UNIVERSIDAD DE JAÉN
Vicerrectorado de Investigación, Desarrollo Tecnológico e Innovación

ACCIÓN 6

SOLICITUD PRIMEROS PROYECTOS DE INVESTIGACIÓN
(Plan de Apoyo a la Investigación, Desarrollo Tecnológico e Innovación 2014-2015

II. *Programa de Apoyo a Actividades de I+D+I*
Subprograma de Proyectos de Investigación

DATOS DEL SOLICITANTE / INVESTIGADOR PRINCIPAL

Apellidos y Nombre	**Fernández Hilario, Alberto**				
				D.N.I.	**74694059H**
Grupo Investigación	**Sistemas Inteligentes y Minería de Datos**				
				Código	**TIC-207**
Departamento	**Informática**				
				Código	**618**
Área de Conocimiento	**Arquitectura y Tecnología de Computadores**				
				Código	**035**
Teléfono	**953213016**	**Fax**	**953 212472**	**e-mail**	**alberto.fernandez@ujaen.es**

DATOS DEL PROYECTO DE INVESTIGACIÓN

Título del Proyecto	Retos Actuales en Minería de Datos: Nuevos Modelos para la Resolución de Problemas con Clases Difíciles

N.º de Investigadores UJA	6
N.º de Investigadores Externos	0

DOCUMENTACIÓN QUE SE ACOMPAÑA o ADJUNTA

Memoria científico-técnica (en soporte informático)	☑
Currículum Vitae de los miembros del equipo investigador en modelo normalizado (en soporte informático)	☑
Autorizaciones de Compatibilidad para la participación del personal externo a la UJA en su caso	☐
Solicitud de Informe favorable de la Comisión de Ética en aquellos proyectos que lo requieran	☐

El/la Investigador/a Principal declara que no ha tenido esta condición en ningún proyecto de Investigación del Plan de Apoyo a la I+D+I de la UJA, Instituto de Estudios Giennenses, PAIDI, Plan Nacional/Estatal de I+D+I, Programa Marco o convocatorias homologadas	☑
Me comprometo a solicitar un Proyecto de Investigación del Programa Marco, del P. Estatal de I+D+I, del PAIDI o de convocatorias homologables o Redes Temáticas antes de que finalice la ejecución del proyecto correspondiente a esta convocatoria, en caso de concesión	☑
Me comprometo a presentar como resultado del Proyecto de Investigación concedido una obra científica (libro, monografía o artículo que aparezca en una publicación indexada) o la identificación de una patente	☑

Jaén a, 30 de Mayo de 2014.
El Investigador o investigadora Principal,

Fdo.: Alberto Fernández Hilario

SRA. VICERRECTORA DE INVESTIGACIÓN, DESARROLLO TECNOLÓGICO E INNOVACIÓN.
Nota: Se presentará en papel (original y copia) este impreso de solicitud, con las firmas de conformidad de todos los miembros del equipo investigador, las autorizaciones de participación del personal externo por parte de su entidad de procedencia y la solicitud de informe favorable

A. EQUIPO INVESTIGADOR

1. INVESTIGADOR/A PRINCIPAL

Apellidos y Nombre	Fernández Hilario, Alberto		
D.N.I.	74694059H	**Fecha de Nacimiento**	19/11/1982
Titulación	Ingeniero en Informática	**Grado**	Doctor
Categoría Profesional	Profesor Contratado Doctor	**Situación Laboral**	Activo
Fecha de Toma de Posesión /Contrato	21/02/2013		
Dedicación en la UJA	Tiempo Completo		
Firma de Conformidad			

2. INVESTIGADORES/AS DE LA UJA

Apellidos y Nombre	del Jesus Díaz, María José		
D.N.I.	26485651R	**Fecha de Nacimiento**	1/11/1971
Titulación	Ingeniero en Informática	**Grado**	Doctor
Categoría Profesional	Titular de Universidad	**Situación Laboral**	Activo
Fecha de Toma de Posesión /Contrato	08/01/ 2001		
Grupo PAIDI	**Sistemas Inteligentes y Minería de Datos**		
		Código	**TIC-207**
Departamento	**Informática**		
		Código	**618**
Área de Conocimiento	**Ciencias de la Computación e Inteligencia Artificial**		
		Código	**075**
Firma de Conformidad			

Apellidos y Nombre	García López, Salvador				
D.N.I.	26234294B			**Fecha de Nacimiento**	24/03/1981
Titulación	Ingeniero en Informática			**Grado**	Doctor
Categoría Profesional	Profesor Titular de Universidad			**Situación Laboral**	Activo
Fecha de Toma de Posesión /Contrato	08/05/2012				
Grupo PAIDI	**Sistemas Inteligentes y Minería de Datos**				
			Código	**TIC-207**	
Departamento	**Informática**				
			Código	**618**	
Área de Conocimiento	**Ciencias de la Computación e Inteligencia Artificial**				
			Código	**075**	
Firma de Conformidad					

Apellidos y Nombre	Rivera Rivas, Antonio Jesús				
D.N.I.	26015863B			**Fecha de Nacimiento**	25/2/1973
Titulación	Ingeniero en Informática			**Grado**	Doctor
Categoría Profesional	Titular de Universidad			**Situación Laboral**	Activo
Fecha de Toma de Posesión /Contrato	01/07/2010				
Grupo PAIDI	**Sistemas Inteligentes y Minería de Datos**				
			Código	**TIC-207**	
Departamento	**Informática**				
			Código	**618**	
Área de Conocimiento	**Arquitectura y Tecnología de Computadores**				
			Código	**035**	
Firma de Conformidad					

Apellidos y Nombre	López Godoy, María Dolores				
D.N.I.	75093149N			**Fecha de Nacimiento**	20/10/1970
Titulación	Ingeniero en Informática			**Grado**	Ingeniero
Categoría Profesional	Titular de Universidad			**Situación Laboral**	Activo
Fecha de Toma de Posesión /Contrato	09/03/2012				
Grupo PAIDI	**Sistemas Concurrentes**				
			Código	**TIC-157**	
Departamento	**Informática**				
			Código	**618**	

Área de Conocimiento	**Lenguajes y Sistemas Informáticos**		
		Código	**570**
Firma de Conformidad			

Apellidos y Nombre	Carmona del Jesus, Cristobal				
D.N.I.	75102603J			**Fecha de Nacimiento**	17/06/1982
Titulación	Ingeniero en Informática			**Grado**	Doctor
Categoría Profesional	Contrato asociado a proyecto de investigación			**Situación Laboral**	Activo
Fecha de Toma de Posesión /Contrato	05/01/2014				
Grupo PAIDI	**Sistemas Inteligentes y Minería de Datos**				
		Código	**TIC-207**		
Departamento	**Informática**				
		Código	**618**		
Área de Conocimiento	**Lenguajes y Sistemas Informáticos**				
		Código	**570**		
Firma de Conformidad					

3. INVESTIGADORES/AS EXTERNOS

Apellidos y Nombre			
D.N.I.		**Fecha de Nacimiento**	
Titulación		**Grado**	
Categoría Profesional		**Situación Laboral**	
Entidad u Organismo de Procedencia			
Grupo PAIDI			
		Código	
Firma de Conformidad			

Apellidos y Nombre			
D.N.I.		**Fecha de Nacimiento**	
Titulación		**Grado**	
Categoría Profesional		**Situación Laboral**	
Entidad u Organismo de Procedencia			
Grupo PAIDI			
		Código	
Firma de Conformidad			

Apellidos y Nombre			
D.N.I.		**Fecha de Nacimiento**	
Titulación		**Grado**	
Categoría Profesional		**Situación Laboral**	
Entidad u Organismo de Procedencia			
Grupo PAIDI			
		Código	
Firma de Conformidad			

4. PERSONAL INVESTIGADOR EN FORMACIÓN

Apellidos y Nombre			
D.N.I.		**Fecha de Nacimiento**	
Titulación	Ingeniero en Informática	**Régimen**	Beca ☐ Contrato en prácticas ☐
Fecha Inicio / Fin (Beca o Contrato en Prácticas)			
Entidad Financiadora			
Firma de Conformidad			

Apellidos y Nombre			
D.N.I.		**Fecha de Nacimiento**	
Titulación		**Régimen**	Beca ☐ Contrato en prácticas ☐
Fecha Inicio / Fin (Beca o Contrato en Prácticas)	Inicio ____/____/____ Fin ____/____/____		
Entidad Financiadora			
Firma de Conformidad			

5. PERSONAL DE ADMINISTRACIÓN Y SERVICIOS

Apellidos y Nombre			
D.N.I.		**Fecha de Nacimiento**	
Titulación			
Categoría			
Duración máxima prevista para la colaboración	de ____/____/____ a ____/____/____	**N.º de horas**	
Puesto Ocupado RPT			
Firma de Conformidad			

Apellidos y Nombre			
D.N.I.		**Fecha de Nacimiento**	
Titulación			
Categoría			
Duración máxima prevista para la colaboración	De ____/____/____ a ____/____/____	**N.º de horas**	
Puesto Ocupado RPT			
Firma de Conformidad			

B. DESCRIPCIÓN DEL PROYECTO

Título

Retos Actuales en Minería de Datos: Nuevos Modelos para la Resolución de Problemas con Clases Difíciles

Resumen Español/Inglés

Las tareas de clasificación y predicción están continuamente presentes en la vida cotidiana. Podemos encontrar diversos ejemplos en la vida real realizadas por expertos en diferentes ámbitos, como por ejemplo en diagnóstico médico, reconocimiento de patrones, calificación de productos, y un largo etcétera. El desarrollo de sistemas automáticos puede ayudar a facilitar la labor a realizar, y permitir efectuar mejor la predicción, ya que el volumen de datos con los que puede trabajar es mucho mayor que la capacidad de una persona.

Cuando se pretende resolver una aplicación dada en el escenario de la clasificación, los expertos e investigadores deben conocer la estructura y características intrínsecas de los datos que gestionan para de este modo alcanzar la máxima precisión para todos los conceptos incluidos en el problema. Un problema en este sentido es el del tratamiento de las *clases difíciles*, definido en términos generales como aquéllas sobre las que la tasa de acierto de los clasificadores es mucho menor que en el resto, lo que puede llevar a que sea ignorada.

Uno de los posibles ejemplos prácticos son aquéllas áreas de trabajo en los que la distribución de las clases no es equilibrada, conocido como la clasificación con conjuntos de datos no balanceados. La mayoría de aproximaciones de aprendizaje estándar consideran un conjunto de entrenamiento equilibrado (o balanceado), esto conlleva a la obtención de un modelo de clasificación sub-óptimo, es decir, una buena cobertura de los ejemplos mayoritarios (también conocida como clase negativa), mientras que los minoritarios (o clase positiva) son más difíciles de discriminar. Este hecho se ve acentuado si trabajamos en un contexto con múltiples clases, pues existe un mayor número de fronteras a considerar.

Debemos enfatizar la importancia de este problema, ya que está relacionado con problemas en dominios del mundo real que implican un alto coste cuando los ejemplos de estas clases difíciles se clasifican de manera errónea. Este escenario tiene una especial relevancia para aplicaciones en el campo de la medicina, en la que los expertos deben discriminar cuando una paciencia dado tiene una enfermedad, siendo el riesgo de un falso negativo más peligroso que el caso contrario. Otro caso interesante de estudio está relacionada con los sistemas de detección de intrusiones, donde los datos se refieren a los accesos anormales de comportamiento son muy raros, siendo además necesario discriminar correctamente entre diferentes clases de ataques al sistema para su posterior tratamiento.

De acuerdo al contexto planteado, nuestro objetivo general del proyecto será el desarrollo de algunos de los enfoques de aprendizaje que abordan el aprendizaje sobre clases difíciles desde el punto de vista de (1) los conjuntos de datos con clases no balanceadas (2) la extensión de este problema de clasificación a un escenario multi-clase, lo que implica una dificultad añadida a la consecución de una precisión media de todos los conceptos representados.

The classification and prediction tasks of are continuously present in everyday life. We may find many examples in real life carried by experts in several fields, such as in medical diagnosis, pattern recognition, product qualification, and so on. The development of automated systems may help to make the work to be easier, and also allow at performing a better prediction, since the volume of data it can work with is much greater than a person's ability.

When trying to solve a given application in the classification scenario, experts and researchers must know the structure and intrinsic characteristics of the data to achieve the maximum accuracy for all items included in the problem. A related problem is that of "difficult classes", which are defined as those ones for which the accuracy of the classifier is much smaller than for the remaining ones, so that it can be ignored.

One of the related practical examples, are those frameworks where the class distribution is imbalanced, known as the classification with imbalanced datasets. Most of the standard learning approaches consider a balanced training set, which leads to the production of sub- optimal classification models, i.e. good coverage of the majority examples (also known as negative class), whereas minority ones (positive class) are more difficult to discriminate. All these problems are further accentuated when working in a context with multiple classes, since there are a larger number of boundaries to consider.

We must emphasize the importance of this issue as it is related to problems in real-world domains that involve a high cost when these difficult class examples are misclassified. This scenario is especially relevant for applications in the medical field, in which the experts must discriminate when a patient has a given disease, being the risk of a false negative more dangerous than the contrary case. Another interesting case study is related to intrusion detection systems, where the data related to abnormal behavior hits is very rare, being also necessary to discriminate correctly among different types of system attacks for their isolate management.

According to the fomer, the aim of the project will be to develop some learning approaches that address learning on difficult classes from the point of view of (1) data sets with imbalanced classes (2) the extension of this problem to a multi-class scenario, which involves an additional difficulty to achieve an average accuracy for all represented concepts.

Palabras Clave

Minería de Datos, Aprendizaje, Clases Difíciles, Clasificación no Balanceada, Múltiples clases, Descomposición de Clasificadores, Ensembles, Big Data.

Áreas científico-técnicas ANEP [1]

Área de Ciencias de la computación y tecnología informática (INF)

Códigos UNESCO

1203.04	1203.12	1203.18	1209.01	1209.03
1209.12	1209.14			

[1] Área de Agricultura (AGR) / Área de Biología Molecular, Celular y Genética (BMC) / Área de Biomedicina (BMED) / Área de Biología Vegetal y Animal, Ecología (BVAE) / Área de Ciencia y Tecnología de Alimentos (TA) / Área de Ciencias de la computación y tecnología informática (INF) / Área de Ciencias de la tierra (CT) / Área de Ciencias sociales (CS) / Área de Derecho (DER) / Área de Economía (ECO) / Área de Filología y Filosofía (FFI) / Área de Física y ciencias del espacio (FI) / Área de Fisiología y Farmacología (FFA) / Área de Ganadería y pesca (GAN) / Área de Transferencia de Tecnología (IND) / Área de Historia y arte (HA) / Área de Ingeniería civil y arquitectura (ICI) / Área de Ingeniería eléctrica, electrónica y automática (IEL) / Área de Ingeniería mecánica, naval y aeronáutica (IME) / Área de Matemáticas (MTM) / Área de Medicina Clínica y Epidemiología (MCLI) / Área de Psicología y Ciencias de la Educación (PS) / Área de Química (QMC) / Área de Tecnología electrónica y de las comunicaciones (COM) / Área de Tecnología Materiales (TM) / Área de Tecnología Química (TQ). ACCESO WEB: http://www.mec.es/ciencia/jsp/plantilla.jsp?area=anep&id=24

Presupuesto Solicitado

	Importe
1. Gastos de Personal [2]	4.000,00€
2. Gastos de Ejecución	6.000,00€
2.i) Material inventariable	**3.000,00€**
2.ii) Material fungible	**350,00€**
2.iii) Material bibliográfico	**150,00€**
2.iv) Viajes, dietas e inscripciones a congresos, exclusivamente para el personal que forme parte del equipo investigador de acuerdo con las NGEP del Presupuesto de la UJA vigente	**2.500,00€**
2.v) Otros gastos complementarios necesarios para el desarrollo del proyecto directamente relacionados con la actividad y debidamente justificados *(Realizar breve descripción):*	**0,00€**
3. TOTAL (1 + 2)	**10.000,00 €**

Nota: La propuesta económica no podrá exceder de 10.000,00 € y la duración será de 2 años.

Memoria científico-técnica

Nota: Adjuntar en documento anexo a esta solicitud, incluyendo en la misma:

1. Antecedentes y estado actual del tema, incluyendo la bibliografía más relevante comentada.
2. Hipótesis y planteamiento de la investigación.
3. Descripción de manera precisa, clara y realista de los Objetivos que se persiguen.
4. Material y métodos.
5. Plan de trabajo comprendiendo cronograma orientativo y reparto de tareas.
6. Medios disponibles y requeridos para la ejecución del proyecto, incluyendo propuesta económica, que no podrá exceder de 10.000,00 €, desglosada por gastos de personal y gastos de ejecución (deberá coincidir con cuadro de presupuesto solicitado detallado en este impreso de solicitud).
7. Plan de difusión de resultados y repercusión esperable de los resultados, tanto en su impacto bibliométrico como impacto por transferencia de conocimiento/tecnología.

[2] Personal colaborador externo nombrado de acuerdo con lo establecido en el Reglamento de Colaboradores con Cargo a Créditos de Investigación. En ningún caso se contemplarán retribuciones para los miembros del equipo investigador solicitante.

Antecedentes y estado actual del tema, incluyendo la bibliografía más relevante comentada

Los avances continuos en generación y almacenamiento de información han hecho posible que en las distintas áreas de conocimiento y negocio existan grandes cantidades de datos. La búsqueda de patrones en datos, habitual en el ser humano, se automatiza en gran medida mediante los algoritmos de extracción de conocimiento en grandes bases de datos (en inglés "*Knowledge Discovery in Databases"* (KDD)).

El KDD fue definido en el año 1996 como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y que finalmente pudiesen ser utilizados, en los datos" [Fay96]. Actualmente desempeña un papel importante en una doble vertiente: es fundamental en investigación científica por ser una herramienta de análisis y descubrimiento de conocimiento a partir de datos de observación, y crece de forma gradual su transferencia con éxito desde aplicaciones tradicionales como finanzas o marketing, a otros dominios como industria, energía, medicina, bioinformática o análisis de información web, entre otros, en los que el volumen de información y la necesidad de extraer conocimiento útil, que produzca beneficio directo, se incrementan casi en la misma medida.

El KDD está formado por un conjunto de pasos interactivos e iterativos, entre los que se incluye el pre-procesamiento de los datos, la búsqueda de patrones de interés con una representación particular y la interpretación de estos patrones (ver Figura 1). Aunque KDD es el nombre apropiado cuando hablamos de este procedimiento, el término Minería de Datos (MDD, en inglés "*Data Mining*") [Tan06] es utilizado con frecuencia para referirse al proceso completo, si bien éste representa la parte encargada de extraer el conocimiento a partir de los datos procesados [Pyl99], siendo realmente la principal de todo el sistema. En MDD se puede distinguir, en función del objetivo, entre tareas predictivas y descriptivas. En las primeras, el objetivo es encontrar un modelo que permita predecir un comportamiento futuro, habitualmente mediante inducción supervisada. En este grupo de tareas de MDD se encuentra la clasificación, regresión y predicción de series temporales. En MDD descriptiva se busca, mediante aprendizaje no supervisado, un modelo que describa información sobre el problema que subyace bajo los datos e incluye la extracción de reglas de asociación, clustering y sumarización, entre otras tareas de MDD.

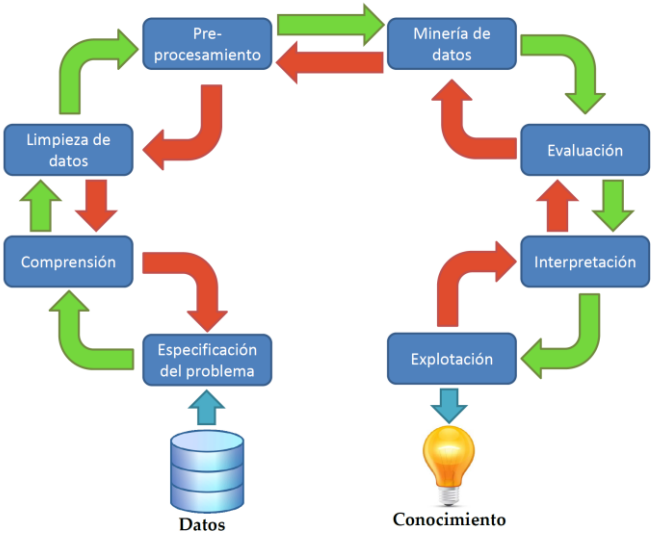

Figura 1. Proceso de KDD

Una de las áreas con una fuerte intersección con la MDD es el aprendizaje automático (en inglés, "*Machine Learning"*) [Alp04] El aprendizaje automático es una rama de la inteligencia artificial que trata el diseño y desarrollo de algoritmos que permiten aprender comportamientos, patrones o conceptos basada en la observación de datos empíricos, como pueden ser datos provenientes de sensores o de bases de datos (como es el caso más relacionado con la MDD). En definitiva, es una herramienta que nos permite extraer conocimiento a partir de una serie de ejemplos del problema a resolver.

En este proyecto vamos a focalizar nuestro trabajo en el escenario del aprendizaje automático supervisado, y más concretamente, en la clasificación. En este contexto, entendemos por clasificación el proceso en el que, sabiendo la existencia de ciertas clases o categorías, establecemos una función o regla para ubicar nuevas observaciones en alguna de las clases existentes (aprendizaje supervisado). Un clasificador recibe como entrada un conjunto de ejemplos, denominado conjunto de entrenamiento, con el que se aprende la regla de clasificación. Además, en el proceso de validación de un clasificador, se utiliza un conjunto de ejemplos, no conocido en el proceso de aprendizaje, denominado conjunto de test y utilizado para comprobar la precisión del clasificador. Las clases resultan de un problema de predicción, donde cada clase corresponde a la salida posible de una función a predecir a partir de los atributos con que describimos los elementos de un conjunto de datos.

Un problema dentro del marco de la clasificación, que aún no ha sido abordado con suficiente detalle, es el *problema de las clases difíciles*, en el que particularmente hemos comenzado a trabajar desde el grupo de investigación [Gal14]. Una clase puede decirse que es difícil cuando la tasa de acierto de los clasificadores sobre ella es mucho menor que sobre el resto, lo que puede llevar a que sea ignorada. Tal y como especificamos anteriormente, este es un problema novedoso sobre el que no existen muchas aportaciones en la literatura especializada. Sin embargo, en muchos problemas reales se puede querer priorizar una clasificación equitativa de las diferentes clases, en lugar de ser muy preciso solo en algunas de ellas. Por ejemplo, en la clasificación de señales de tráfico [Puj06], la identificación de autores [Sta08], la predicción del cáncer [Pau09], la detección de patrones en vídeo [Yua11] o la clasificación de texturas [Liu12].

Para ilustrar este problema, mostraremos un ejemplo sencillo que consiste en considerar un problema de tres clases, con exactamente el mismo número de ejemplos en cada una de ellas. Este problema sufrirá el problema de las clases difíciles si alguna de ellas tiene un ratio de reconocimiento del 50%, por ejemplo, debido a su solapamiento con las otras clases, cuyo ratio de reconocimiento es del 90% (por tanto, siendo la precisión global del 76,66 %). Por contra, podríamos obtener el mismo resultado global del 76,66% sobre todas las clases, teniendo ese mismo porcentaje de acierto para cada una de ellas, lo cual es sin duda una situación completamente diferente, pero seguramente mucho más deseable, ya que en este caso, todas las clases son igualmente reconocidas, mientras que en el primero, una de ellas tenía un porcentaje de acierto mucho menor.

Las características que definen a cada clase en un problema de clasificación suelen ser normalmente diferentes: el número de instancias (distribución de ejemplos), las relaciones entre clases, las relaciones entre los ejemplos de la propia clase y el solapamiento entre clases. Debido a las características que definen a cada clase, algunas pueden ser más fáciles de distinguir que otras. En este contexto, las clases difíciles podemos definirlas como aquellas sobre las que obtenemos un menor ratio de aciertos. Si consideramos la precisión global como la medida de evaluación para los clasificadores, estas clases pueden perder su importancia, ya que es más fácil mejorar el reconocimiento de las clases fáciles a cambio de clasificar erróneamente alguna de las instancias de las clases difíciles.

Dentro de los problemas con clases difíciles, podemos encontrar aquéllos cuya distribución de ejemplos en las distintas clases o conceptos que representan el conjunto de datos no es uniforme [He09, Sun09]. Este problema es observable en multitud de ejemplos, como pueden ser la detección de fraudes [Oen14], estimación de riesgos bancarios [Hua06], clasificación de textos [Ngu11], diagnosis médica [Zie14], y muchos otros dominios en los que esta característica es inherentemente implícita al problema ya que, afortunadamente suelen existen muy pocos casos "anómalos" en comparación con los casos normales. Otra situación que puede derivar en la aparición de este tipo de conjuntos ocurre cuando el proceso de colección de datos está limitado (debido a razones económicas o privadas). Es importante hacer notar que este tipo de conjuntos de datos con *clases no balanceadas* difieren de los conjuntos estándar no solo en el desequilibrio entre las clases, sino también en la importancia creciente de la clase minoritaria, tradicionalmente identificada como "clase positiva".

A pesar de mostrar una ocurrencia bastante frecuente y un fuerte impacto en las aplicaciones del día a día, el problema de las clases no balanceadas no ha sido solventado de manera apropiada por los algoritmos de aprendizaje automático, puesto que asumen distribuciones de clases balanceadas o costes de clasificación iguales para todas las clases. En efecto, la mayoría de los algoritmos de aprendizaje tienen como objetivo obtener un modelo con un alto acierto en predicción y una buena capacidad de generalización. Sin embargo, aquéllos algoritmos que obtienen un buen comportamiento en el marco de la clasificación estándar no necesariamente alcanzan el mejor rendimiento para conjuntos de datos no balanceados [Fer10b].

Observamos por tanto que el sesgo de los algoritmos de clasificación estándar para los ejemplos de la clase mayoritaria es la consecuencia más directa derivada de la distribución desigual de clases. Cuando el proceso de búsqueda se guía mediante la tasa de acierto estándar, beneficia la cobertura de los ejemplos mayoritarios. Segundo, las reglas de clasificación que predicen la clase positiva son a menudo altamente especializadas y así su grado de cobertura es muy bajo, por lo tanto son descartadas en favor de reglas más generales, por ejemplo aquellas que predicen la clase negativa.

Estudios recientes realizados por parte nuestro equipo actual de investigación, muestran que el problema de las clases no balanceadas suele aparecer en combinación con diversas características intrínsecas de los datos [Lop13]. Esto impone restricciones adicionales durante la etapa de aprendizaje. En primer lugar, destacamos la presencia de áreas con un alto solapamiento entre las clases, cuyo efecto es mucho más negativo cuando queremos discriminar los ejemplos de la clase positiva [Gar08, Den10]. Asimismo pueden existir pequeños grupos de ejemplos (en inglés "small-disjuncts") de la clase minoritaria que pueden ser tratados erróneamente como ruido, y por tanto ignorados por el clasificador [Orr09, Wei10]. La existencia de incluso pocos ejemplos ruidosos pueden degradar la identificación de la clase minoritaria, ya que tiene de por sí un menor número de ejemplos [Sei14]. Por último, debemos indicar el caso del "dataset shift" o "cambio en conjunto de datos", referido a la diferente distribución de los datos entre las particiones de entrenamiento y test [Lop14].

Surge por tanto una alta dificultad para poder alcanzar el objetivo final de desarrollar un clasificador que alcance una elevada precisión tanto para las clases positiva como negativa del problema. Por ello, un gran número de soluciones han sido desarrolladas para esta tarea, pudiendo categorizarse en tres grandes grupos [Lop12]:

- Muestreo de datos: en el que las instancias de entrenamiento se modifican para alcanzar una distribución de clases más equilibrada que permita a los clasificadores trabajar de un modo similar a la clasificación estándar [Bat04].
- Modificación algorítmica: este procedimiento está orientado hacia la adaptación de los modelos de aprendizaje base, para así poder condicionarlos a las condiciones del desequilibrio o desbalanceo de clases [Lin13, Zon13].
- Aprendizaje sensible al coste: este tipo de soluciones incorporan aproximaciones a nivel de los datos, a nivel algorítmico, o incluso a ambos niveles en conjunto. Se consideran costes más altos por la mala clasificación de ejemplos de la clase positiva con respecto a la clase negativa y, por tanto, trata de minimizar el nivel de coste asociado al problema global [Bat10, Zad03].

Adicionalmente a las técnicas anteriores, es posible combinar un conjunto de clasificadores para mejorar la precisión del sistema final. Esta precisión se mide mediante el porcentaje de aciertos obtenidos sobre los ejemplos no utilizados en la fase de aprendizaje del sistema [Dud01]. Este tipo de sistemas, que combinan varios clasificadores, se denominan multi-clasificadores o ensembles [Pol06, Rok10], modificando o adaptando mediante la combinación entre el propio algoritmo de aprendizaje ensemble y cualquiera de las técnicas descritas anteriormente, a saber, bien a nivel de los datos o mediante enfoques algorítmicos basados en el aprendizaje sensible al coste, siendo todas soluciones evaluadas por el equipo de investigación de este proyecto [Gal12].

En efecto, una amplia línea de investigación seguida por los miembros del grupo de investigación ha estado versada en la clasificación con datos no balanceados mediante el uso de técnicas Soft Computing, por ejemplo utilizando sistemas basados en reglas de clasificación difusas [Fer08, Fer09, Fer10, Vil12, Lop13b], redes neuronales [Per10], o mediante el uso de ensembles de clasificadores [Gal13].

De manera tradicional, en clasificación no balanceada la investigación se ha basado en el caso de estudio binario, si bien en este contexto particular [Fer13] y en el de las "clases difíciles" en general, nos encontramos mayormente con aplicaciones en *problemas de clasificación multi-clase*. En efecto, el dominio de aplicación de las técnicas para problemas multi-clase es diverso, por ejemplo, en el campo de la bioinformática, la clasificación de microarrays [Liu09] y tejidos [Tao04] tratan con múltiples clases; en visión por computador el reconocimiento de objetos en imágenes y vídeos [Tor07] y el reconocimiento del lenguaje de signos [Ara10], así como en medicina pueden existir varias clases en la clasificación de un tipo de cáncer [Ana09] o de señales de electroencefalogramas [Gul07].

Generalmente, es más sencillo construir un clasificador para distinguir entre dos clases, que considerar más de dos clases en un problema, ya que las fronteras de decisión en el primer caso son más simples. Ésta es una de las principales razones por la cual surgen las técnicas de binarización, para abordar los problemas multi-clase dividiendo el problema original en problemas binarios más sencillos de resolver, que pueden ser afrontados por clasificadores binarios independientes. Para clasificar una nueva instancia del problema, tomamos las salidas de cada uno de los clasificadores del ensemble y las agregamos para decidir la clase final con la que es etiquetada la instancia. La aplicación de ensembles a problemas de multi-clasificación (con múltiples clases) permite mejorar los resultados obtenidos cuando tratamos el problema con un único clasificador que trata de distinguir entre todas las clases, debido a la simplificación del problema inicial [Lor08, Fur02, Rif04].

Los esquemas diseñados en este sentido pueden codificarse en el marco de trabajo de los Error Correcting Output Code (ECOC) [Die95, All00]. De entre todas, dos de las estrategias más comúnmente utilizadas son la estrategia uno-contra-uno (One-vs-One, OVO) [Kne90] y uno-contra-todos (One-vs-All, OVA) [Cla91, Ana95]. El funcionamiento de ambos modos quedo ilustrado, sobre un problema multi-clase (de 3 clases) en la Figura 2.

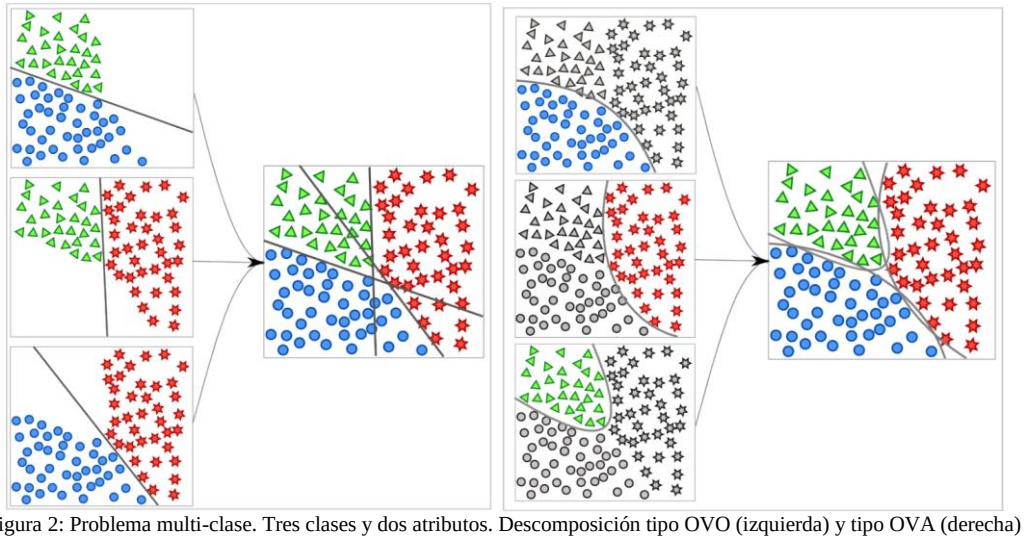

Figura 2: Problema multi-clase. Tres clases y dos atributos. Descomposición tipo OVO (izquierda) y tipo OVA (derecha)

Este tipo de descomposiciones han mostrado su potencial para mejorar las precisiones obtenidas respecto a los clasificadores que soportan problemas multi-clase de manera directa, tal y como se resume en nuestros estudios [Gal11]. Además de esta gran ventaja, su utilización está muy extendida en comunidades científicas como la de las Máquinas de Soporte Vectorial (en inglés Support Vector Machine o SVM), ya que de esta forma es posible afrontar problemas multi-clase mediante clasificadores binarios que no tienen un soporte multi-clase (o que su extensión no está establecida). Sin embargo, si bien cada problema binario es más simple que el multi-clase original, una vez construidos los clasificadores, es necesario utilizar alguna agregación para combinar sus salidas.

La manera en la que agregamos estas salidas puede ser tan importante como la forma en la que descomponemos el problema. En efecto, pudimos destacar cómo el esquema de binarización tipo OVO era el más efectivo en este caso, y ello dio pie al desarrollo de nuevos modelos para mejorar el comportamiento de estos sistemas [Gal14].

Sin embargo, volviendo al tema de trabajo destacado inicialmente, como es el problema de las "clases difíciles", debemos volver a resaltar que apenas hay alguna aproximación inicial que se haya aplicado a esta área de trabajo. De esta forma, aún contamos con un alto margen de mejora mediante la adaptación de los algoritmos de los subproblemas antes citados. En resumen, hemos de destacar dos condiciones fundamentales que dan especial importancia a la consecución de este proyecto:

1. En primer lugar, con respecto a la clasificación no balanceada binaria, observamos que es un caso particular del problema de las clases difíciles. Puesto que el inconveniente este escenario general no está causado únicamente por la propia distribución de clases, sino que existen hay más factores que hacen que una clase sea difícil, hace que muchas de las propuestas desarrolladas sobre dicho problema, no sean válidas para afrontar las clases difíciles (como puede ser el pre-procesamiento de datos para balancear la distribución de clases). Por ello, es necesario estudiar nuevas propuestas para resolver este problema, como por ejemplo realizar una metodología sensible al coste realizando una ponderación de las instancias del conjunto de datos, bien a partir de un proceso de aprendizaje genético, bien mediante el uso de ensembles de clasificadores diseñados "ad-hoc".
2. En segundo lugar, en el contexto de múltiples clases, sería interesante manejar las clases difíciles en el proceso de agregación de un multi-clasificador tipo OVO, donde podríamos intentar beneficiar las clases positivas con un proceso de decisión por "torneo", o directamente agrupando clases durante el proceso de aprendizaje inicial, para mejorar la discriminación entre las fronteras.

Por último, debemos destacar que en los últimos años se ha observado la aparición de un factor adicional que condiciona el desarrollo de potenciales programas para la inducción del conocimiento, como es la *Explosión de datos* en la que actualmente nos vemos envueltos. De este modo, se ha acuñado el término de "Big Data" como referencia a aquéllos desafíos (o incluso ventajas) derivados de la recogida y tratamiento de estas grandes cantidades de datos [Mar13].

Para ser capaces de gestionar problemas en este marco de trabajo, se ha establecido una nueva generación de sistemas, cuyos exponentes más representativos de la industria y el mundo académico son MapReduce [Dea08] y su implementación de código abierto Hadoop MapReduce [Lam11] (no confundir con la pila de protocolos Hadoop, que añade todas las funcionalidades de un entorno tipo Cloud Computing [Buy11]). Este nuevo paradigma elimina las restricciones anteriores para precargar los datos, uso de esquemas de almacenamiento fijos, o incluso el uso de SQL (en tareas analíticas). En cambio, los desarrolladores podrán codificar sus programas utilizando este nuevo modelo permitirá paralelizar automáticamente la ejecución del modelo implementado.

Por tanto, como línea de trabajo futuro podría resultar de un alto interés la extensión tanto de los modelos usados como base en este proyecto, como de los actualmente desarrollados para su aplicación en problemas Big Data.

Hipótesis y planteamiento de la investigación

La hipótesis de trabajo se basa en la mejora de los modelos actuales de clasificación en diferentes escenarios de aplicación para los que aún no se ha incidido con suficiente profundidad como son, en términos generales, la resolución del problema de la clasificación sobre clases difíciles, y en particular la predicción sobre clases no balanceadas, y la gestión de conjuntos con múltiples clases.

En este sentido, primero debemos ser capaces de asentar los fundamentos que definen aquéllas clases difíciles como tales. Ello podría llevarse a cabo con respecto a su representación en el problema, medidas de complejidad de datos asociadas, la precisión obtenida por un clasificador genérico, o con cualquier otra metodología similar.

Por otro lado, destacamos que la medida de rendimiento seleccionada marca el objetivo final de cualquier estudio experimental. Cuando la precisión se mide mediante la precisión estándar (*accuracy*) es bien sabido que no aporta información sobre la calidad en las distintas clases del problema, dado que es muy genérica. Otras medidas como "kappa" [Ben07, Ben08] no llegan a instaurarse del todo y en ocasiones causan controversia.

De este modo, resulta positivo estudiar otro tipo de medidas que ponderen la importancia de estas clases difíciles. Diferentes expresiones de las métricas de calidad, pueden ser utilizadas para guiar el proceso de optimización y aprendizaje de los modelos, lo cual puede realizarse dentro de un proceso de descomposición para problemas multi-clase (binarización OVO).

Destacar que asimismo el esquema OVO muestra ciertas debilidades en la precisión sobre las clases difíciles [Gal14]. Así pues, sería interesante estudiar modelos de agregación que permitan alterar el comportamiento de la estrategia OVO para así beneficiar a las clases difíciles. Uno de los ejemplos de actuación sería la obtención de un tratamiento más adecuado cuando existen clasificadores no competentes, es decir, aquéllos clasificadores base o binarios que son disparados para dar una salida sobre una instancia de una clase para la que no han sido entrenados. Otra línea de actuación significativa en este contexto sería justamente la de aplicar esquemas que manejen la clasificación no balanceada en conjuntos de datos sobre múltiples clases.

Para el segundo de los subproblemas tratados, debemos destacar que el esfuerzo principal sobre clasificación no balanceada ha estado siempre ligado a resolver el desequilibrio entre los ejemplos de distintas clases del problema, sin atender realmente al resto de problemas intrínsecos que aparecen íntimamente relacionados con el anterior. De este modo, pensamos que trabajar sobre el solapamiento entre los datos, o el descubrimiento y tratamiento de los "pequeños datos disjuntos" puede dar lugar a la generación de modelos óptimos en esta tipología de problemas.

Para ello podemos realizar modelos a nivel de los datos, bien relativos al preprocesamiento de instancias y/o de variables. Adicionalmente, podemos trabajar también a nivel algorítmico creando soluciones más concretas que apliquen una metodología de aprendizaje orientada a gestionar los problemas anteriormente citados, incluso trabajando a nivel de ensembles de clasificadores, aprovechando las ventajas ofrecidas por este tipo agregación de métodos.

En concreto, siguiendo el fundamento de las técnicas de "boosting" en ensembles, y los buenos resultados ofrecidos en la literatura especializada, parece lógico que se pueda seguir una tendencia similar mediante la "ponderación de instancias" en lugar de hacer una simple "selección", como se ha hecho tradicionalmente. Básicamente, la "selección de instancias" puede resultar ligeramente agresiva eliminando ejemplos. Sabemos que su finalidad es la de reducir ruido y deshacerse de instancias redundantes. Sin embargo, sería mucho más positivo diseñar un modelo que "fortifique" o de importancia a las instancias claves del problema asignándoles un coste. En concreto, estas instancias deberán ser aquéllas que estén rodeadas por muchas de la clase contraria (serían "small disjuncts" o "borderline") pero que asimismo tengan una representación adecuada en su área del problema, es decir, que no sean ruido ni outliers.

Por último, comentar la posibilidad de extender todos estos estudios a un escenario de trabajo en Big Data, el cual permite el tratamiento sobre grandes volúmenes de información y la gestión trasparente de la escalabilidad para mejorar la actuación de los algoritmos de MDD desarrollados.

Objetivos

Los objetivos generales perseguidos en este proyecto se dirigen a la construcción de nuevos modelos y algoritmos de clasificación para la gestión de problemas con clases difíciles en un sentido general, como particularmente sobre conjuntos no balanceados, tanto binarios como multi-clase. De este modo, se diseñarán sistemas de aprendizaje que permitan obtener información más útil y completa frente a problemas para los que en un principio la precisión alcanzada, según diferentes métricas de calidad, sea inferior a la de los umbrales requeridos por los usuarios o expertos en diversas aplicaciones del mundo real.

Para alcanzar dichos objetivos, hemos destacado un conjunto de subobjetivos, que se definen a continuación:

1. Tratamiento de las clases difíciles: Una aproximación inicial consiste en utilizar aprendizaje sensible al coste una vez definidos los conceptos del problema en los que prestar mayor atención. Con objeto de no crear un sesgo demasiado importante hacia estas clases de interés, debemos hacer uso de técnicas más sofisticadas para equilibrar en lo posible la precisión global. Otras opciones estarían enmarcadas en una correcta gestión durante el proceso de decisión en el esquema OVO, o incluso una agrupación óptima de las clases para realizar mejor la labor de aprendizaje en la binarización.

2. Clasificación con datos no balanceados: para ello necesitaremos solventar diferentes cuestiones inherentes a este escenario. El problema fundamental reside en la distribución inicial de ejemplos dentro del conjunto de datos, que fragmenta el problema en grupos con un ratio diferente entre las clases, lo que afecta al aprendizaje en sí. Nosotros queremos ir un paso más allá y gestionar distintas características intrínsecas a los datos para focalizar los esfuerzos en aplicaciones más específicas sobre las que mejorar la calidad de los modelos aprendidos. En concreto nos centraremos en el tratamiento de los pequeños datos disjuntos, el solapamiento, y el ruido, mediante el uso de técnicas de boosting en ensembles, así como un aprendizaje genético para la ponderación de las instancias.

3. Uso de esquemas de descomposición para un entorno multi-clase: debemos identificar nuevas aproximaciones para afrontar la etapa de decisión, es decir, la agregación de la salida de todos los clasificadores binarios. También podemos gestionar el proceso de construcción del propio modelo para ponderar la importancia de las diferentes clases del problema. En efecto este punto estaría en cierto modo ligado al objetivo 3.a para las clases difíciles.

4. Estudiar posibles adaptaciones de los modelos desarrollados al entorno de Big Data mediante uso de paradigmas de paralelización, por ejemplo MapReduce.

Material y métodos

La metodología propuesta abarca una vertiente teórica y otra práctica. En el aspecto teórico debemos estudiar el comportamiento de los modelos actuales sobre los problemas planteados en el tratamiento de las clases difíciles, clasificación no balanceada, y multi-clase. En la vertiente práctica, una vez diseñados los algoritmos y procedimientos para resolver las casuísticas anteriores, se desarrollarán la implementación de los mismos por los distintos miembros del equipo.

En el aspecto teórico, el método de estudio a seguir es el habitual método científico:

- Formulación de hipótesis, que en nuestro caso implica el desarrollo de propuestas nuevas para algoritmos en el escenario mencionado previamente. Además, será necesario la identificación de mecanismos de comparación de los algoritmos para la correcta validación de la metodología.
- Recogida de observaciones, que en nuestro contexto supone disponer de varias bases de datos con características dentro del contexto a analizar, a saber, clases difíciles, clasificación no balanceada, multi-clase. Es importante recordar que estos problemas están estrechamente relacionados, por lo que podremos contar con conjuntos de datos genéricos para analizar todos los casos.
- Contraste de hipótesis con las observaciones, es decir, evaluación de la calidad de los algoritmos de extracción de conocimiento con respecto a las bases de datos empleadas. Para ello emplearemos la herramienta KEEL; y finalmente,
- Readaptación de las hipótesis iniciales a la luz de los resultados obtenidos; que implicará la modificación y refinamiento de los algoritmos de aprendizaje y las características intrínsecas a su codificación, así como de los mecanismos de análisis de su comportamiento como consecuencia de las pruebas realizadas y la experiencia acumulada.

Para el desarrollo del software seguiremos el paradigma del ciclo de vida clásico: análisis de requisitos, para obtener la especificación del problema en términos informáticos; análisis del sistema y diseño, que delimitarán el sistema a construir, determinarán las estructuras de datos, el esquema funcional y los mecanismos de comunicación entre módulos; codificación, que en nuestro caso utilizará el lenguaje de programación JAVA debido a las especificaciones del entorno KEEL que está desarrollado en este lenguaje. Durante todo el proceso se aplicarán mecanismos de verificación y validación para asegurar el correcto desarrollo y funcionamiento del sistema.

Varios de los algoritmos de MDD que usaremos como base están ya disponibles a través de la herramienta KEEL [Alc09, Alc11], desarrollada por varios miembros del equipo y disponible en la web http://www.keel.es En particular, KEEL dispone de un módulo íntegramente desarrollado para clasificación con clases no balanceadas, junto con un bloque de algoritmos de multi-clasificación, tanto basados en descomposición, como tipo ensemble.

Para realizar correctamente la experimentación y evaluación de los sistemas diseñados, haremos en primer lugar uso del servidor de cálculo disponible por el grupo de investigación SIMIDAT en el que todos los miembros del equipo de trabajo participan activamente. Adicionalmente, para tareas que requieran un mayor uso de escalabilidad, podremos hacer uso del servidor de cálculo dispuesto en el Centro de Estudios Avanzados en Tecnologías de la Información y Comunicación (CEATIC). Este clúster de máquinas forma una plataforma de computación al servicio de la comunidad investigadora de la Universidad de Jaén para su fácil aprovechamiento de los recursos.

Plan de trabajo comprendiendo cronograma orientativo y reparto de tareas

Como indicamos en la sección de objetivos, se aborda el análisis de la adaptación de algoritmos de Minería de Datos en clasificación no balanceada y multi-clase con las etapas que la componen. De este modo, la división en tareas se lleva a cabo de la siguiente forma:

Tarea A) Análisis inicial del problema de clases difíciles.

Coordinación: A. Fernández

Colaboradores: M.J. del Jesus, S. García, A. Rivera,

Descripción: profundizar en la definición de "clases difíciles" y propuestas de metodologías orientadas a mejorar la precisión local en dichas clases.

Temporización: M1-M4: Análisis del problema

Objetivo Desarrollado: Objetivo 1.

Resultado previsible: Descubrimiento de medidas que identifiquen problemas con clases difíciles. Identificación de técnicas y algoritmos para la resolución del problema

Tarea B) Estudio de algoritmos sensibles al coste en clasificación no balanceada.

Coordinación: M.J. del Jesus

Colaboradores: A. Fernández, A. Rivera, M.D. Pérez

Descripción: Estudio del caso particular de la clasificación no balanceada y su resolución con preprocesamiento a nivel de instancias mediante "ponderación de ejemplos", junto con el desarrollo de modelos tipo ensemble.

Temporización: M2-M5: Estudio clasificación no balanceada

Objetivo Desarrollado: Objetivo 2.

Resultado previsible: Diseño de nuevos modelos para clasificación no balanceada que resuelvan los problemas intrínsecos de los datos en solapamiento, pequeños datos disjuntos, y ruido.

Tarea C) Tratamiento de los clasificadores no competentes en un esquema tipo OVO y el aumento de la precisión en el escenario de múltiples clases no balanceadas.

Coordinación: A. Fernández

Colaboradores: C. Carmona, S. García, A. Rivera

Descripción: Identificar nuevas aproximaciones para afrontar la etapa de decisión del multi-clasificador. Estudio del proceso de construcción del propio modelo para ponderar la importancia de las diferentes clases del problema.

Temporización: M3-M6: Análisis de los problemas multi-clase

Objetivo Desarrollado: Objetivo 3.

Resultado previsible: Definición de sistemas alternativos para agregar las salidas en el esquema de multi-clasificación. Nuevos esquemas de agrupación de clases durante la binarización.

Tarea D) Desarrollo e implementación de los modelos.

Coordinación: A. Fernández, M.J. del Jesus

Colaboradores: C. Carmona, S. García, M.D. Pérez, A. Rivera

Descripción: En esta etapa se efectúa la codificación de algoritmos de procesamiento y extracción de conocimiento definidos durante las tareas A, B y C.

Temporización: M6-M16: Codificación de los algoritmos

Objetivo Desarrollado: Objetivos 1, 2 y 3

Resultado previsible: Obtención de diferentes modelos para la mejora del comportamiento y prestaciones en los diferentes marcos de trabajo definidos.

Tarea E) Evaluación.

Coordinación: A. Fernández

Colaboradores: M.J. del Jesús, S. García, A. Rivera, M.D. Pérez

Descripción: Llegados a esta etapa es importante revisar concienzudamente tanto la técnica como los resultados proporcionados. Se revisan los pasos llevados a cabo en la aplicación de los algoritmos de clasificación para asegurarnos que se adecúan, tanto los métodos como los modelos obtenidos, a los objetivos perseguidos.

Temporización: M16-M19: Evaluación de los modelos obtenidos.

Objetivo Desarrollado: Objetivos 1, 2 y 3

Resultado previsible: Tras esta etapa se decide la utilidad de los modelos obtenidos con las técnicas de clasificación implementadas con respecto a los requerimientos de precisión y capacidad de generalización que se han fijado como base al principio de la realización del proyecto con respecto al estado del arte.

Tarea F) Experimentación y análisis

Coordinación: A. Fernández, M.J. del Jesus

Colaboradores: C. Carmona, A. Rivera, M.D. Pérez

Descripción: Una vez que la codificación quede libre de errores, y los resultados sean a priori satisfactorios, continuaremos con una validación de las implementaciones. Para ello, generaremos diferentes resultados para obtener la precisión de los modelos aprendidos, y realizaremos un análisis estadístico de los mismos para dar un mayor soporte al análisis experimental. Por otro lado se difundirán los resultados de dicha investigación a través de diferentes medios como pueden ser congresos nacionales e internacionales, revistas científicas y vía web.

Temporización: M19-M24: Implantación de los modelos obtenidos.

Objetivo Desarrollado: Objetivos 1, 2 y 3

Resultado previsible: Implantación de los modelos obtenidos y publicaciones de carácter científico.

Tarea G) Extensión de los modelos al entorno Big Data

Coordinación: A. Fernández

Colaboradores: C. Carmona, S. García, A. Rivera, M.D. Pérez

Descripción: Estudio preliminar sobre la adaptación de los métodos del estado del arte, junto con los desarrollados en el proyecto, para su aplicación en problemas tipo Big Data.

Temporización: M20-M24: Extensión a Big Data.

Objetivo Desarrollado: Objetivo 4

Resultado previsible: Directrices para la implementación en un entorno paralelo (tipo MapReduce) de los algoritmos diseñados.

Cronograma asociado al desarrollo del proyecto:

Actividades/Tareas	Persona responsable y otras involucradas	Primer año	Segundo año
T.A. Análisis clases difíciles		x x x x	
T.B. Estudio problemas clasificación no balanceada.		x x x x	
T.C. Análisis problemas Con múltiples clases		x x x x	
T.D. Desarrollo e implementación de los modelos		x x x x x x x	x x x
T.E. Evaluación			x x x x
T.F. Experimentación y análisis			x x x x x
T.G. Extensión a Big Data			x x x x

Medios disponibles y requeridos para la ejecución del proyecto

Los medios necesarios para la correcta puesta en marcha del proyecto son, fundamentalmente, informáticos. Así, tanto el almacenamiento, preprocesamiento y tratamiento de datos, generación de modelos de clasificación, obtención y análisis de resultados se realizarán mediante herramientas informáticas que incluyen hardware y software.

Actualmente, los medios disponibles son:

- Hardware
  - Computadores de los investigadores implicados (sistemas bi-procesador y quad, con antigüedades que oscilan de 1 a 4 años).
  - 1 Servidor de Cálculo perteneciente al grupo de investigación: 8 máquinas Intel(R) Xeon(TM) CPU 3.20GHz (64 bits) con 1GB Ram (400 Mhz), con discos de 120 GB y 8 máquinas Intel(R) Xeon(R) CPU 3065 @ 2.33GHz (64 bits) con 7 GB Ram (800 Mhz), con discos de 250 GB, 6 máquinas Intel® 24 núcleos (64 bits) con 64GB Ram (800 Mhz) con discos de 250GB.
  - Servidores de Cálculo de CEATIC: 1 nodo central, 16 nodos de cómputo (16 núcleos y 32 GB Ram) + Inifiniband
  - Periféricos (fundamentalmente para impresión)
- Software
  - Sistemas operativos Linux y Windows 7 Professional.
  - Entornos de desarrollo para el lenguaje de programación JAVA: Netbeans, Eclipse.
  - Entornos de análisis estadístico de los resultados obtenidos: SPSS, StatGraphics, R.
  - Software ofimático: OpenOffice, MS-Office.

Los medios requeridos dependen, en gran medida, tanto de los entornos de desarrollo que se deben utilizar como (muy especialmente) del tipo de algoritmos a desarrollar. En efecto, todo este software consume una muy considerable cantidad de recursos informáticos, fundamentalmente memoria principal, CPU y memoria masiva.

Ante la creciente mejora de las prestaciones de los equipos informáticos, se considerará la valoración de incorporar nuevos sistemas en el segundo año del proyecto, de forma que puedan acortarse los tiempos de ejecución de los algoritmos y acelerar la obtención de resultados.

De forma resumida, los medios requeridos para llevar a cabo el proyecto son:

- Hardware
  - Al menos un nuevo equipo informático que será adquirido durante en segundo año con lo cual dispondrá de mayores prestaciones que los existentes en la actualidad
  - Sistemas de almacenamiento masivo para distribuir y compartir las bases de datos objeto de estudio.
- Software
  - Actualizaciones de las herramientas disponibles en la actualidad
  - Paquetes estadístico y de minería de datos comerciales como SAS, CLEMENTINE, etc…

Dado que todo el software requerido es o bien Software Libre o bien Propietario pero con licencia para la Universidad de Jaén, los gastos requeridos en la adquisición/actualización del mismo son mínimos. En caso de necesitar licencias de software avanzado, la propuesta económica también recoge dicha necesidad.

Propuesta Económica

Gastos de Personal	4.000,00 €
Gastos de Ejecución	6.000,00€
Material Inventariable	3.000,00 €
Material Fungible	350,00 €
Material bibliográfico	150,00 €
Inscripción, desplazamiento y dietas para 1 congreso internacional	1.750,00 €
Inscripción, desplazamiento y dietas para 1 congresos nacional	750,00 €
Otros gastos complementarios necesarios para el desarrollo del proyecto directamente relacionados con la actividad y debidamente justificados (Realizar breve descripción)	0,00 €
TOTAL	**10.000,00 €**

Plan de difusión de resultados

La difusión de conocimientos y transferencia de los resultados se llevará a cabo en una doble vertiente: por un lado los habituales canales de divulgación científica: publicaciones en revistas, congresos, etc., y por otro con la publicación en prensa y vía web.

Con la finalidad de centralizar toda la información y logros alcanzados se ha desarrollará una web del proyecto, donde se expondrán todas las publicaciones asociadas al proyecto así como los resultados más relevantes alcanzados. Del mismo modo incluiremos en esta web una recopilación de la bibliografía más relevante en el ámbito del proyecto lo cual puede ser un buen mecanismo de difusión de la bibliografía básica en este tema de interés y puede fomentar la colaboración entre otros grupos de investigación interesados.

Es importante destacar que contamos con la distribución gratuita de los prototipos software base que servirán como punto de partida para generar modelos acorde con los objetivos planteados. Adicionalmente, los nuevos algoritmos implementados se pondrán también a disposición de la comunidad investigadora de con el sistema de "código abierto". Por su estrecha relación, para facilitar la gestión de toda la información y los logros que se generen en el proyecto se utilizará la Web KEEL (http://www.keel.es), donde se expondrán todas las publicaciones, los modelos desarrollados que se podrán descargar, y se mantendrán actualizados listados de referencias bibliográficas en las diferentes áreas de interés para el proyecto.

KEEL ya se conoce en el presente por tres sitios web muy prestigiosos y conocidos que contienen referencias con herramientas software.

Google Directory **Software**

Computers > Artificial Intelligence > Machine Learning > Software

http://www.google.com/Top/Computers/Artificial_Intelligence/Machine_Learning/Software/

KDnuggets : Software : Suites

Software Suites for Data Mining and Knowledge Discovery

Free and Shareware http://www.kdnuggets.com/software/suites.html

The Data Mine

http://www.the-data-mine.com/bin/view/Software/DataMiningSoftware

Nos proponemos seguir el siguiente criterio a la hora de difundir los resultados obtenidos:

- Asistencia a congresos nacionales e internacionales que aborden la temática tratada en el proyecto. La finalidad es doble, por un lado presentar los resultados y por otro contactar con grupos de investigación que aborden temas similares. En el ámbito de la informática, daremos especial importancia a los congresos españoles en Minería de Datos (TAMIDA), sistemas inteligentes (CAEPIA) y lógica difusa (ESTYLF), y los congresos internacionales de aprendizaje automático y Soft-computing (ISDA, HAIS, HIS, FUSION).
- Publicaciones en revistas especializadas donde se suelen publicar las investigaciones relativas a minería de datos, soft computing y sus aplicaciones. Por ejemplo, podemos resaltar: Information Sciences, Soft Computing, Pattern Recognition, Knowledge based Systems, WIREs: Data Mining and Knowledge Discovery, Applied Soft Computing, Expert Systems with Applications, Expert Systems, Applied Intelligence, Applied Artificial Intelligence, entre otros. En algunas de ellas los miembros del proyecto ya han publicado en ocasiones anteriores.
- Capítulos de libros de investigación de editoriales internacionales, tal y como hemos hecho en los últimos años por invitación expresa de autores de prestigio en el área.

Repercusión esperable de los resultados, tanto en su impacto bibliométrico como impacto por transferencia de conocimiento/tecnología

Por último, y no menos importante, destacar que el contexto de trabajo de este proyecto, permite un alto grado de internacionalización debido a diversas características. En concreto, la repercusión esperable puede verse a diferentes niveles:

a) Las características asociadas a este tema de investigación permite desarrollar una línea de trabajo con un alto grado de continuidad por el grupo de trabajo asociado. De esta forma, no se trata de un proyecto aislado para una aplicación concreta, si no la posibilidad de arrancar, mediante este estudio detallado, una vía de investigación de gran productividad de cara a los próximos años.
b) Desde la perspectiva comercial, cualquiera de las técnicas desarrolladas en Minería de Datos, y particularmente en clasificación, tiene un carácter totalmente aplicado. En particular, debemos destacar que la temática elegida entra dentro del contexto de un gran número de aplicaciones de carácter interdisciplinar sobre problemas en diferentes campos de ingeniería, salud, comunicaciones, etc.
c) Adicionalmente, la gestión de cualquier proyecto en un escenario comercial real, no solo nos otorga una trasferencia clara de conocimiento, sino que además puede llevar a la mejora de los métodos propuestos y al diseño de nuevos y mejores métodos.
d) Con respecto a los puntos anteriores, encontramos un gran interés de las empresas en el campo de la Inteligencia de Negocio (Business Intelligence). De este modo, puede dar lugar a un número importante de colaboraciones y/o propuestas de trabajo, por ejemplo en el marco de los Proyectos Horizonte 20-20.
e) Una de las vías futuras de trabajo está relacionada con la resolución en problemas de tipo Big Data, es decir, aquéllas aplicaciones que presentan un alto volumen de información, velocidad en la llegada de los datos y respuesta, y variedad de la propia información. Siendo ésta una temática altamente prometedora, así como muy demandada por parte de las empresas, la adaptación de los modelos desarrollados en esta línea puede conducir a un importante número de proyectos futuros.
f) El impacto bibliométrico se verá reflejado en la participación en congresos nacionales e internacionales tanto relacionadas con ciencias de la computación, como también aplicaciones informáticas. Mucho más importante es destacar la difusión y publicación en revistas situadas en el marco del índice JCR. Del mismo modo se publicaran los resultados vía web para su acceso público y en prensa.

Referencias:

[Alc09] Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M.J., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, J., Rivas, V.M., Fernández, J.C., Herrera, F. (2009) KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing 13(3): 307-318

[Alc11] Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L., 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.

[All00] Allwein, E. L., Schapire, R. E., Singer, Y., Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research 1 (2000) 113–141.

[Alp04] Alpaydin E. (2004) Introduction to Machine Learning. The MIT Press

[Ana95] Anand R., Mehrotra K., Mohan C. K., y Ranka S. (1995) Efficient classification for multiclass problems using modular neural networks. IEEE Transactions on Neural Networks 6(1): 117–124

[Ana09] Anand A. y Suganthan P. N. (2009) Multiclass cancer classification by support vector machines with class-wise optimized genes and probability estimates. Journal of Theoretical Biology 259(3): 533–540

[Ara10] Aran O. y Akarun L. (2010) A multi-class classification strategy for fisher scores: Application to signer independent sign language recognition. Pattern Recognition 43(5): 1776–1788

[Bat04] Batista G. E. A. P. A., Prati R. C., y Monard M. C. (2004) A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations Newsletter 6(1): 20-29.

[Bat10] Batuwita, R., Palade, V. (2010). FSVM-CIL: Fuzzy support vector machines for class imbalance learning. IEEE Transactions on Fuzzy Systems 18(3):558–571.

[Ben07] Ben-David, A. (2007) A lot of randomness is hiding in accuracy. Engineering Applications of Artificial Intelligence 20: 875–885

[Ben08] Ben-David, A. (2008) Comparison of classification accuracy using Cohen's Weighted Kappa. Expert Systems with Applications 34: 825–832

[Buy11] Buyya, R., Broberg, J., Goscinski, A. (2011) Cloud Computing: Principles and Paradigms. Wiley, Chichester.

[Cla91] Clark P. y Boswell R. (1991) Rule induction with CN2: Some recent improvements. En EWSL'91: Proc. of the European Working Session on Machine Learning, páginas 151–163. London, UK.

[Dea08] Dean J. and Ghemawat S. (2008) MapReduce: simplified data processing on large clusters. Commununications of the ACM 51(1): 107-113.

[Den10] Denil, M., Trappenberg, T., 2010. Overlap versus imbalance. In: Proceedings of the 23rd Canadian conference on Advances in artificial intelligence (CCAI'10). Vol. 6085 of Lecture Notes on Artificial Intelligence. pp. 220–231.

[Die95] Dietterich T. G. y Bakiri G. (1995) Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2: 263–286

[Dud01] Duda R. O., Hart P. E., y Stork D. G. (2001) Pattern Classification. John Wiley, 2nd edition

[Fay96] Fayyad U., Piatetsky-Shapiro G., and Smyth P. (1996) From data mining to knowledge discovery in databases. AI Magazine 17(3): 37-54.

[Fer08] Fernandez A., García S., del Jesus M. J., and Herrera F. (2008) A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. Fuzzy Sets and Systems 159(18): 2378-2398.

[Fer09] Fernandez A., del Jesus M. J., and Herrera F. (2009) Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. International Journal of Approximate Reasoning 50(3): 561-577.

[Fer10] Fernandez, A., del Jesus, M.J., 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

[Fer10b] Fernandez, A., García, S., Luengo, J., Bernadó-Mansilla, E., Herrera, F. (2010) Genetics-Based Machine Learning for Rule Induction: State of the Art, Taxonomy and Comparative Study. IEEE Transactions on Evolutionary Computation 14(6): 913-941

[Fer13] Fernández, A., López, V., Galar, M., del Jesus, M. J., Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. Knowledge-Based Systems 42:97–110.

[Fur02] Fürnkranz J. (2002) Round robin classification. Journal of Machine Learning Research 2: 721–747

[Gal11] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F. (2011) An Overview of Ensemble Methods for Binary Classifiers in Multi-class Problems: Experimental Study on One-vs-One and One-vs-All Schemes. Pattern Recognition 44(8): 1761-1776.

[Gal12] Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F. (2012). A review on ensembles for class imbalance problem: Bagging, boosting and hybrid based approaches. IEEE Transactions on Systems, Man, and Cybernetics–part C: Applications and Reviews 42(4): 463–484.

[Gal13] Galar, M., Fernández, A., Barrenechea, E., Herrera, F. (2013) EUSBoost: Enhancing Ensembles for Highly Imbalanced Data-sets by Evolutionary Undersampling. Pattern Recognition 46(12): 3460–3471

[Gal14] Galar, M., Fernández, A., Barrenechea, E., Herrera, F. (2014). Empowering difficult classes with a Similarity-based aggregation in multi-class classification problems. Information Sciences 264: 135-157

[Gar08] García, V., Mollineda, R. A., Sánchez, J. S., (2008). On the k–NN performance in a challenging scenario of imbalance and overlapping. Pattern Analysis Applications 11(3–4): 269–280.

[Gul07] Guler I. y Ubeyli E. D. (2007) Multiclass support vector machines for EEG-signals classification. IEEE Transactions on Information Technology in Biomedicine 11(2): 117–126

[Has98] Hastie, T., Tibshirani, R., (1998) Classification by pairwise coupling. The Annals of Statistics 26(2): 451–471.

[He09] He H. y Garcia E. A. (2009) Learning from imbalanced data. IEEE Transactions On Knowledge And Data Engineering 21(9): 1263-1284.

[Hua06] Huang, Y.-M., Hung, C.-M., Jiau, H.C. (2006) Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem Nonlinear Analysis: Real World Applications 7 (4), pp. 720-747

[Hul10] Hüllermeier, E., Vanderlooy, S. (2010) Combining predictions in pairwise classification: an optimal adaptive voting strategy and its relation to weighted voting, Pattern Recognition 43(1): 128–142.

[Hul09] J. Van Hulse and T.M. Khoshgoftaar. Knowledge discovery from imbalanced and noisy data. Data and Knowledge Engineering 68:12 (2009), 1513-1542

[Kne90] Knerr S., Personnaz L., y Dreyfus G. (1990) Single-layer learning revisited: A stepwise procedure for building and training a neural network. En Fogelman Soulié F. y Hérault J. (Eds.) Neurocomputing: Algorithms, Architectures and Applications, volumen F68 of NATO ASI Series, páginas 41–50. Springer-Verlag

[Lam11] Lam C. (2011) Hadoop in action. Manning Press.

[Lin13] Lin, M., Tang, K., Yao, X., (2013). Dynamic sampling approach to training neural networks for multiclass imbalance classification. IEEE Transactions on Neural Networks and Learning Systems 24(4): 647–660

[Liu09] Liu K. H. y Xu C. G. (2009) A genetic programming-based approach to the classification of multiclass microarray datasets. Bioinformatics 25(3): 331–337

[Liu12] Liu L. y Fieguth P. (2012) Texture classification from random features. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(3):574-586

[Lop12] López, V., Fernandez, A., Moreno-Torres J.G., Herrera, F. (2012) Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. Expert Systems with Applications 39(7): 6585-6608

[Lop13] López, V., Fernandez, A., García, S., Palade, V., Herrera, F. (2013). An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. Information Sciences 250: 113-141

[Lop13b] López, V., Fernandez, A., del Jesus, M.J., 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: 85-104,

[Lop14] López, V., Fernandez, A., Herrera, F. (2014) On the Importance of the Validation Technique for Classification with Imbalanced Datasets: Addressing Covariate Shift when Data is Skewed. Information Sciences 257: 1-13

[Lor08] Lorena A. C., Carvalho A. C., y Gama J. M. (2008) A review on the combination of binary classifiers in multiclass problems. Artificial Intelligence Review 30(1-4): 19–37.

[Mar13] Marx, V. (2013) The big challenges of big data. Nature 498(7453):255-260

[Ngu11] Nguyen, T.T., Chang, K., Hui, S.C. (2013) Supervised term weighting centroid-based classifiers for text categorization Knowledge and Information Systems 35 (1), pp. 61-85

[Oen14] Oentaryo R., Lim, E.-P., Finegold, M., Patel, D., Berrar, D. (2014) Detecting click fraud in online advertising: A data mining approach, Journal of Machine Learning Research 15 pp 99-140

[Orr09] Orriols-Puig A., Bernadó-Mansilla E., Goldberg D. E., Sastry K., and Lanzi P. L. (2009) Facetwise analysis of XCS for problems with class imbalances. IEEE Transactions on Evolutionary Computation 13:260-283

[Pau09] Paul T. K. y Iba H. (2009) Prediction of cancer class with majority voting genetic programming classifier using gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6(2):353–367.

[Per10] Pérez-Godoy, M.D., Fernández, A., Rivera, A.J., del Jesus, M.J. (2010) Analysis of an Evolutionary RBFN Design Algorithm CO2RBFN for Imbalanced Data-Sets. Pattern Recognition Letters 31(15): 2375-2388

[Pol06] Polikar R. (2006) Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 6(3): 21 – 45.

[Puj06] Pujol O., Radeva P., y Vitria J. (2006) Discriminant ECOC: a heuristic method for application dependent design of error correcting output codes. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(6): 1007–1012

[Pyl99] Pyle D. (1999) Data Preparation for Data Mining. Morgan Kaufmann.

[Rif04] Rifkin R. y Klautau A. (2004) In defense of one-vs-all classification. Journal of Machine Learning Research 5: 101–141

[Rok10] Rokach L. (2010) Ensemble-based classifiers. Artificial Intelligence Review 33: 1–39.

[Sei14] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., Folleco, A., (2014). An empirical study of the classification performance of learners on imbalanced and noisy software quality data. Information Sciences 259: 571-595.

[Sta08] Stamatatos E. (2008) Author identification: Using text sampling to handle the class imbalance problem. Information Processing & Management 44(2): 790–799

[Sun09] Sun Y., Wong A. K. C., y Kamel M. S. (2009) Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence 23(4): 687-719.

[Tan06] Tan P. N., Steinbach M., and Kumar V. (2006) Introduction to Data Mining. Addison-Wesley

[Tao04] Tao L., Chengliang Z., y Mitsunori O. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20(15): 2429–2437

[Tor07] Torralba A., Murphy K. P., y Freeman W. T. (2007) Sharing visual features for multiclass and multiview object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(5): 854–869.

[Vil12] Villar, P., Fernandez, A., Carrasco, R.A., 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(3): 369-397

[Wei10] Weiss G. M. (2010) The impact of small disjuncts on classifier learning. In Stahlbock R., Crone S. F., and Lessmann S. (Eds.) Data Mining, volumen 8 of Annals of Information Systems, pp. 193-226. Springer

[Yua11] Yuan J., Liu Z., y Wu Y. (2011) Discriminative video pattern search for efficient action detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(9): 1728–1743

[Zad03] Zadrozny, B., Langford, J., Abe, N., (2003). Cost–sensitive learning by cost–proportionate example weighting. In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03). pp. 435–442.

[Zho10] Zhou, Z.-H., Liu, X.-Y. (2010). On multi-class cost-sensitive learning, Computational Intelligence 26:3 232–257

[Zie14] Zięba, M., Tomczak, J.M., Lubicz, M., Świątek, J. (2014) Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied Soft Computing Journal 14 (PART A), pp. 99-108

[Zon13] Zong, W., Huang, G.-B., Chen, Y., (2013). Weighted extreme learning machine for imbalance learning. Neurocomputing 101, 229–242.