

Índice

Capítulo I: Introducción	1
1.1 Contextualización	2
1.2 Problemática	6
1.3 Objetivos.....	8
1.3.1 Objetivo general	8
1.3.2 Objetivos específicos.....	8
1.4 Resultados esperados	9
1.5 Estructura de la Tesis	10
Capítulo II: Aprendizaje Automático.....	11
2.1 Descubrimiento de conocimiento en bases de datos	12
2.1.1 Proceso KDD.....	14
2.2 Clasificación.....	18
2.2.1 Métodos para la minimización del riesgo	24
2.2.1.1 Modelos de Optimización.....	24
2.2.1.2 Métodos de descenso por gradiente instantáneo.....	25
2.2.1.3 Método de retro propagación (Back-propagation).....	26
2.2.2 Medidas de rendimiento	26
Capítulo III: Desbalance de Clases y Traslape en bases de datos.....	30
3.1 El Problema de las clases desbalanceadas	31
3.1.1 Métodos de solución a nivel de datos	32
3.1.1.1 Undersampling.....	33
3.1.1.2 Oversampling.....	34
3.1.1.3 SMOTE	35
3.1.2 Métodos de solución a nivel algorítmico	37
3.1.2.1 Aprendizaje sensible al costo	37
3.1.2.2 Múltiples clasificadores	38
3.2 Traslape de clases en la base de datos.....	40
Capítulo IV: Enfoque de Clasificación Discriminante	43
4.1 Naive Bayes.....	44
4.2 Support Vector Machines	46
4.2.1 SVM lineal: Caso separable	47
4.2.2 SVM lineal: Caso no separable.....	52
4.2.3 SVM No Lineal: Inclusión de funciones Kernel	53
Capítulo V: Enfoque de Clasificación basado en Reconocimiento	58
5.1 Clasificación de una clase (OCC).....	59
5.2 Support Vector Data Description (SVDD).....	62
5.2.1 SVDD Lineal	63
5.2.2 SVDD con ejemplos negativos.....	66
5.2.3 SVDD No Lineal	68
5.3 Estimación de densidad de Parzen.....	69

Capítulo VI: Metodología de Clasificación para Desbalance y Traslapo de Clases.....	71
6.1 Definición.....	72
6.2 Justificación	73
6.3 Diseño.....	76
6.3.1 Estrategias de combinación	77
6.3.2 Estimación de las probabilidades a posteriori	81
Capítulo VII: Aplicación de la metodología y comparación de resultados.....	83
7.1 Aplicación sobre base artificial.....	84
7.1.1 Descripción de la base de datos	84
7.1.2 Resultados de los modelos individuales	87
7.1.2.1 Support vector machines	87
7.1.2.2 Support Vector Data Description	91
7.1.2.3 Parzen.....	94
7.1.2.4 Resumen modelos individuales	95
7.1.3 Resultados de la combinación de modelos.....	97
7.1.3.1 Diversidad de los modelos de clasificación	97
7.1.3.2 Reglas fijas de combinación	99
7.1.3.3 Stacking	100
7.2 Aplicación de la metodología al problema de Predicción de Fugas de Clientes.....	103
7.2.1 Descripción de la base de datos y selección de atributos	104
7.2.2 Resultados de los modelos individuales	106
7.2.2.1 Support Vector Machines.....	106
7.2.2.2 Support Vector Data Description	107
7.2.2.3 Parzen.....	108
7.2.3 Resultados de la combinación de modelos.....	109
7.2.3.1 Diversidad de modelos de clasificación	109
7.2.3.2 Reglas fijas de combinación	110
7.2.3.3 Stacking	111
7.2.4 Aplicación de undersampling y oversampling a la base de datos	112
Capítulo VIII: Conclusiones	118
8.1 Conclusiones generales	119
8.2 Trabajo futuro.....	123
Referencias Bibliográficas	125
Anexos.	130
Anexo 1: Condiciones de Mercer	131
Anexo 2: Resultados métodos individuales sobre bases artificiales	132
Anexo 3: Medidas de diversidad	133
Anexo 4: Selección de atributos base fuga.....	134
Anexo 5: Resultado métodos individuales base fuga.....	134
Anexo 6: Resultados métodos individuales (base fuga muestreo)	135
Anexo 7: Diversidad de ensembles de clasificadores (base fuga muestreo).	136

Anexo 8: Resultados reglas fijas de combinación (base fuga muestreo) 137

Anexo 9: Resultados método ‘Stacking’ (base fuga muestreo) 138

Ilustraciones

Ilustración 2.1 Proceso KDD	14
Ilustración 2.2 Función clasificadora convencional	20
Ilustración 2.3 Trade-off Sesgo-Varianza en función de la complejidad del modelo	23
Ilustración 2.4 Método de descenso por gradiente instantáneo	25
Ilustración 3.1 Undersampling de la clase mayoritaria	33
Ilustración 3.2 Oversampling de la clase minoritaria.....	34
Ilustración 3.3 Traslape de clases en un conjunto de datos	41
Ilustración 3.4 Problema de decisión en el contexto de desbalance y traslape de clases	42
Ilustración 4.1 Hiperplano óptimo de separación HOS en SVM	46
Ilustración 4.2 Hiperplanos canónicos en SVM.....	49
Ilustración 4.3 Ejemplo de SVM no lineal.....	53
Ilustración 5.1 Descripción de un conjunto de datos	60
Ilustración 5.2 Support Vector Data Description	63
Ilustración 6.1 Motivo representacional para la generación de ensembles	74
Ilustración 6.2 Arquitectura global de la metodología propuesta	77
Ilustración 6.3 Estrategias de combinación de clasificadores.....	78
Ilustración 7.1 Bases de datos artificiales	86
Ilustración 7.2 Evolución acierto respecto a parámetro C, SVM lineal	88
Ilustración 7.3 Evolución acierto respecto a nivel de rechazo de positivos, SVDD lineal.....	92
Ilustración 7.4 Acierto de métodos individuales sobre conjunto de test	96
Ilustración 7.5 Acierto sobre conjunto de test de reglas fijas de combinación	100
Ilustración 7.6 Acierto sobre conjunto de test de métodos individuales y ensembles	102

Tablas

Tabla 2.1 Matriz de confusión para un problema de dos clases.....	26
Tabla 3.1 Matriz de Costos	37
Tabla 6.1 Reglas fijas de combinación	79
Tabla 6.2 Ejemplo regla de combinación ‘mínimo’.....	79
Tabla 7.1 Acierto SVM lineal sobre conjunto de test, base artificial	89
Tabla 7.2 Acierto SVM polinomial sobre conjunto de test, base artificial	89
Tabla 7.3 Acierto SVM RBF sobre conjunto de test, base artificial	90
Tabla 7.4 Acierto SVM gaussiano sobre conjunto de test, base artificial	91
Tabla 7.5 Acierto SVDD lineal sobre conjunto de test, base artificial	93
Tabla 7.6 Acierto SVDD polinomial sobre conjunto de test, base artificial	93
Tabla 7.7 Acierto SVDD RBF sobre conjunto de test, base artificial	94
Tabla 7.8 Acierto método de densidad de parzen sobre conjunto de test, base artificial.....	95
Tabla 7.9 Resumen acierto de métodos individuales (g_mean), base artificial.....	97
Tabla 7.10 Medidas de diversidad para conjunto de clasificadores, base artificial	98
Tabla 7.11 Acierto de reglas fijas de combinación (g_mean), base artificial.....	99
Tabla 7.12 Acierto método Stacking Naive Bayes, base artificial.....	101
Tabla 7.13 Acierto método Stacking SVM RBF, base artificial	102
Tabla 7.14 Resumen acierto (g_mean) métodos individuales y ensembles, base artificial	103

Tabla 7.15 Descripción de atributos para el problema de predicción de fugas de clientes.....	105
Tabla 7.16 Acierto métodos individuales	109
Tabla 7.17 Medidas de diversidad para conjuntos de clasificadores.....	109
Tabla 7.18 Acierto de reglas fijas de combinación.....	110
Tabla 7.19 Acierto método Stacking	111
Tabla 7.20 Balance y traslape de clases base de datos fuga, muestreo	113
Tabla 7.21 Acierto métodos individuales base de datos fuga (g_mean), muestreo	113
Tabla 7.22 Acierto reglas fijas de combinación base de datos 'fuga' (g-mean), muestreo	114
Tabla 7.23 Acierto método Stacking Naive Bayes base de datos fuga, muestreo	115
Tabla 7.24 Acierto método Stacking SVM RBF base de datos fuga, muestreo	116
Tabla 7.25 Ganancia métodos de ensembles sobre máximo rendimiento individual.....	116
Tabla 7.26 Resumen acierto (g_mean) métodos individuales y ensembles, base fuga.....	117