowel0	SMOTE	0.934 (0.039)	0.937 (0.037)	9.000 (0.000)	16.800 (0.837)	0.327 (0.139)
vowel0	Borderline_- SMOTE	0.917 (0.012)	0.938 (0.023)	7.000 (3.240)	12.000 (7.348)	0.126 (0.120)
vowel0	ADASYN	0.957 (0.036)	0.975 (0.008)	9.600 (2.966)	21.400 (11.261)	0.153 (0.169)
vowel0	Safe_Level_- SMOTE	0.964 (0.009)	0.965 (0.007)	11.000 (1.000)	23.600 (3.847)	0.191 (0.118)
vowel0	TL	0.838 (0.080)	0.863 (0.056)	6.200 (1.789)	9.200 (4.764)	0.146 (0.138)
vowel0	OSS	0.898 (0.071)	0.902 (0.065)	5.800 (1.789)	10.200 (4.764)	0.078 (0.064)
vowel0	NCL	0.863 (0.070)	0.879 (0.057)	6.000 (1.581)	8.800 (3.962)	0.256 (0.429)
vowel0	SBC	0.751 (0.230)	0.758 (0.236)	4.400 (2.510)	7.600 (5.899)	0.433 (0.521)
vowel0	SMOTE_TL	0.948 (0.037)	0.950 (0.037)	9.600 (2.302)	19.400 (6.066)	0.108 (0.123)
vowel0	SMOTE_- ENN	0.936 (0.056)	0.955 (0.026)	9.800 (1.095)	20.000 (2.449)	0.196 (0.176)

Tabla A.13: Valores medios obtenidos por el método propuesto para la base de datos no balanceada vowel0.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
yeast-0-5-6-7-9_vs_4	None	0.497 (0.005)	0.507 (0.022)	2.200 (1.304)	3.800 (3.271)	0.650 (0.485)
yeast-0-5-6-7-9_vs_4	SMOTE	0.831 (0.035)	0.841 (0.032)	6.600 (1.517)	13.400 (3.362)	0.113 (0.148)
yeast-0-5-6-7-9_vs_4	Borderline_-SMOTE	0.821 (0.050)	0.825 (0.049)	4.400 (0.894)	9.200 (2.280)	0.232 (0.153)
yeast-0-5-6-7-9_vs_4	ADASYN	0.811 (0.051)	0.813 (0.048)	6.000 (0.707)	13.000 (3.162)	0.183 (0.124)
yeast-0-5-6-7-9_vs_4	Safe_Level_-SMOTE	0.816 (0.057)	0.823 (0.057)	4.200 (0.837)	7.600 (2.702)	0.129 (0.149)
yeast-0-5-6-7-9_vs_4	TL	0.563 (0.073)	0.582 (0.082)	4.200 (0.837)	9.000 (3.162)	0.190 (0.143)
yeast-0-5-6-7-9_vs_4	OSS	0.726 (0.155)	0.726 (0.150)	2.600 (1.140)	4.000 (2.550)	0.298 (0.407)
yeast-0-5-6-7-9_vs_4	NCL	0.608 (0.078)	0.682 (0.050)	4.600 (0.548)	10.200 (1.643)	0.130 (0.061)
yeast-0-5-6-7-9_vs_4	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
yeast-0-5-6-7-9_vs_4	SMOTE_TL	0.812 (0.030)	0.834 (0.037)	6.000 (1.581)	12.000 (1.225)	0.106 (0.168)
yeast-0-5-6-7-9_vs_4	SMOTE_-ENN	0.814 (0.052)	0.823 (0.036)	5.600 (1.140)	10.000 (1.225)	0.182 (0.185)

Tabla A.14: Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-0-5-6-7-9_vs_4.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
yeast-1-2-8-9_vs_7	None	0.495 (0.004)	0.496 (0.004)	1.400 (0.548)	1.600 (0.894)	0.607 (0.538)
yeast-1-2-8-9_vs_7	SMOTE	0.697 (0.056)	0.710 (0.065)	7.400 (2.702)	14.200 (6.140)	0.149 (0.105)
yeast-1-2-8-9_vs_7	Borderline_-SMOTE	0.686 (0.076)	0.690 (0.076)	6.200 (1.789)	13.200 (4.764)	0.272 (0.153)
yeast-1-2-8-9_vs_7	ADASYN	0.748 (0.055)	0.758 (0.065)	6.600 (1.817)	15.000 (3.742)	0.163 (0.125)
yeast-1-2-8-9_vs_7	Safe_Level_-SMOTE	0.706 (0.108)	0.736 (0.119)	6.800 (1.643)	12.600 (3.912)	0.097 (0.133)
yeast-1-2-8-9_vs_7	TL	0.497 (0.002)	0.498 (0.002)	1.800 (0.447)	1.800 (0.447)	0.603 (0.543)
yeast-1-2-8-9_vs_7	OSS	0.577 (0.099)	0.578 (0.099)	1.800 (0.837)	2.200 (1.304)	0.607 (0.538)
yeast-1-2-8-9_vs_7	NCL	0.497 (0.002)	0.498 (0.002)	1.800 (0.447)	1.800 (0.447)	0.575 (0.394)
yeast-1-2-8-9_vs_7	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
yeast-1-2-8-9_vs_7	SMOTE_TL	0.717 (0.072)	0.725 (0.067)	7.000 (2.236)	13.800 (4.658)	0.210 (0.251)
yeast-1-2-8-9_vs_7	SMOTE_-ENN	0.697 (0.096)	0.703 (0.107)	6.200 (1.924)	15.200 (5.805)	0.223 (0.123)

Tabla A.15: Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-1-2-8-9_vs_7.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
yeast-1-4-5-8_vs_7	None	0.498 (0.002)	0.498 (0.002)	1.800 (0.837)	1.800 (0.837)	1.000 (0.000)
yeast-1-4-5-8_vs_7	SMOTE	0.660 (0.056)	0.662 (0.055)	5.200 (2.387)	10.200 (5.263)	0.256 (0.164)
yeast-1-4-5-8_vs_7	Borderline_-SMOTE	0.620 (0.069)	0.630 (0.068)	7.000 (3.162)	15.000 (7.176)	0.187 (0.084)
yeast-1-4-5-8_vs_7	ADASYN	0.635 (0.037)	0.640 (0.030)	6.000 (2.000)	11.800 (5.805)	0.363 (0.205)
yeast-1-4-5-8_vs_7	Safe_Level_-SMOTE	0.632 (0.050)	0.636 (0.047)	7.200 (2.168)	15.200 (5.630)	0.237 (0.137)
yeast-1-4-5-8_vs_7	TL	0.497 (0.002)	0.497 (0.002)	1.200 (0.447)	2.200 (2.683)	0.803 (0.441)
yeast-1-4-5-8_vs_7	OSS	0.497 (0.002)	0.497 (0.002)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
yeast-1-4-5-8_vs_7	NCL	0.496 (0.000)	0.496 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
yeast-1-4-5-8_vs_7	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
yeast-1-4-5-8_vs_7	SMOTE_TL	0.627 (0.031)	0.621 (0.027)	7.000 (3.317)	14.600 (7.893)	0.217 (0.125)
yeast-1-4-5-8_vs_7	SMOTE_-ENN	0.641 (0.083)	0.643 (0.083)	8.200 (2.864)	18.800 (8.167)	0.266 (0.154)

Tabla A.16: Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-1-4-5-8_vs_7.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
yeast-1_vs_7	None	0.498 (0.003)	0.498 (0.003)	1.200 (0.447)	1.600 (1.342)	0.807 (0.431)
yeast-1_vs_7	SMOTE	0.640 (0.060)	0.642 (0.060)	8.400 (2.793)	16.200 (7.225)	0.393 (0.165)
yeast-1_vs_7	Borderline_- SMOTE	0.649 (0.075)	0.667 (0.097)	9.000 (1.225)	18.400 (3.286)	0.146 (0.229)
yeast-1_vs_7	ADASYN	0.659 (0.057)	0.684 (0.060)	8.600 (3.130)	17.600 (8.989)	0.228 (0.153)
yeast-1_vs_7	Safe_Level_- SMOTE	0.635 (0.042)	0.650 (0.054)	7.000 (2.739)	13.800 (8.871)	0.390 (0.134)
yeast-1_vs_7	TL	0.499 (0.003)	0.499 (0.003)	1.600 (0.894)	1.800 (1.095)	0.806 (0.434)
yeast-1_vs_7	OSS	0.504 (0.050)	0.542 (0.063)	1.600 (0.894)	2.200 (2.168)	0.638 (0.500)
yeast-1_vs_7	NCL	0.495 (0.008)	0.529 (0.067)	1.800 (0.837)	2.400 (1.342)	0.811 (0.422)
yeast-1_vs_7	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
yeast-1_vs_7	SMOTE_TL	0.632 (0.087)	0.683 (0.032)	8.200 (2.168)	15.400 (4.037)	0.192 (0.184)
yeast-1_vs_7	SMOTE_- ENN	0.647 (0.044)	0.650 (0.045)	7.600 (2.302)	14.800 (6.301)	0.293 (0.164)

Tabla A.17: Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-1_vs_7.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
yeast-2_vs_4	None	0.730 (0.104)	0.806 (0.083)	4.200 (1.095)	8.600 (3.507)	0.037 (0.037)
yeast-2_vs_4	SMOTE	0.903 (0.023)	0.910 (0.027)	5.400 (1.342)	10.600 (2.510)	0.147 (0.125)
yeast-2_vs_4	Borderline_- SMOTE	0.894 (0.031)	0.907 (0.032)	5.000 (1.225)	9.600 (2.881)	0.125 (0.156)
yeast-2_vs_4	ADASYN	0.891 (0.033)	0.892 (0.032)	5.200 (0.837)	8.200 (1.483)	0.280 (0.202)
yeast-2_vs_4	Safe_Level_- SMOTE	0.894 (0.038)	0.915 (0.028)	4.400 (0.894)	8.200 (2.588)	0.050 (0.074)
yeast-2_vs_4	TL	0.777 (0.115)	0.843 (0.081)	4.000 (1.000)	6.800 (2.168)	0.056 (0.056)
yeast-2_vs_4	OSS	0.881 (0.043)	0.888 (0.044)	4.000 (0.707)	7.800 (0.447)	0.109 (0.087)
yeast-2_vs_4	NCL	0.758 (0.111)	0.850 (0.070)	4.000 (0.707)	7.200 (1.483)	0.079 (0.092)
yeast-2_vs_4	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
yeast-2_vs_4	SMOTE_TL	0.905 (0.025)	0.904 (0.023)	5.800 (1.304)	11.000 (3.317)	0.150 (0.103)
yeast-2_vs_4	SMOTE_- ENN	0.910 (0.029)	0.910 (0.022)	5.800 (1.304)	11.400 (2.793)	0.226 (0.232)

Tabla A.18: Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-2_vs_4.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
yeast-2_vs_8	None	0.499 (0.002)	0.550 (0.112)	1.800 (0.447)	2.000 (0.707)	0.675 (0.457)
yeast-2_vs_8	SMOTE	0.749 (0.152)	0.809 (0.122)	2.600 (0.894)	4.600 (2.074)	0.001 (0.001)
yeast-2_vs_8	Borderline_- SMOTE	0.570 (0.108)	0.747 (0.085)	3.800 (1.483)	6.200 (4.087)	0.097 (0.072)
yeast-2_vs_8	ADASYN	0.564 (0.059)	0.619 (0.104)	9.000 (2.345)	20.000 (5.612)	0.187 (0.166)
yeast-2_vs_8	Safe_Level_- SMOTE	0.749 (0.088)	0.764 (0.103)	2.600 (0.548)	4.800 (2.280)	0.095 (0.105)
yeast-2_vs_8	TL	0.499 (0.002)	0.499 (0.002)	1.600 (0.548)	1.600 (0.548)	1.000 (0.000)
yeast-2_vs_8	OSS	0.599 (0.106)	0.746 (0.089)	3.000 (0.707)	4.000 (1.871)	0.006 (0.008)
yeast-2_vs_8	NCL	0.499 (0.002)	0.499 (0.002)	1.600 (0.548)	1.600 (0.548)	1.000 (0.000)
yeast-2_vs_8	SBC	0.549 (0.111)	0.530 (0.067)	1.200 (0.447)	1.400 (0.894)	0.801 (0.444)
yeast-2_vs_8	SMOTE_TL	0.786 (0.103)	0.788 (0.104)	2.800 (0.837)	5.600 (3.286)	0.123 (0.152)
yeast-2_vs_8	SMOTE_- ENN	0.791 (0.106)	0.810 (0.105)	2.800 (0.447)	4.200 (1.643)	0.041 (0.067)

Tabla A.19: Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast-2_vs_8.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
yeast4	None	0.500 (0.001)	0.500 (0.000)	2.800 (0.447)	4.400 (0.894)	0.222 (0.435)
yeast4	SMOTE	0.853 (0.025)	0.854 (0.023)	7.000 (2.739)	13.800 (9.365)	0.101 (0.107)
yeast4	Borderline_- SMOTE	0.856 (0.030)	0.858 (0.029)	5.800 (1.483)	12.000 (2.739)	0.133 (0.115)
yeast4	ADASYN	0.842 (0.031)	0.854 (0.032)	7.400 (1.140)	15.800 (4.919)	0.205 (0.127)
yeast4	Safe_Level_- SMOTE	0.848 (0.028)	0.857 (0.022)	4.800 (0.837)	8.800 (2.775)	0.240 (0.179)
yeast4	TL	0.499 (0.001)	0.500 (0.001)	2.800 (0.447)	4.400 (0.894)	0.046 (0.036)
yeast4	OSS	0.498 (0.002)	0.498 (0.002)	1.400 (0.548)	1.600 (0.894)	0.801 (0.444)
yeast4	NCL	0.498 (0.003)	0.500 (0.001)	3.200 (0.447)	4.800 (0.837)	0.261 (0.414)
yeast4	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
yeast4	SMOTE_TL	0.851 (0.024)	0.859 (0.022)	7.200 (2.588)	13.200 (5.119)	0.224 (0.205)
yeast4	SMOTE_- ENN	0.848 (0.029)	0.849 (0.028)	7.200 (1.643)	14.000 (3.391)	0.204 (0.200)

Tabla A.20: Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast4.

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
yeast5	None	0.500 (0.000)	0.500 (0.000)	4.200 (0.837)	5.000 (1.000)	0.417 (0.532)
yeast5	SMOTE	0.954 (0.032)	0.954 (0.032)	5.000 (1.000)	9.200 (2.280)	0.202 (0.171)
yeast5	Borderline_- SMOTE	0.952 (0.034)	0.965 (0.011)	4.600 (0.548)	7.600 (1.817)	0.162 (0.099)
yeast5	ADASYN	0.949 (0.033)	0.964 (0.009)	5.000 (1.225)	9.000 (3.464)	0.220 (0.225)
yeast5	Safe_Level_- SMOTE	0.957 (0.014)	0.959 (0.014)	4.200 (0.447)	6.800 (1.304)	0.078 (0.029)
yeast5	TL	0.500 (0.000)	0.500 (0.000)	4.600 (0.548)	5.800 (0.837)	0.802 (0.444)
yeast5	OSS	0.891 (0.025)	0.924 (0.022)	3.000 (0.707)	4.400 (1.673)	0.029 (0.050)
yeast5	NCL	0.652 (0.104)	0.707 (0.155)	6.000 (0.707)	9.800 (2.280)	0.237 (0.428)
yeast5	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	0.800 (0.447)
yeast5	SMOTE_TL	0.951 (0.034)	0.964 (0.009)	4.200 (0.447)	7.000 (1.732)	0.085 (0.088)
yeast5	SMOTE_- ENN	0.959 (0.016)	0.964 (0.013)	4.200 (0.447)	7.200 (1.924)	0.187 (0.187)

Tabla A.21: Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast5.

A. RESULTADOS OBTENIDOS CON EL MÉTODO PROPUESTO

BDNB	MB	AUC_BR	AUC_BD	# Regl.	# Cond.	IIS
yeast6	None	0.500 (0.000)	0.500 (0.000)	3.000 (0.000)	3.400 (0.548)	0.805 (0.435)
yeast6	SMOTE	0.902 (0.070)	0.905 (0.069)	9.200 (1.095)	18.400 (2.302)	0.336 (0.185)
yeast6	Borderline_- SMOTE	0.914 (0.069)	0.917 (0.072)	7.400 (0.548)	12.600 (2.191)	0.177 (0.147)
yeast6	ADASYN	0.893 (0.065)	0.896 (0.067)	7.000 (1.225)	12.400 (2.510)	0.175 (0.077)
yeast6	Safe_Level_- SMOTE	0.893 (0.071)	0.901 (0.064)	6.800 (1.095)	11.400 (2.510)	0.209 (0.192)
yeast6	TL	0.500 (0.000)	0.500 (0.000)	3.000 (0.000)	3.400 (0.548)	0.848 (0.339)
yeast6	OSS	0.636 (0.156)	0.620 (0.146)	2.400 (1.342)	4.200 (3.347)	0.509 (0.458)
yeast6	NCL	0.500 (0.000)	0.500 (0.000)	3.000 (0.000)	3.800 (0.447)	1.000 (0.000)
yeast6	SBC	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
yeast6	SMOTE_TL	0.887 (0.070)	0.897 (0.063)	7.400 (1.949)	13.400 (3.362)	0.175 (0.189)
yeast6	SMOTE_- ENN	0.903 (0.067)	0.905 (0.065)	7.000 (1.414)	12.400 (3.209)	0.162 (0.167)

Tabla A.22: Valores medios obtenidos por el método propuesto para la base de datos no balanceada yeast6.

Referencias

- Abadeh, M. S. and Habibi, J. (2007). Computer intrusion detection using an iterative fuzzy rule learning approach. In *FUZZ-IEEE*, pages 1–6. IEEE. 2, 107
- Alcala, R., Gacto, M., and Herrera, F. (2011). A fast and scalable multiobjective genetic fuzzy system for linguistic fuzzy modeling in high-dimensional regression problems. *Fuzzy Systems, IEEE Transactions on*, 19(4):666–681. 90
- Alcalá, R., Nojima, Y., Herrera, F., and Ishibuchi, H. (2011). Multiobjective genetic fuzzy rule selection of single granularity-based fuzzy classification rules and its interaction with the lateral tuning of membership functions. *Soft Computing*, 15(12):2303–2318. 82
- Alcalá-Fdez, J., Fernandez, A., J. Luengo, J. D., S. García, L. S., and Herrera, F. (2011). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2-3):255 – 287. 120
- Alshomrani, S., Bawakid, A., Shim, S.-O., Fernández, A., and Herrera, F. (2015). A proposal for evolutionary fuzzy systems using feature weighting: Dealing with overlapping in imbalanced datasets. *Knowledge-Based Systems*, 73:1 – 17. 101
- Antonelli, M., Ducange, P., Lazzerini, B., and Marcelloni, F. (2011a). Learning concurrently data and rule bases of mamdani fuzzy rule-based systems by exploiting a novel interpretability index. *Soft Computing*, 15(10):1981–1998. 89

REFERENCIAS

- Antonelli, M., Ducange, P., Lazzerini, B., and Marcelloni, F. (2011b). Learning knowledge bases of multi-objective evolutionary fuzzy systems by simultaneously optimizing accuracy, complexity and partition integrity. *Soft Computing*, 15(12):2335–2354. 89
- Antonelli, M., Ducange, P., and Marcelloni, F. (2014a). An experimental study on evolutionary fuzzy classifiers designed for managing imbalanced datasets. *Neurocomputing*, 146(0):125 – 136. Bridging Machine learning and Evolutionary Computation (BMLEC) Computational Collective Intelligence. 100
- Antonelli, M., Ducange, P., and Marcelloni, F. (2014b). A fast and efficient multi-objective evolutionary learning scheme for fuzzy rule-based classifiers. *Inf. Sci.*, 283:36–54. 94
- Antonelli, M., Ducange, P., Marcelloni, F., and Segatori, A. (2013). Evolutionary fuzzy classifiers for imbalanced datasets: An experimental comparison. In *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint*, pages 13–18. xiii, 92, 93, 98, 120
- Batista, G., Prati, R., and Monard, M. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6(1):20–29. xi, xii, 13, 26, 27
- Berlanga, F. J., Rivera, A. J., del Jesus, M. J., and Herrera, F. (2010). Gp-coach: Genetic programming-based learning of compact and accurate fuzzy rule-based classification systems for high-dimensional problems. *Inf. Sci.*, 180(8):1183–1200. 99
- Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD09)*, volume 5476 of *Lecture Notes on Computer Science*, pages 475–482. Springer-Verlag. 17, 18
- Cannone, R., Alonso, J., and Magdalena, L. (2011). Multi-objective design of highly interpretable fuzzy rule-based classifiers with semantic cointension. In

- Genetic and Evolutionary Fuzzy Systems (GEFS), 2011 IEEE 5th International Workshop on*, pages 1–8. 91
- Cao, H., Li, X.-L., Woon, D.-K., and Ng, S.-K. (2013). Integrated oversampling for imbalanced time series classification. *Knowledge and Data Engineering, IEEE Transactions on*, 25(12):2809–2822. xi, 10, 15
- Cardenas, E. and Carmago, H. (2012). Multiobjective genetic optimization of fuzzy partitions and t-norm parameters in fuzzy classifiers. In *Neural Networks (SBRN), 2012 Brazilian Symposium on*, pages 154–159. 84
- Casillas, J. and Carse, B. (2009). Special issue on "genetic fuzzy systems: Recent developments and future directions". *Soft Comput.*, 13(5):417–418. 3
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357. xv, 10, 11, 12
- Chen, C.-H., He, J.-S., and Hong, T.-P. (2013). A two-stage multi-objective genetic-fuzzy mining algorithm. In *Genetic and Evolutionary Fuzzy Systems (GEFS), 2013 IEEE International Workshop on*, pages 16–20. 92
- Chi, Z., Yan, H., and Pham, T. (1996). *Fuzzy algorithms : with applications to image processing and pattern recognition*. World Scientific Singapore ; River Edge, N.J. 97
- Cococcioni, M., Ducange, P., Lazzerini, B., and Marcelloni, F. (2007). A pareto-based multi-objective evolutionary approach to the identification of mamdani fuzzy systems. *Soft Computing*, 11(11):1013–1031. 87
- Coello Coello, C. A., Lamont, G. B., and Van Veldhuizen, D. A. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation)*. Springer, 2nd edition. 46
- Cordón, O. (2011). A historical review of evolutionary learning methods for mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems. *International Journal of Approximate Reasoning*, 52(6):894 – 913. 74

REFERENCIAS

- Cordon, O., Gomide, F. A. C., Herrera, F., Hoffmann, F., and Magdalena, L. (2004). Ten years of genetic fuzzy systems: current framework and new trends. *Fuzzy Sets and Systems*, 141(1):5–31. 66
- de Oliveira, D. V. R. (2009). *Análise comparativa de algoritmos de seleção de protótipos em bases desbalanceadas*. Centro de Informática. xi, 21, 22
- Deb, K., Agrawal, R. B., and Agrawal, R. B. (1995). Simulated Binary Crossover for Continuous Search Space. 39, 41
- Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002a). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Trans. Evol. Comp*, 6(2):182–197. 2
- Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002b). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Trans. Evol. Comp*, 6(2):182–197. xii, 46, 47, 48
- Deb, K. and Tiwari, S. (2008). Omni-optimizer: A generic evolutionary algorithm for single and multi-objective optimization. *European Journal of Operational Research*, 185(3):1062 – 1087. 41
- Ducange, P., Lazzerini, B., and Marcelloni, F. (2010a). Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets. *Soft Computing*, 14(7):713–728. 88
- Ducange, P., Lazzerini, B., and Marcelloni, F. (2010b). Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets. *Soft Computing*, 14(7):713–728. 93
- Eiben, A. and Smith, J. (2003). *Introduction to Evolutionary Computing*. Natural Computing Series. Springer. xii, xv, 31, 32, 33, 34
- Fazzolari, M., Alcalá, R., Nojima, Y., Ishibuchi, H., and Herrera, F. (2013a). Improving a fuzzy association rule-based classification model by granularity learning based on heuristic measures over multiple granularities. In *Genetic and Evolutionary Fuzzy Systems (GEFS), 2013 IEEE International Workshop on*, pages 44–51. 87

- Fazzolari, M., Alcalá, R., Nojima, Y., Ishibuchi, H., and Herrera, F. (2013b). A review of the application of multiobjective evolutionary fuzzy systems: Current status and further directions. *Fuzzy Systems, IEEE Transactions on*, 21(1):45–65. xiii, 3, 72, 73
- Fernandez, A., del Jesus, M., and Herrera, F. (2008). Sistemas basados en reglas difusas en clasificación: Nuevos retos. In *XIV Congreso Español sobre Tecnologías y Lógica Fuzzy(ESTYLF08)*, pages 493–500. 7
- Fernández, A., del Jesus, M. J., and Herrera, F. (2010). On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets. *Information Sciences*, 180(8):1268 – 1291. 96
- Fonseca, C. M. and Fleming, P. J. (1993). Genetic algorithms for multiobjective optimization: Formulation discussion and generalization. *ICGA*, 93:416–423. 2
- Gacto, M., Alcalá, R., and Herrera, F. (2009a). Adaptation and application of multi-objective evolutionary algorithms for rule reduction and parameter tuning of fuzzy rule-based systems. *Soft Computing*, 13(5):419–436. 70
- Gacto, M., Alcalá, R., and Herrera, F. (2009b). Adaptation and application of multi-objective evolutionary algorithms for rule reduction and parameter tuning of fuzzy rule-based systems. *Soft Computing*, 13(5):419–436. 76
- Gacto, M., Alcalá, R., and Herrera, F. (2009c). Adaptation and application of multi-objective evolutionary algorithms for rule reduction and parameter tuning of fuzzy rule-based systems. *Soft Computing*, 13(5):419–436. 78
- Gacto, M., Alcalá, R., and Herrera, F. (2010). Integration of an index to preserve the semantic interpretability in the multiobjective evolutionary rule selection and tuning of linguistic fuzzy systems. *Fuzzy Systems, IEEE Transactions on*, 18(3):515–531. xiii, 78, 79, 114, 115, 116
- Gacto, M., Galende, M., Alcalá, R., and Herrera, F. (2013). Obtaining accurate task fuzzy rule-based systems by multi-objective evolutionary learning in high-dimensional regression problems. In *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, pages 1–7. 91

REFERENCIAS

- Gaffer, S., Yahia, M., and Ragab, K. (2012). Genetic fuzzy system for intrusion detection: Analysis of improving of multiclass classification accuracy using kddcup-99 imbalance dataset. In *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on*, pages 318–323. 96
- Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4):42–47. 7, 9, 20
- Ganji, M. F., Abadeh, M. S., Hedayati, M., and Bakhtiari, N. (2010). Fuzzy classification of imbalanced data sets for medical diagnosis. In *Iranian Conference of Biomedical Engineering*. 6
- Gonzalez, A. and Perez, R. (1999). Slave: A genetic learning system based on an iterative approach. *IEEE Transactions on Fuzzy Systems*, 7:176–191. 2, 68, 108
- Gonzalez, A. and Perez, R. (2009). Improving the genetic algorithm of slave. *Mathware and Soft Computing*, 16(1):59–70. 68, 107
- Greene, D. and Smith, S. (1993). Competition-based induction of decision models from examples. *Machine Learning*, 13(2-3):229–257. 70
- Hamdan, M. (2012). On the disruption-level of polynomial mutation for evolutionary multi-objective optimisation algorithms. *COMPUTING AND INFORMATION SCIENCES*, 29(5). 41
- Han, H., Wang, W., and Mao, B. (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *2005 International Conference on Intelligent Computing(ICIC05)*, volume 3644 of *Lecture Notes on Computer Science*, pages 878–887. Springer-Verlag. xi, 13, 15
- He, H., Bai, Y., Garcia, E., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 International Joint Conference on Neural Networks(IJCNN08)*, pages 1322–1328. xi, 14, 16
- He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284. xi, 14

- Herrera, F. (2005). Genetic fuzzy systems: Status, critical considerations and future directions. *International Journal of Computational Intelligence Research*, 1(1):59–67. 50, 64, 66
- Herrera, F. (2008). Genetic fuzzy systems: taxonomy, current research trends and prospects. *Evolutionary Intelligence*, 1(1):27–46. xiii, 50, 66, 67
- Herrera, F., Martínez, L., and Herrera, F. (1999). A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on Fuzzy Systems*, 8:746–752. 130
- Hühn, J. and Hüllermeier, E. (2009). Furia: an algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, 19(3):293–319. 100
- Hinojosa, E. and Camargo, H. (2011). Geração Genética Multiobjetivo de Sistemas Fuzzy Usando a Abordagem Iterativa. Master’s thesis, Universidade Federal de São Carlos - UFSCAR, Brazil. 107
- Hinojosa Cardenas, E. and Camargo, H. (2013). Multi-objective iterative genetic approach for learning fuzzy classification rules with semantic-based selection of the best rule. In *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint*, pages 292–297. 88
- Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI. 35
- Holland, J. H. and Reitman, J. S. (1977). Cognitive systems based on adaptive algorithms. *SIGART Bull.*, (63):49–49. 68
- Huang, J. and Ling, C. (2005). Using auc and accuracy in evaluating learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, 17(3):299–310. 28
- Ishibuchi, H. (2007). Multiobjective genetic fuzzy systems: Review and future research directions. In *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*, pages 1–6. xiii, 75, 83

REFERENCIAS

- Ishibuchi, H., Murata, T., and Türken, I. (1997). Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems. *Fuzzy Sets and Systems*, 89(2):135 – 150. 74
- Ishibuchi, H. and Nojima, Y. (2006). Evolutionary multiobjective optimization for the design of fuzzy rule-based ensemble classifiers. *Int. J. Hybrid Intell. Syst.*, 3(3):129–145. 85
- Ishibuchi, H. and Nojima, Y. (2007). Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning. *International Journal of Approximate Reasoning*, 44:4–31. 85
- Jang, J.-S. R. and Sun, C.-T. (1997). *Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 59
- Karr, C. (1991). Genetic algorithms for fuzzy controllers. *AI Expert*, 6(2):26–33. 70
- Klir, G. J. and Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 50
- Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *14th International Conference on Machine Learning(ICML97)*, pages 179–186. xi, 21, 22
- Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In *8th Conference on AI in Medicine in Europe(AIME01)*, volume 2001 of *Lecture Notes on Computer Science*, pages 63–66. Springer Berlin / Heidelberg. xii, 23, 24, 25, 26
- Leon, J. A. (2005). Metodología para la detección de requerimientos subjetivos en el diseño de producto. In *Tesis Doctoral, Universidad Politecnica de Catalunya, Espana*, pages 18–106. 50
- Liu, Y.-H. and Chen, Y.-T. (2007). Face recognition using total margin-based adaptive fuzzy support vector machines. *IEEE Transactions on Neural Networks*, 18(1):178–192. 6

- Longadge, R., Dongre, S., and Malik, L. (2013). Class imbalance problem in data mining review. *International Journal of Computer Science and Network*, 2(1). xi, 9, 20
- López, V., del Río, S., Benítez, J. M., and Herrera, F. (2015). Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data. *Fuzzy Sets and Systems*, 258:5 – 38. Special issue: Uncertainty in Learning from Big Data. 102
- López, V., Fernández, A., del Jesus, M. J., and Herrera, F. (2013). A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets. *Knowledge-Based Systems*, 38(0):85 – 104. Special Issue on Advances in Fuzzy Knowledge Systems: Theory and Application. 93, 99, 120, 126
- López, V., Fernández, A., and Herrera, F. (2010). A first approach for cost-sensitive classification with linguistic genetic fuzzy systems in imbalanced data-sets. In *Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on*, pages 676–681. 95
- Lopez, V., Fernandez, A., and Herrera, F. (2013). Addressing covariate shift for genetic fuzzy systems classifiers: A case of study with farc-hd for imbalanced datasets. In *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, pages 1–8. 98
- Lopez, V., Fernandez, A., Moreno-Torres, J. G., and Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. *Expert Syst. Appl.*, 39(7):6585–6608. 7
- Luke, S. (2013). *Essentials of Metaheuristics*. Lulu, second edition. xii, 35, 36, 37, 38, 39, 41, 43, 44
- Mahdizadeh, M. and Eftekhari, M. (2013). Designing fuzzy imbalanced classifier based on the subtractive clustering and genetic programming. In *Fuzzy Systems (IFSC), 2013 13th Iranian Conference on*, pages 1–6. 97

REFERENCIAS

- Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60. 88, 122
- Marquez, A., Marquez, F., and Peregrin, A. (2010). A multi-objective evolutionary algorithm with an interpretability improvement mechanism for linguistic fuzzy systems with adaptive defuzzification. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pages 1–7. 83
- Marquez, A., Marquez, F., and Peregrin, A. (2012). An efficient multi-objective evolutionary adaptive conjunction for high dimensional problems in linguistic fuzzy modelling. In *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, pages 1–8. 84
- Márquez, A. A., Márquez, F. A., and Peregrín, A. (2009). Rule base and inference system cooperative learning of mamdani fuzzy systems with multiobjective genetic algorithms. In *European Society for Fuzzy Logic and Technology*, pages 1045–1050. 82
- Morales, V. L. (2014). *Sistemas de clasificación basados en reglas difusas para problemas no balanceados. Aproximaciones y uso de nuevas estrategias para resolver problemas intrínsecos a los datos no balanceados*. PhD thesis, Ciencias de la Computación e Inteligencia Artificial (Universidad de Granada). 101
- Nicoletti, M. and Camargo, H. (2004). Fundamentos da teoria de conjuntos fuzzy. In *EdUFSCar*. 50
- Nojima, Y. and Ishibuchi, H. (2013). Multiobjective genetic fuzzy rule selection with fuzzy relational rules. In *Genetic and Evolutionary Fuzzy Systems (GEFS), 2013 IEEE International Workshop on*, pages 60–67. 86
- Nuovo, A. G. D. and Catania, V. (2009). Linguistic modifiers to improve the accuracy-interpretability trade-off in multi-objective genetic design of fuzzy rule based classifier systems. In *ISDA*, pages 128–133. IEEE Computer Society. 2

- Palacios, A., Sanchez, L., and Couso, I. (2010). Preprocessing vague imbalanced datasets and its use in genetic fuzzy classifiers. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pages 1–8. 95
- Palacios, A., Trawiński, K., Cordón, O., and Sánchez, L. (2014). Cost-sensitive learning of fuzzy rules for imbalanced classification problems using furia. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 22(05):643–675. 100
- Pareto, V. (1896a). *Cours d'Économie Politique*. Droz, Genève. 2
- Pareto, V. (1896b). *Cours d'économie politique professé à l'Université de Lausanne*. 45
- Pedrycz, W. and Gomide, F. (1998). *An Introduction to Fuzzy Sets*. MIT Press. 50
- Pulkkinen, P., Hytönen, J., and Koivisto, H. (2008). Developing a bioaerosol detector using hybrid genetic fuzzy systems. *Engineering Applications of Artificial Intelligence*, 21(8):1330 – 1346. 76
- Ramyachitra, D. and Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research*, 5(4). xi, 7, 8, 9
- Rey, M., Galende, M., Sainz, G., and Fuente, M. (2013). Selection of rules by orthogonal transformations and genetic algorithms to improve the interpretability in fuzzy rule based systems. In *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, pages 1–8. 86
- Silverman, B. W. and Jones, M. C. (1989). E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). *International Statistical Review / Revue Internationale de Statistique*, 57(3):233–238. 22
- Smith, S. F. (1980). *A Learning System Based on Genetic Adaptive Algorithms*. PhD thesis, Pittsburgh, PA, USA. AAI8112638. 68

REFERENCIAS

- Srinivas, N. and Deb, K. (1994). Multiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation*, 2:221–248. 2
- Tomek, I. (1976). Two modifications of cnn. *IEEE Transactions on Systems and Man and Cybernetics*, 6:769–772. 20
- Tsai, C.-H., Chang, L.-C., and Chiang, H.-C. (2009). Forecasting of ozone episode days by cost-sensitive neural network methods. *Sci Total Environ*, 407(6):2124–35. 6
- VILLAR, P., FERNÁNDEZ, A., CARRASCO, R. A., and HERRERA, F. (2012). Feature selection and granularity learning in genetic fuzzy rule-based classification systems for highly imbalanced data-sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 20(03):369–397. 92, 97, 120
- Villar, P., Fernández, A., and Herrera, F. (2011). Studying the behavior of a multiobjective genetic algorithm to design fuzzy rule-based classification systems for imbalanced data-sets. In *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, pages 1239–1246. 91
- Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83. 88, 122
- Williams, D. P., Myers, V., and Silvious, M. S. (2009). Mine Classification With Imbalanced Data. 6(3):528–532+. 6
- Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *Systems, Man and Cybernetics, IEEE Transactions on*, 2(3):408–421. 27
- Yao, X. (1999). *Evolutionary Computation: Theory and Applications*. World Scientific. 31
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353. 49