Convocatoria de ayudas de Proyectos de Investigación Fundamental no orientada

TECHNICAL ANNEX FOR TYPE A or B PROJECTS

1. SUMMARY OF THE PROPOSAL (the summary must be also filled in Spanish)

PROJECT TITLE: New Problems in Knowledge Discovery: A Genetic Programming Approach

PRINCIPAL INVESTIGATOR: Sebastián Ventura Soto

SUMMARY
(brief and precise, outlining only the most relevant topics and the proposed objectives):

The Project iNsPIrED (New Problems In knowlEdge Discovery) has the main objective of developing new knowledge discovery methodologies using genetic programming and other evolutionary computation approaches, as well as their application in several real-world problems. This main objective can be split into the following secondary objectives:

1) Development of genetic programming models for solving different problems in knowledge discovery: multiple instance learning, relational learning, multi-label classification and mining association rules.
2) Adaptation of the models developed to deal with new challenges associated to high dimensionality problems, large datasets and imbalanced data.
3) Application of the developed models to real problems in the context of educational data mining (new representation models for predicting student’s performance in LMS, modeling students drop-out, categorization of learning objects) and web mining (intrusion detection and web categorization problems).
4) Development of data repositories enabling the scientific community to compare our findings with other existing proposals, and integration of the models developed in the KEEL and WEKA software platforms, in order to facilitate their promotion.

TITULO DEL PROYECTO: Nuevos Retos en Descubrimiento de Conocimiento: Un Enfoque Basado en Programación Genética

RESUMEN
(breve y preciso, exponiendo sólo los aspectos más relevantes y los objetivos propuestos):

El proyecto iNsPIrED (New Problems In knowlEdge Discovery) tiene como principal objetivo el desarrollo de nuevas metodologías de extracción de conocimiento mediante técnicas de programación genética, y su aplicación a distintos problemas reales. Este objetivo fundamental puede desglosarse en los siguientes objetivos secundarios:

1) Desarrollo de modelos para la resolución de distintos problemas de extracción de conocimiento: aprendizaje con multi-instancias, aprendizaje relacional, aprendizaje multi-etiqueta y obtención de reglas de asociación.
2) Adaptación de los modelos anteriores a nuevos problemas con diferentes tipos de datos: datos con gran número de variables, conjuntos de datos de gran tamaño y datos no balanceados.
3) Aplicación de los modelos desarrollados a problemas reales en el ámbito de la minería de datos educativos y de la minería de datos web. En el primer caso, pretendemos abordar problemas tales como la predicción del rendimiento académico de estudiantes, la predicción del abandono de estudiantes (*drop-out*) y la categorización de los elementos disponibles en repositorios de objetos de aprendizaje. En el segundo caso, nos centraremos en problemas de detección de intrusos y categorización web.
4) Desarrollo de repositorios de datos que permitan a la comunidad científica la comparación de resultados e integración de los modelos desarrollados en las plataformas software de mayor relevancia actual (KEEL y WEKA), para facilitar la difusión de los mismos.

2. INTRODUCTION
(maximum 5 pages)

PROPOSAL, STATE OF THE ART OF SCIENTIFIC AND TECHNICAL KNOWLEDGE

Evolutionary Algorithms (EA) [Eib10, Yu10] are search algorithms inspired in genetics and the natural processes of evolution. This kind of algorithms has been successfully used in multiple complex problems and, they have recently revealed their suitability as a technique for knowledge discovery (KDD) [Han06], having been employed in tasks such as prediction (classification and regression), pattern extraction, clustering, feature selection and instance selection. A proof of this fact is the plethora of scientific articles published, as well as the monographs recently edited, that show the latest contributions to the field [Fre02, Bull10, Gho10, Mai10].

From the paradigms that comprise evolutionary computation (EC), genetic programming (GP) [Pol08] is a technique that offers great potential for solving knowledge discovery problems. In GP, individual representation is more flexible than in other EC paradigms, directly allowing the representation of decision trees, rule bases or complex mathematical expressions [Ban98, Iba09]. This represents an advantage over other EC paradigms, like genetic algorithms, where the lineal structure of individuals imposes an important limitation on representing KDD entities [Esp10]. On the other hand, if the syntactic form of the desired solution is known in advance, we can use GP variants that use grammars to express restrictions in the representation space. In these paradigms, known generically as Grammar-Based Genetic Programming [McK10], the grammar avoids the possibility of generating invalid individuals, biasing the creation of new individuals by crossover and mutation, increasing the efficiency of the evolutionary process. Furthermore, the combined use of the grammar and other mechanisms for controlling the size of the expressions that evolve allows an improvement in the interpretability of results, a highly desirable property in knowledge discovery algorithms [Han06].

In the last few years, new problems of great interest have arisen in machine learning and data mining communities. Some of them, like multiple instance learning [Die97], multi-label learning [Tso09] or relational learning [Dze03] are related to presenting information in a more flexible way. In other cases, like learning from imbalanced data [Sun09] or the identification of infrequent association rules [Wei04], the difficulty has to do with an asymmetrical distribution of learning data. Finally, problems like learning from high dimensionality data or learning from huge datasets have been solved with preprocessing techniques like feature selection [Guy06, Liu07b] or instance selection [Liu10], or defining easy-to-scale proposals, based on high performance computing [Hag10]. Some of these new problems have already been broached from an evolutionary learning perspective, obtaining competitive solutions. Nevertheless, there is work to be done in this field.

Based on the foregoing considerations, the main objective in this project is the application of genetic programming techniques to resolve these new problems in knowledge discovery. The project will be developed from a double perspective: first, new genetic programming proposals for solving different knowledge discovery problems will be proposed, tested and compared with other state-of-the-art algorithms; second, the developed proposals will be applied to real problems in two application domains: educational data mining and web mining. The following sections briefly state the problems and application domains that are the subject of this proposal.

Multiple instance learning (MIL)

Multiple instance or multi-instance (MI) learning, introduced by Dietterich et al. [Die97], has attracted a significant amount of interest since its inception, primarily because several real-world learning tasks, among them drug activity prediction [Die97], content-based image retrieval [Pao08] and web content mining [Zaf09], can be very naturally represented in the MI setting, but also because MI learning is a stepping stone on the way from propositional to relational learning [Reu04]. MI learning differs from standard single-instance (SI) inductive learning in that examples for learning are more complex. In SI learning, an example is described by a fixed-length vector of attribute values, normally called an instance. In MI learning, an example consists of a bag of instances, rather than a single instance, and different examples can contain different numbers of instances. Often these different instances represent different views of the same object (e.g. possible shapes of a molecule as in the case of drug activity prediction).

Recent years have seen the development of algorithms for binary classification data from MI. Reviewing the literature[1] we can find algorithms designed specifically for this problem, such as APR [Die97] and Diverse Density (DD) [Mar97, Pao08] and algorithms inspired on traditional supervised learning methods such as KNN [Zho05], decision trees [Che01], classification rules [Che01], logistic regression [Xu04], neural networks [Zha06a], support vector machines [Gar04, Che06] and ensembles [Zho07a]. Almost all these models are based on the hypothesis made by Dietterich, where a pattern is considered positive if at least one of the instances contained in the bag represents the concept that we want to learn while it is considered negative otherwise [Die97]. However, other hypotheses have shown to be more appropriate for certain specific problems [Wei03, Sco05].

Recently, we have carried out grammar guided genetic programming models for multiple instance learning [Zaf10, Zaf11a], and feature selection methods particularly suitable for this learning paradigm [Zaf11b]. However, there are still several open issues that we intend to tackle in this project:

[1] A complete bibliography about the topic, compiled by our research group can be found in http://www.uco.es/grupos/kdis/kdiswiki/mil.

- *Development of new models for MIL with imbalanced data sets*. Imbalanced datasets have not been addressed in the MIL context. However, there are a significant number of problems (for example, semantic object categorization) affected by this question.
- *Development of new proposals to work with generalized multiple instance problems*. Most models developed to date are based on classic hypotheses formulated by Dietterich et al. [Die97]. The aim is to extend these models to take into account the hypotheses based on thresholds and ranges [Wei03, Sco05].
- *New feature selection methods for MI data*. The feature selection methods developed to date use the Dietterich et al. Hypothesis [Die97]. The aim is to study if these methods are also valid in generalized multiple instance problems, and if not, new specific methods will be designed.

Multi-label classification (MLC)

In multi-label (ML) classification, as opposed to the traditional single-classification problem, the examples are associated with more than one class label. In the past, the problem was mainly motivated by tasks of text categorization and medical diagnosis. Nowadays, we notice that MLC methods are increasingly required by modern applications, such as protein function classification, music categorization and semantic scene classification [Tso09].

Multi-label classification methods can be categorized into two different groups: problem transformation methods and algorithm adaptation methods [Tso07]. The former approach transforms a multi-label dataset into one [Tso09] or several [Tso07] single-label datasets to be processed by a classical classification algorithm while the latter group of methods extends specific learning algorithms in order to handle multi-label data directly. Regarding algorithm adaptation methods, a number of proposals based on well-known single-label algorithms have been studied: support vector machines (SVM) [Eli01], lazy learning [Zha05], neural networks [Zha06b], decision trees [Cla01, Noh04], and ensemble methods [Tso07, Rea08]. A complete bibliography on the topic has been compiled by our research group and can be found at http://www.uco.es/grupos/kdis/kdiswiki/mlc.

Our research group has developed several specific models for MLC which are based on genetic programming and have obtained very competitive results with respect to other proposals [Avi10a, Avi10b]. The experience obtained to date leads us to consider new challenges related to this topic of interest such as:

- *Development of new methods for MLC*. The proposals developed to date by this group can be improved. In this respect, our aim is to improve the scalability and efficiency of the algorithms developed. Another important goal is the development of more precise models by taking into account the relationships between labels.
- *Dimensionality reduction.* Most multi-label datasets present a great number of patterns and attributes. Due to this fact, it is essential to study the dimensionality reduction methods specifically developed to carry out multi-label classification and to take into account aspects such as the relationship between labels.
- *MI-ML classification*. There are problems, like text categorization, where documents can be seen simultaneously from the perspective of MI and ML. It is a very recent research line [Zho07b, Zha08a] that represents the convergence of our research interests in MLC and MIL.

High Performance Genetic Programming with GPGPU

As mentioned, GP has been used successfully to obtain models for classification [Esp10]. However, there is still work to be done in improving the scalability and efficiency of such algorithms with respect to their application in large datasets. GPU computing or GPGPU [Owe08] is a computer model that uses general purpose graphics processing units (GPU) to perform tasks with a high computational cost. GPUs are high-performance and many-core processors capable of performing tasks over multiple data in parallel. In recent years there have been numerous publications and software that use the potential of the GPU to improve the efficiency of algorithms in many fields [Che08]. Data mining techniques especially benefit from the GPU to solve high-dimensional and parallelizable problems such as classification and association. These techniques include evolutionary algorithms [Won06] and, specifically, genetic programming [Lan10a, Lan10b, Wil10].

The group has started this research line recently and is working on developing new implementations based on the use of the GPGPU [Can10]. The results obtained are very promising, producing a speedup near 600 in some cases, and allowing us to address problems with datasets of over 1 million records. We intend to advance this line by developing highly efficient versions of our algorithms, which allows us to address high dimensionality problems (many examples and attributes) such as those encountered in the field of web content mining.

Association rule mining (ARM)

One point of particular interest in any data mining task is to find repeating patterns, trends or rules that explain the behavior of the data in a given context. More specifically, if we concentrate on association rule mining (ARM) [Agra93, Han04], the aim is to find common relationships between elements (called items) in large databases. An association rule is defined as an implication of the form $A \rightarrow C$, where A and C are itemsets that have no attributes in common, i.e. $A \cap C = \emptyset$, A being A the antecedent and C the consequent of the rule. Thus, an association rule is interpreted as, if all the items in A are present in a transaction, then it is quite likely that the items in C are also present in the transaction. These techniques have already been successfully used in many different application contexts, including medicine [Ord06], biology [Eom06], credit fraud detection [San08], and intrusion detection [Lun10c], among others[2].

ARM classical algorithms such as Apriori [Agr93] or FP-Growth [Han04] carry out an exhaustive search for items whose frequency of occurrence is greater than or equal to a given threshold. This frequent itemset is then used to extract association rules. Both steps of the process have important limitations, such as the high computational cost required for the extraction of frequent items, or the large amount of memory used in the construction of the rules. Furthermore, it is often necessary to apply intensive preprocessing techniques before running these algorithms (e.g. the discretization of the dataset), which may cause accuracy problems.

In recent years genetic algorithms (GA) have been used to address the problem of ARM [Mat02, Yan05, Sal07, Yan09]. These proposals have major advantages over previous approaches, especially when covering large search spaces, since they allow its execution to involve different representations and obtain optimal solutions in tolerable dimensions of time. However, its main drawback is related to the fixed size of individuals, which forces association rules to have a preset size, both in the antecedent and the consequent. So, with the use of these proposals for ARM, the problem of high computational cost and the need for large amounts of memory required by classical algorithms is eliminated, as mentioned above. As already discussed, genetic programming (GP) [Ban98] permits us to represent rules in a more flexible way, allowing the establishment of syntactic constraints [McK10] to define the structure of association rules. This type of approach, namely the so-called Grammar Guided Genetic Programming (G3P) has already been employed by our research group [Lun10a, Lun10b], obtaining very competitive results in terms of scalability, adaptability, accuracy and execution time. Throughout this project, we intend to extend our previous research in this field, resolving also the following new challenges related to the theme of the ARM:

- *Rare Association Rule Mining* (*RARM*). DM is not only useful for discovering frequent patterns, but also for those infrequent items that are often even more relevant than those that appear more frequently (e.g. diseases, frauds, etc.). This is the main reason why this recent research area has been considered of great interest in this field.
- *Relational Association Rule Mining*. The extraction of meaningful knowledge from relational databases comprising complex relationships among data has become an important area of interest. We plan to explore the use of new evolutionary exploration techniques in this context, mainly to improve results in terms of accuracy and time.

Educational Data Mining

Educational Data Mining (EDM) is a field that exploits statistical, machine learning and Data Mining (DM) algorithms in different types of educational data [Rom10a, Rom10b]. Its main objective is to analyze these types of data in order to resolve educational research issues. EDM is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn [Bak09, Bak10]. Some examples of tasks of interest in this research area that have been successfully resolved using data mining are: to visualize [Maz04] educational data in order to highlight useful information and to support decision making, to automatically develop student modeling [Fri06], to create groups of students according to their customized features and personal characteristics using clustering [Aye09] and classification [Sup06]. More references and information can be found in http://www.uco.es/grupos/kdis/kdiswiki/edm.

Our interest in this project is to continue our research in the tasks involved in predicting student performance and providing feedback to support instructors and students, and also to deal with new trends such as predicting student drop-out and automatically classifying learning objects. In continuation, we briefly introduce our proposals:

- *Relational and MI representations for predicting student performance in LMS*. One of the most important problems related to creating user modeling in LMS for predicting student marks or scores is the existence of sparse data. This is because there is not usually available the same information about all the activities or interactions done by each student. Therefore, there is a great deal of missing data when all the information is gathered into one single table. This fact can negatively affect the quality of the model obtained from these data. In order to resolve this problem we propose the use of a more flexible form of data representation (relational and MI). Thus, we try to improve the prediction process even in cases with high percentages of missing data.
- *Predicting student drop-out*. The objective is to look for educational, social, family and personal factors and reasons that affect student drop out (which is when a student abandons a specific academic programme). These data are usually imbalanced so we have to try to rebalance data or to use cost sensitive classification algorithms. We will also compare classical algorithms with the G3P algorithms developed in in other tasks of this project.
- *Providing feedback for supporting instructors and students*. The objective is to provide, on the one hand, feedback to support course authors, teachers and/or administrators in their decision making (about how to improve students' learning, organize instructional resources more efficiently, etc.) and, on the other hand, to make recommendations directly to the students with respect to their personalized activities, as well as to be able to adapt learning contents, interfaces and sequences to each particular student. We have successfully applied traditional association rule mining but in this project we are interested in using new trends such as relational data mining and rare association rules.
- *Classifying learning objects*. The objective is to automatically assign or classify learning resources or Learning Objects (LOs) into one or several subjects, fields or related knowledge domain areas so that they can be found and reused more easily by

[2] A complete bibliography on the topic, compiled by our research group can be found at http://www.uco.es/grupos/kdis/kdiswiki/arm-biblio.

other instructors in other courses. In this project, we will face this problem by using multi-label classification algorithms, in which each LO can be classified in one or several classes (subjects, fields or areas).

Web Mining

Web mining (WM) is the use of data mining techniques to automatically discover and extract information from Web documents and services [Cha03, Liu07a, Mak07]. Although WM uses many data mining techniques, it is not an application of traditional data mining due to the heterogeneity and semi-structured or unstructured nature of the Web data. Web mining tasks can be categorized in three types: web structure mining (WSM – discovery of knowledge from hyperlinks), web content mining (WCM – discovery of knowledge from web page contents) and web usage mining (WUM – discovery of user access patterns from web usage logs). In the last few years, the number of publications in this line has increased enormously. Some interesting bibliographic surveys have also been published [Zha08b, Qi09].

In the field of WCM, traditional classification methods have been applied to several problems and, recently, multi-label classification techniques have been applied to problems of categorization of emotions [Alm05]. Other interesting examples in the field of WUM are: to discover association rules in order to analyze the user's preferences for recommending resources adapted to his/her profile [Bed09, For10, Cas11] and to analyze the behavior of social networks users in order to obtain strategic knowledge about developing new applications or in order to analyze user's browsing/navigation sequences [Boz10].

In this area, we have applied our MIL algorithms for recommending index web pages [Zaf09, Zaf11a]. We have also used algorithms for obtaining association rules in order to analyze the behavior of students in e-learning systems [Rom04] and, more recently, we have applied techniques for obtaining infrequent association rules in problems of intrusion detection in web recommendation systems [Lun10c]. We want to proceed with the previously described lines and also to develop the following new lines:

- *Using MI for representing web pages*. There are references related to the use of MI for document representation [He09]. We want to find out if this representation is also suitable for representing web page content and to apply our MIL and MIL-FS algorithms in different web page categorization tasks such as Web spam detection.
- *Models for intrusion detection.* In general, the problem of obtaining intrusion detection models is a classification task. In this specific problem, there is information available about normal and anomalous behaviors and the model has to discriminate between them. This is an example of a problem with imbalanced data because anomalous behaviors are much less frequent than normal ones, and so we want to evaluate if our algorithms work well with this type of problems and data.

International and National Teams working in the field

Many of the international groups can be found in the bibliographical references given at the end of this section. We must note that most of them publish their research in international conferences such as Parallel Problem Solving from Nature (PPSN) (considered the European conference on Evolutionary Computation), the IEEE Conference on Evolutionary Computation (CEC) and the Genetic and Evolutionary Computation Conference (GECCO) and also in journals such as Evolutionary Computation (MIT), IEEE Transactions on Evolutionary Computation or Genetic Programming and Evolvable Machines. We can also find published contributions in conferences and journals on Pattern Recognition, Machine Learning and Data Mining and Knowledge Discovery.
Regarding the national groups, there are a set of (full or partially) Spanish research groups working on evolutionary learning. The Spanish Conference on Metaheuristics and Evolutionary and Bioinspired Algorithms (MAEB) is one of the meeting places for these researchers. In continuation, we enumerate some of these groups, together with the university to which they belong, the contact person and research areas:

- Carlos III University (Pedro Isasi, evolutionary neural networks and instance selection)
- European Center of Soft Computing, Mieres (Oscar Cordón and Luis Magdalena, genetic fuzzy systems)
- Polytechnic University of Madrid (Pedro Larrañaga, evolutionary learning of Bayesian netwoks, José M. Peña, parallel data mining)
- Ramón Llull University (Josep Mª Garrell and Ester Bernardó, learning classifier systems and data complexity)
- University of Asturias (Luciano Sánchez, genetic fuzzy systems)
- University of Castilla la Mancha (José Antonio Gámez, causal networks and fuzzy systems)
- University of Córdoba (César Hervás, evolutionary neural networks)
- University of Extremadura (Francisco Fernández, genetic programming based learning)
- University of Granada (Francisco Herrera – knowledge discovery with soft computing techniques, Antonio González - genetic fuzzy systems, and Juan Julián Merelo – evolutionary neural networks, evolutionary data streams)
- University of Jaén (María José del Jesus, association rule mining, genetic fuzzy systems)
- University of Málaga (Enrique Alba and Carlos Cotta, evolutionary neural networks and data analysis of bioinformatics)
- University of País Vasco (Jose Antonio Lozano, evolutionary learning of Bayesian networks and EDA)
- University of Santiago de Compostela (Alberto Bugarín and Manuel Mucientes, genetic fuzzy systems and robotics)
- University of Sevilla (José Riquelme, evolutionary learning of rules)

In the framework of learning and data mining we must emphasize the expertise network called *Red Española de Minería de Datos y Aprendizaje*[3] (Spanish Network for Data Mining and Learning), which is subsidized by the Spanish Ministry of Technology and Science and coordinated by Dr. Riquelme. The network meetings serve as meeting and discussion forums for Spanish researchers in this field. These meetings are organized around the TAMIDA workshop (Taller de Minería de Datos). The last editions were organized as workshops of the CEDI Congress, Granada 2005, Zaragoza 2007 and Valencia 2010.

RELEVANT BIBLIOGRAPHY

[Agr93] R. Agrawal, T. Imielinski and A.N. Swami. Mining association rules between sets of items in large databases. *1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C. May 1993, pp. 207–216.

[Alm05] C. O. Alm, D. Roth, and R. Sproat, Emotions from Text: Machine Learning for Text-based Emotion Prediction, *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada, October 2005, pp. 579-586.

[Avi10a] J. L. Ávila, E. L. Gibaja Galindo and S. Ventura. Evolving Multi-label Classification Rules with Gene Expression Programming: A Preliminary Study. *Fifth International Conference on Hybrid Artificial Intelligent Systems,* vol. (2) 2010: 9-16

[Avi10b] J. L. Ávila, E. L. Gibaja, A. Zafra and S. Ventura. A Niching Algorithm to Learn Discriminant Functions with Multi-Label Patterns. *Journal of Multiple-Valued Logic and Soft Computing*, 2010 (accepted).

[Aye09] E. Ayers, R. Nugent and N. Dean. A Comparison of Student Skill Knowledge Estimates. *International Conference on Educational Data Mining*, Córdoba, 2009.

[Bak09] R. Baker and K. Yacef. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1:1 (2009), 3-17.

[Bak10] Baker, R (2010). Data Mining for Education. *International Encyclopedia of Education (3rd edition)*. Oxford, UK: Elsevier, 112-118.

[Ban98] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone, *Genetic Programming: An Introduction*. Morgan Kaufmann, 1998.

[Bed09] P. Bedi, H. Kaur, B. Gupta, J. Talreja and M. Sood. A website recommender system based on an analysis of the user's access log. *Journal of Intelligent Systems* 18:4 (2009) 333-352.

[Boz10] A. S. Bozkir, S. Güzin Mazman and E. Akçapinar Sezer. Identification of user patterns in social networks by data mining techniques: Facebook case. *Communications in Computer and Information Science* 96 (2010), 145-153.

[Bull10] L. Bull, E. Bernadó-Mansilla and J. Holmes (Eds.) . *Learning Classifier Systems in Data Mining*. Springer, 2010.

[Can10] A. Cano, A. Zafra and S. Ventura. Solving Classification Problems Using Genetic Programming Algorithms on GPUs. *Fifth International Conference on Hybrid Artificial Intelligence Systems*, pages 17-26, 2010.

[Cas11] G. Castellano, A. M. Fanelli and M. A. Torsello. NEWER: A system for NEuro-fuzzy WEb recommendation. *Applied Soft Computing* 11:1 (2011), 793-806.

[Cha03] S. Chakrabarti. *Mining the WEB. Discovering Knowledge form Hypertext Data*. Morgan Kaufmann, 2003.

[Che01] Y. Z. Chevaleyre, J. D. Zucker, Solving multiple-instance and multiple-part learning problems with decision trees and decision rules. Application to the mutagenesis problem, in: *AI'01: 14th Conference of the Canadian Society for Computational Studies of Intelligence.* LNCS 2056, 204-214. 2001.

[Che06] Y. Chen, J. Bi, J. Wang, Miles: Multiple-instance learning via embedded instance selection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28:12 (2006) 1931–1947.

[Che08] Che, S. et al., A performance study of general-purpose applications on graphics processors using CUDA, *Journal of Parallel and Distributed Computing*, 68 (10), 1370-1380, 2008.

[Cla01] A. Clare and R. D. King. Knowledge discovery in multi-label phenotype data. *Lecture Notes in Computer Science*, 2168:42_53, 2001.

[Dze03] S. Dzeroski. Multi-Relational Data Mining: An Introduction. *SIGKDD Explorations* 5:1 (2003), 1-16.

[Die97] T. G. Dietterich, R. H. Lathrop and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* 89:1-2 (1997) 31-71.

[Eib10] A. E. Eiben and J.E. Smith. *Introduction to Evolutionary Computing*. Springer 2010.

[Eom06] J. H. Eom and B. T. Zhang. Prediction of Protein Interaction with Neural Network-based Feature Association Rule Mining. *Lecture Notes in Computer Science*, 4234, 30-39, 2006.

[Esp10] P.G. Espejo, S. Ventura and F. Herrera. A Survey on the Application of Genetic Programming to Classification. *IEEE Transactions on Systems, Man and Cybernetics. Part C: Applications and Reviews*, 40:2 (2010), 121-144.

[Eli01] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. *Advances in Neural Information processing Systems*, 14:681_687, 2001.

[For10] R. Forsati and M. R. Meybodi. Effective page recommendation algorithms based on distributed learning automata and weighted association rules. *Expert Systems with Applications* 37:2 (2010), 1316-1330.

[Fre02] A. A. Freitas. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer, 2002.

[Fri06] E. Frias-Martinez, S. Chen and X. Liu. Survey of Data Mining Approaches to User Modeling for Adaptive Hypermedia. *IEEE Transactions on Systems, Man, and Cybernetics-Part C* 36:6 (2006), 734-749.

[Guy06] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh (Eds.). *Feature Extraction: Foundations and Applications*. Springer, 2006

[Gar04] T. Gärtner, J. W. Lloyd and P. A. Flach, Kernels and distances for structured data, *Machine Learning* 57:3 (2004) 205–232.

[Gho10] A. Ghosh, S. Dehuri and S. Ghosh (Eds.). *Multi-Objective Evolutionary Algorithms for Knowledge Discovery from Databases*. Springer, 2010.

[Hag10] G. Hager and G. Wellein. *Introduction to High Performance Computing for Scientists and Engineers.* CRC Press, 2010.

[Han04] J. Han, J. Pei, Y. Yin and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery* 8:53–87, 2004.

[Han06] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*, 2nd ed. Morgan Kaufmann Publishers, 2006.

[He09] W. He and Y. Wang. Text Representation and Classification Based on Multi-Instance Learning. *Sixteenth International Conference on Management Science & Engineering*. Moscow, 2009.

[Iba09] H. Iba, Y. Hasegawa and T. K. Paul. *Applied Genetic Programming and Machine Learning*. CRC Press, 2009.

[Lan10a] Langdon, W.B., Harrison, A.P., GP on SPMD parallel graphics hardware for mega bioinformatics data mining, *Soft Computing*, 12 (12), 1169-1183, 2008.

[Lan10b] Langdon, W.B., Large scale bioinformatics data mining with parallel genetic programming on graphics processing units, *Studies in Computational Intelligence*, 269, pp. 113-141, 2010.

[Liu07a] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2007.

[Liu07b] H. Liu, H. Motoda (Eds.). *Computational Methods of Feature Selection*. CRC Press, 2007.

[Liu10] H. Liu, H. Motoda (Eds.). *Instance Selection and Construction for Data Mining*. Springer, 2010.

[Lun10a] J.M. Luna, J.R. Romero, S. Ventura. Analysis of the Effectiveness of G3PARM Algorithm. In *Proceedings of the 5th International Conference on Hybrid Artificial Intelligent Systems (HAIS)*. San Sebastián, Spain. 2010.

[Lun10b] G3PARM: A Grammar Guided Programming for Mining Association Rules. In Proceedings of the IEEE World Congress on Computational Intelligent (WCCI). Barcelona, Spain. 2010.

[Lun10c] J.M. Luna, A. Ramírez, J.R. Romero, S. Ventura. An Intrusion Detection Approach Based on Infrequent Rating Pattern Mining. *Tenth International Conference on Intelligent Systems, Design and Applications* (ISDA). El Cairo, Egypt. 2010.

[3] http://www.lsi.us.es/redmidas/index.html

[McK10] R. I. McKay, N. X.Hoai, P. A. Whigham, Y. Shan and M. O'Neill. Grammar-based Genetic Programming: a survey. *Genetic Programming and Evolvable Machines*, 11:3-4 (2010), 365-396.

[Mai10] O. Maimon and L. Rokach (Eds.). *Soft Computing for Knowledge Discovery and Data Mining*. Springer, 2010.

[Mak07] Z. Markov, D.T. Larose. Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage. Wiley, 2007.

[Mar97] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, *NIPS'97: Neural Information Processing Systems*, 570-576, 1997.

[Mat02] J. Mata, J. L. Álvarez, and J. C. Riquelme. Discovering Numeric Association Rules via Evolutionary Algorithm, *Lecture Notes in Computer Science* 2336, 40-51. 2002.

[Maz04] R. Mazza and C. Milani. GISMO: a Graphical Interactive Student Monitoring Tool for Course Management Systems, *International Conference on Technology Enhanced Learning*, Milan, 1-8. 2004.

[Noh04] H. G. Noh, M. S. Song, and S. H. Park. An unbiased method for constructing multilabel classification trees. *Computational Statistics & Data Analysis*, 47:1 (2004), 149-164.

[Ord06] C. Ordoñez, N. Ezquerra and C. Santana. Constraining and Summarizing Association Rules in Medical Data. *Knowledge and Information Systems*, 9(3):259-283, 2006.

[Owe08] Owens, J.D. et al., GPU computing, *Proceedings of the IEEE*, 96:5 (2008), 879-899.

[Pao08] H. T. Pao, S. C. Chuang, Y. Y. Xu, H. . Fu, An EM based multiple instance learning method for image classification. *Expert Systems with Applications*, 35:3 (2008), 1468–1472.

[Pol08] R. Poli, W. B. Langdon and N. F. McPhee. *A Field Guide to Genetic Programming*. Lulu Enterprises, 2008.

[Qi09] X. Qi and B.D. Davison. Web page classification: Features and algorithms. *ACM Computing Surveys* 41:2 (2009), article 12.

[Rea08] J. Read, B. Pfahringer and G. Holmes. Multi-label Classification using Ensembles of Pruned Sets. *IEEE International Conference on Data Mining (ICDM 2008)*. Pisa, Italy, 2008.

[Reu04] P. Reutemann, B. Pfahringer and E. Frank. A Toolbox for Learning from Relational Data with Propositional and Multi-Instance Learners. *17th Australian Joint Conference on Artificial Intelligence*, 1017-1023. Springer, 2004.

[Rom04] C. Romero, S. Ventura and P. de Bra. Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware Authors. *User Modelling and User Adapted Interaction*, 14:5 (2004), 425-464.

[Rom07] C. Romero and S. Ventura. Educational Data Mining: a Survey from 1995 to 2005. *Expert Systems with Applications* 33:1 (2007), 135-146.

[Rom10a] C. Romero and S. Ventura. Educational Data Mining: A Review of the State-of-the-Art. *IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews*. 40:6 (2010), 601 – 618.

[Rom10b] C. Romero, S. Ventura, M. Pechenizkiy and R. Baker (Eds.). *Handbook of Educational Data Mining*. Taylor & Francis, 2010.

[Sal07] A. Salleb-Aouissi, C. Vrain and C. Nortet. QuantMiner: a Genetic Algorithm for Mining Quantitative Association Rules. *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'97)*, Hyberadad, India, 1035-1040, 2007.

[San08] D. Sánchez, J.M. Serrano, L. Cerda and M. A. Vila. Association Rules Applied to Credit Card Fraud Detection. *Expert systems with applications* 36 (2008), 3630-3640.

[Sco05] S. Scott, J. Zhang, J. Brown. On generalized MIL. *International Journal of Computational Intelligence and Applications* 5 (2005) 21–35.

[Sun09] Y. Sun, A. C. Wong, M. S. Kamel. Classification of Imbalanced Data: A Review. *International Journal of Pattern Recognition and Artificial Intelligence* 23:4 (2009), 687–719.

[Sup06] J. F. Superby, J. P. Vandamme and N. Meskens. Determination of factors influencing the achievement of the first-year university students using data mining methods. In *International conference on intelligent tutoring systems, Educational Data Mining Workshop*, Taiwan, 1-8. 2006.

[Tso07] G. Tsoumakas and I. Katakis. Multi label classification: An overview. *International Journal of Data Warehousing and Mining* 3:3 (2007), 1-13.

[Tso09] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining Multi-label Data. In O. Maimon and L. Rokach (Eds.). *Data Mining and Knowledge Discovery Handbook*. Springer, 2009.

[Wei03] N. Weidmann, E. Frank and B. Pfahringer, A two-level learning method for generalized multi-instance problems, in: *14th European Conference on Machine Learning (ECML'03)*, Cavtat-Dubrovnik, Croatia, 2003.

[Wei04] G. M. Weiss. Mining with Rarity: A Unifying Framework. *SIGKDD Explorations* 6:1 (2004), 7-19.

[Wil10] Wilson, G., Banzhaf, W., Deployment of parallel linear genetic programming using GPUs on PC and video game console platforms, *Genetic Programming and Evolvable Machines*, 11 (2), pp. 147-184, 2010.

[Won06] Wong, T.-T., Wong, M.L., Parallel evolutionary algorithms on consumer-level graphics processing unit, Studies in Computational Intelligence, 22, pp. 133-155, 2006.

[Xu04] X. Xu, E. Frank, Logistic regression and boosting for labeled bags of instances, *8th Conference of Pacific-Asia (PAKDD'04)*. LNCS 3056, 272-281, 2004.

[Yan05] X. Yan, C. Zhang and S. Zhang. ARMGA: Identifying Interesting Association Rules with Genetic Algorithms. *Applied Artificial Intelligence*, 19:7 (2005), 677-689.

[Yan09] X. Yan, D. Zhang and S. Zhang. Genetic Algorithm-based Strategy for Identifying Association Rules without Specifying Actual Minimum Support. *Expert Systems with Applications*, 36:2 (2009), 3066-3076.

[Yu10] X. Yu and M. Gen. *Introduction to Evolutionary Algorithms*. Springer, 2010.

[Zaf09] A. Zafra, C. Romero, S. Ventura and E. Herrera-Viedma. Multi-Instance Genetic Programming for Web Index Recommendation. *Expert Systems with Applications*, 36 (2009), 11470-11479.

[Zaf10] A. Zafra and S. Ventura. G3P-MI: A Genetic Programming Algorithm for Multiple Instance Learning. *Information Sciences*, 180:23 (2010), 4496-4513.

[Zaf11a] A. Zafra, E. Gibaja and S. Ventura. Multi-instance Learning with Multi-Objective Genetic Programming for Web Mining. *Applied Soft Computing* 11:1 (2011), 93-102.

[Zaf11b] A. Zafra, M. Pechenizkiy and S. Ventura. HyDR-MI: A Hybrid Algorithm to Reduce Dimensionality in Multiple Instance Learning. *Information Sciences* 2011 (accepted).

[Zha05] M. L. Zhang and Z. H. Zhou. A k-nearest neighbor based algorithm for multi-label classification. *IEEE international Conference on Granular Computing*, vol. 2, pages 718-721 Vol. 2. The IEEE Computational intelligence Society, 2005.

[Zha06a] M. L. Zhang and Z. H. Zhou. Adapting RBF Neural Networks to multi-instance learning. *Neural Processing Letters* 23 (1) (2006) 1–26.

[Zha06b] M. L. Zhang and Z. H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338_1351, 2006.

[Zha08a] Z.-J. Zha, X.-S.Hua, T. Mei, J. Wang, G.-J. Qi and Z. Wang. Joint multi-label multi-instance learning for image classification. *IEEE Conference on Computer Vision and Pattern Recognition*, pp 1-8, 2008.

[Zha08b] Q. Zhang and R. S. Segall. Web Mining: A Survey of Current Research, Techniques and Software. *International Journal of Information Technology & Decision Making* 7:4 (2008) 683–720.

[Zho05] Z.-H. Zhou, K. Jiang and M. Li. Multi-instance learning based web mining. *Applied Intelligence* 22:2 (2005) 135–147.

[Zho07a] Z.-H. Zhou and M.-L. Zhang. Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowledge and Information Systems 11:*2 (2007) 155–170.

[Zho07b] Z.-H. Zhou, M.-L. Zhang. Multi-instance multi-label learning with applications to scene classification. In: Advances of Neural Information Processing Systems 20 (NIPS'06), Vancouver, Canada, 2007, 1609-1616.

3. OBJETIVES
(maximum 2 pages)

3.1 Describe the reasons to present this proposal and the **initial hypothesis** which support its objectives (maximum 20 lines)

The aim of designing methods of knowledge extraction is to obtain useful models, not only in terms of high performance (understood as accuracy in classification problems), but also in terms of other relevant characteristics such as robustness, versatility, interpretation, ease of updates, and coherence with previous knowledge. Furthermore, such systems should be able to model and manage all the information within their reach, including incomplete and imprecise information, imbalanced data in classification problems, etc.

Our departure hypothesis is that the use of Evolutionary Computing techniques, and particularly those based on Genetic Programming, should allow us to design knowledge extraction models that fulfill the aforementioned features. GP allows a much more flexible individual representation than other EC paradigms, allowing representing decision trees, rule bases or complex mathematical expression directly. It can also express restrictions in the representation space, increasing the efficiency of the evolutionary process, and give more control over the size of the resulting expressions, improving the interpretability of results. Finally, the development of methods based on the use of massive parallel computers opens possibilities for facing large scale data mining problems (high dimensional problems and large scale datasets), by significantly speeding up the processing of data mining algorithms.

Although significant progress has been made in the last few years, there is still much work to be done in this line. Some of them will be approached in the four objectives proposed in the current document. We firmly believe that the team can achieve these objectives, given the experience and results obtained previously. Please see sections 3.3 and 4, where the objectives and tasks are described.

3.2. Indicate the **background and previous results** of your group or the results of other groups that support the initial hypothesis

As has already been mentioned in the introduction, our group has had experience in some of the topics covered in this proposal. We proceed now to comment in more detail on the most significant contributions to date. The list of associated published journal articles can be found in section 6 of this proposal (journal articles) and the rest of the associated publications can be found on the web page of the research group (http://www.uco.es/grupos/kdis).

In the field of developing learning models based on GP, the group has proposed various models to obtain classification rules and discriminating functions. At the same time it has recently published an article which deals with a revision of all proposals made to date (until January 2010). Furthermore, it has also obtained preliminary results in the field of implementing genetic programming algorithms in GPU, and sent various communications to international congresses.

In the field of multi-instance learning, the group has developed a model to obtain rules for learning MIL, whose results are better than those found to date in most of the techniques published. The model, called G3P-MI, has been extended by using multi-objective considerations, to get a model (MO G3P-MI) that maintains a trade-off between the sensitivity and specificity of the results, although it is similar to G3P-MI in accuracy. Previous models were applied to the classification problem of web index pages (a problem within the framework of web content-mining and web usage mining), obtaining excellent results. This research resulted in the thesis of Dr. Amelia Zafra and to date has also included publications in such journals as *Information Sciences*, *Applied Soft Computing* and *Expert Systems with Applications*.

As for multi-label classification, the group has developed a learning model for discriminating functions for MLC (GEP-MLC), especially indicated in problems with numerical variables, but that has also been applied successfully to data sets with nominal variables. The model has been found to be competitive, better than many algorithms in state of the art lines (BR, RAKEL and others). This result has been published in various communications to conferences, as well as in an article in the *Journal of Multi-Valued Logic and Soft Computing*. There are also some preliminary results on the new model for the learning of classification rules for MLC, more easily interpreted than the previous one and more appropriate for data with nominal variables. The results indicate that this model improves on GEP-MLC even in problems with numeric variables.

With respect to the learning of association rules, the group began its research in 2002 in the thesis of Dr. Cristóbal Romero, applying algorithms to obtain association rules to improve e-learning courses, based on genetic programming. In the last few years different models have been developed to obtain association rules based on grammatical genetic programming (G3P-ARM) whose main characteristic is that they can deal with numerical data and its high scalability and performance, as compared to other proposals based as much on genetic as on traditional algorithms. With respect to the topic of the extraction of infrequent association rules, the group has obtained promising preliminary results by adapting the G3P-ARM algorithm so that the rules discovered are infrequent instead of frequent.

As mentioned above, the members of the group obtained their first results in 2002 in the field of educational data mining. Similarly, the group has achieved results in the field of LMS personalization, using rules extracted from information generated in the environment through data mining techniques. Work in the field of hyperlink personalization in adaptive systems has also been successful, as well as the use of techniques to discover subgroups in order to obtain rules that could contribute to improving the LMS (an approach different to that of the research done in the doctoral thesis of Dr. Romero). The results in this line have inspired several journal publications. Furthermore two introductory articles to this theme have been published, one of which receiving an award of the journal *Expert Systems with Applications* for being one of the most quoted in the period 2006-2008.

In the web mining field, the group applied the G3P-MI and MO G3P-MI algorithms to the problem of categorizing web index pages, and achieved excellent results. As well, preliminary results have been obtained about obtaining models for intruder detection based on techniques that obtain infrequent association rules.

3.3. Describe briefly the objectives of the project.

1. Development of genetic programming models for solving several knowledge discovery problems: multiple instance learning, relational learning, multi-label classification and learning association rules:
   a. Development of models for multiple instance learning.
   b. Development of models for multi-label classification.
   c. Development of models for multi-instance multi-label classification.
   d. Development of models for learning classical association rules.
   e. Development of models for learning relational association rules.
2. Adaptation of models previously developed for problems with different kinds of datasets: imbalanced, high dimensionality and large scale datasets.
   a. Development of models for learning with imbalanced data in multiple instance learning and multi-label classification.
   b. Development of models for learning infrequent association rules.
   c. Development of models based on GPGPU for large scale knowledge discovery problems.
   d. Development of feature selection methods for multi-instance and multi-label classification problems.
   e. Development of instance selection methods for multiple instance learning and multi-label classification problems.
3. Application of models previously developed for educational data mining and web mining problems:
   a. Relational and MI representations for predicting student's performance in LMS.
   b. Predicting student drop-out.
   c. Categorization of learning objects.
   d. Providing feedback to support instructors and students.
   e. Web page categorization. Development of web spam models.
   f. Intrusion detection.
4. Development of data repositories oriented to facilitate the comparison of our algorithms with others previously published and integrate our knowledge discovery models in the most prominent KDD software platforms (KEEL and WEKA) in order to make them available to the scientific community:
   a. Development of repository datasets to compare results in MIL, MLC and RL.
   b. Development of a Simulator Tool for Generalized Multiple Instance Problems.
   c. Integration of algorithms in KEEL.
   d. Integration of algorithms in WEKA.

3.4. For Coordinated projects only, the coordinator must indicate (maximum 2 pages):

- the global objectives of the coordinated project, the need for coordination, and the added value provided by this coordination
- the specific objectives of each subproject
- the interaction among the objectives, activities and subprojects
- the mechanisms of coordination for an effective execution of the project.

4. METHODOLOGY AND WORKING PLAN
(in the case of coordinated projects this title must include all the subprojects)

METHODOLOGY

The proposed methodology has a theoretical and a practical part. Concerning the theoretical one, we will develop new algorithms for knowledge extraction and to analyze the behavior of learning algorithms. Concerning the practical part, we will integrate our algorithms into software environments (KEEL and WEKA). The method of study for the theoretical part is commonly used by the scientific community:

- Setting the hypotheses: for the current project, this implies the development of new algorithms for knowledge extraction, the improvement and adaptation of the available algorithms for specific problems, and the analysis of mechanisms to compare learning algorithms, together with the development of new methodologies;
- Collecting data, which in our context means to have public and real databases;
- Testing between hypotheses and observations, i.e., evaluation of the quality of evolutionary algorithms for knowledge extraction with respect to the databases used; and finally,
- Re-adaptation of the initial hypothesis taking into account the results obtained; this will imply the modification and tuning of the evolutionary algorithms and the mechanisms to analyze their behavior as a result of the test carried out and the experience accumulated.

The completion of this project will be undertaken following the guidelines of the Action Research methodological approach, which aims at solving problems through their identification, following a collaborative philosophy of analysis of the problem in order to understand its cause by combining theory and practice. Initially designed for medical and social problems, and often used for the construction of experimental information systems, this method is of particular interest when developing proposals that will actually be used and tested by real developers and end users, as is the case at hand, and using data extracted from real contexts as well. Note that JCLEC[4], a library developed by our research group, provides an excellent source of developers and end-users, respectively, who contribute significant feedback and criticism. So their comments and findings, as well as our research work and experimentation, will guide us through the process steps to adapt and evolve the work. The Action Research strategy, which is widely used by Spanish research groups because of its suitability and simplicity, prescribes an iterative approach of continuous improvement. Thus, in each cycle we have the following steps:

1. Diagnosis: identifying the problem
2. Action planning: considering different alternatives and choosing the most appropriate
3. Taking action: designing and implementing the solution and prototype
4. Evaluating: studying the solution by through exhaustive testing using the prototype
5. Specifying learning: analyzing and assessing the results obtained to identify the changes required to improve the solution

The steps prescribed by this methodology have been adopted in the tasks described below.

WORKING PLAN

The project is structured in class A tasks (improvement and development of GP algorithms for knowledge extraction), class B tasks (models for imbalanced and high dimensionality data), class C tasks (application areas), and class D tasks (development of data repositories and algorithm integration in WEKA and KEEL). Every task is assigned to a coordinator and several participants. In the following, the tasks and the associated timing are detailed.

A – Genetic Programming Models for Knowledge Extraction

This block comprises the development of GP models for solving the problems presented in the introduction of this proposal. These tasks are the following:

T.A.1 – Development of models for MIL

Coordinator: A. Zafra

Participants: A. Cano, E. Gibaja, M. Pechenizkiy, S. Ventura, A. Zafra

Description: Development of new GP models for multiple instance learning, focusing on the following topics: co-evolutionary models, and the use of ensemble models to improve accuracy. Adaptation of the evolutionary process to this learning paradigm.

[4] JCLEC is a Java class library for evolutionary computation. Among its main features, this library contains a module for knowledge discovery with evolutionary algorithms, with some of the algorithms developed by our research group. More information can be found in http://jclec.sourceforge.net

Timing: M01-M02 Study and analysis of the latest methods applied to MIL.
M03-M06 Design and implementation of new algorithms.
M07-M09 Integration, test and evaluation of the algorithms developed.

Achieved Objective: 1.a
Expected results: New GP models for multiple instance learning to improve the performance and to achieve a high interpretability of the results.

T.A.2 - Development of models for MLC

Coordinator: E. Gibaja

Participants: J.L. Ávila, E. Gibaja, J.M. Luna, S. Ventura, A. Zafra

Description: Performance improvement of our MLC methods and development of new ones.

Timing: M01-M02 Study of classical and recent proposals in the area.
M03-M06 Study and development of new proposals
M07-M09 Evaluation of new proposals and their results with respect to others

Achieved Objective: 1.b
Expected results: New models for multi-label classification

T.A.3 - Development of models for MI-MLC

Coordinator: S. Ventura

Participants: J.L. Ávila, A. Cano, E. Gibaja, S. Ventura, A. Zafra,

Description: Development of new proposals for multi-instance multi-label classification based on (a) category-wise decomposition and (b) transformation representation methods.

Timing: M10-M12 Study of previous proposals in the area.
M13-M18 Analysis and design of methods.
M19-M21 Evaluation and study of new proposals.

Achieved Objective: 1.c
Expected results: New models for multi-label classification from multiple instance data.

T.A.4 - Development of models for learning classical association rules

Coordinator: J. R. Romero

Participants: J.M. Luna, J.L. Olmo, M. Pechenizkiy, C. Romero, J. R. Romero

Description: Development of new evolutionary models for learning association rules from frequent item sets applied to different datasets in terms of their form (i.e. comprising categorical, nominal, range data, etc.) and size. More specifically, we focus on different datasets in models guided by context-free grammars (CFG) implementing G3P.

Timing: M01-M02 Study of the state-of-the-art in current approaches, including both classical and evolutionary proposals.
M03-M06 Design and development of a G3P-based model for mining association rules.
M07-M09 Assessment of the proposal and consideration of alternatives.

Achieved Objective: 1.d
Expected results: New G3P-based models for association rule mining.

T.A.5 - Development of models for learning relational association rules

Coordinator: J. R. Romero

Participants: J.M. Luna, J.L. Olmo, M. Pechenizkiy, J. R. Romero, S. Ventura

Description: Development of new models for learning complex association rules, which express relationships between elements that are blurry in multiple tables in a relational database.

Timing: M07-M08 Study of the current proposals in this field.
M09-M13 Based on the highlights and results found in T.A.4, development of models for association rule mining on relational systems.
M14-M16 Application and analysis of the new grammar-guided models, and proposal for alternative GP-based models.
M17-M18 Validation of models and comparison with existing approaches.

Achieved Objective: 1.e
Expected results: New models for mining relational association rules.

B – Genetic Programming models for learning with imbalanced and large scale data

Type B tasks are focused on the adaptation of algorithms developed in the previous task A to problems with imbalanced data sets and large scale data (high number of variables and/or examples) problems

T.B.1 - Development of models for learning with imbalanced data in MIL, MLC and MI-MLC

Coordinator: S. Ventura

Participants: J.L. Ávila, E. Gibaja, M. Pechenizkiy, S. Ventura, A. Zafra

Description: Development of a new proposal for learning with imbalanced data in MI, ML and MI-ML problems. Adaptation of the existing GP algorithms and design of new models where the evolutionary process could adapt to this type of problems, use of specific metrics in the fitness function, multiple objectives measuring the performance in each of the classes and for multi class problems.

Timing:	M01-M02	Study of previous proposals in classification with imbalanced data (classical problems).
	M03-M12	Analysis and design of new proposals.
	M13-M27	Evaluation of the proposals.

Achieved Objective: 2.a
Expected results: New models for learning with imbalanced data in MI, ML and MI-ML problems.

T.B.2 - Development of models for learning infrequent association rules

Coordinator: J.R. Romero

Participants: J.M. Luna, J.L. Olmo, M. Pechenizkiy, C. Romero, J. R. Romero, S. Ventura

Description: Development of new models for infrequent (or rare) mining association rules We are primarily interested in the application of GP as a mechanism to optimize the performance achieved by current algorithms, mainly in terms of runtime and accuracy; furthermore, the use of G3P-based methods allows the generation process of rules to be restricted as well as their application domain.

Timing:	M10-M12	Study of previous algorithms.
	M13-M17	Analysis, design and implementation of GP-based algorithms for mining infrequent rules.
	M18-M21	Analysis, design and implementation of a G3P-based proposal.
	M22-M24	Assessment of the proposals and comparison of results with existing approaches.

Achieved Objective: 2.b
Expected results: New evolutionary models for rare association rule mining.

T.B.3 - Development of models based on GPGPU for large scale problems

Coordinator: S. Ventura

Participants: J.L. Ávila, A. Cano, J.M. Luna, S. Ventura, A. Zafra

Description: Design and implementation of high performance algorithms based on GPGPU.

Timing:	M01-M02	Study of the GP algorithms to implement.
	M03-M12	Analysis, design and implementation.
	M13-M27	Testing and evaluation of the algorithms.

Achieved Objective: 2.c
Expected results: New versions of GP models for large scale problems.

T.B.4 – Feature Selection Methods in MIL, MLC and MI-MLC

Coordinator: A. Zafra

Participants: J.L. Ávila, A. Cano, E. Gibaja, M. Pechenizkiy, S. Ventura, A. Zafra

Description: Development of new methods for feature selection in multiple instance and multi-label learning problems. Study about the use of classical feature selection algorithms in MIL and MLC problems and the development of new proposals for this task.

Timing:	M13-M14	Study of classical feature selection algorithms.
	M15-M23	Analysis, design and implementation of new methods.

M24-M27 Evaluation and improvement of the algorithms developed.

Achieved Objective: 2.d
Expected results: New models for feature selection in multi-instance and multi-label problems.

T.B.5 – Instance selection methods in MIL, MLC and MI-MLC

Coordinator: E. Gibaja

Participants: J.L. Ávila, A. Cano, E. Gibaja, M. Pechenizkiy, A. Zafra

Description: Development of instance selection methods specifically designed to deal with MIL, MLC and MI-MLC problems.

Timing: M13-M16 Study of instance selection proposals in single-label, MLC, MIL and MI-MLC.
M17-M23 Analysis, design and implementation of our proposals.
M24-M27 Testing and evaluation of our proposals.

Achieved Objective: 2e
Expected results: New instance selection models for MLC, MIL and MI-MLC

C – Applying the models developed to problems in the scope of educational data mining and wed data mining.

In this group of tasks, we will apply the developed algorithms in the previous task A and B, and we will also do a survey about the state of arts in the problems described in the introduction of this project.

T.C.1 – Relational and MI representations for predicting student's performance in Learning Management Systems (LMS)

Coordinator: C. Romero

Participants: A. Cano, M. Pechenizkiy, C. Romero, J. R. Romero, A. Zafra

Description: To develop a more flexible form of data representation using relational and MI in order to improve the prediction of student's performance/marks/scores in LMS.

Timing: M01-M06 Study of previous proposals on predicting student performance (classical algorithms).
M06-M12 Analysis and design of new proposals.
M12-M18 Evaluation of the proposals

Achieved Objective: 3.a
Expected results: New models of classification for predicting student performance.

T.C.2– Predicting student drop-out

Coordinator: C. Romero

Participants: J.M. Luna, J.L. Olmo, M. Pechenizkiy, C. Romero, J. R. Romero

Description: To apply algorithms taking into account imbalanced data problems in order to predict student drop outs.

Timing: M1-M04 Study of previous proposals in prediction of student drop outs (classical algorithms).
M04-M10 Analysis and design of new proposals.
M10-M12 Evaluation of the proposals

Achieved Objective: 3.b
Expected result: New models for predicting student drop outs.

T.C.3 – Categorization of learning objects

Coordinator: C. Romero

Participants: J.L. Ávila,E. Gibaja,J.M. Luna, C. Romero, J. R. Romero

Description: To automatically classify Learning Objects (LOs) using ML in one or several subjects, fields or related knowledge domain areas.

Timing: M25-M28 Study of previous proposals about classifying learning objects (classical algorithms).
M29-M34 Analysis and design of new proposals.
M35-M36 Evaluation of the proposals

Achieved Objective: 3.c
Expected results: New models for classifying learning objects.

T.C.4 – Providing feedback for supporting instructors and students

Coordinator: C. Romero

Participants: J.L. Ávila, J.L. Olmo, M. Pechenizkiy, C. Romero, J. R. Romero

Description: To apply relational data mining and rare association rules to provide feedback for both instructors and students.

Timing:	M19-M21	Study of previous proposals about association rule mining and providing feedback.
	M22-M28	Analysis and design of new proposals.
	M29-M30	Evaluation of the proposals.

Achieved Objective: 3.d
Expected results: New models for making recommendations for instructors and students.

T.C.5 – Web Page Categorization

Coordinator: A. Zafra

Participants: A. Cano, E. Gibaja, J.M. Luna, J.L. Olmo, A. Zafra

Description: Development of evolutionary learning algorithms for web page classification tasks. Study of alternative representations of web pages and its suitability to the web page categorization tasks. Application to the development of user profiles in recommendation systems and to web spam problems.

Timing:	M19-M21	Study of different representations of the problem.
	M22-M27	Analysis and design of new methods.
	M28-M30	Evaluation and improvement of the algorithms developed.

Achieved Objective: 3.d
Expected results: New evolutionary models for categorization of web pages and study of more flexible representations to tackle web page categorization problems.

T.C.6 – Intrusion detection

Coordinator: J. R. Romero

Participants: J.M. Luna, J.L. Olmo, J. R. Romero, S. Ventura, A. Zafra

Description: Development of new models for intrusion detection in user-based collaborative recommender systems. This will be addressed from two different perspectives: firstly, using associative classification over imbalanced data for detection; secondly using rare association rule mining to detect infrequent rating patterns. This will then be automated in an overall integrated, tool-supported process.

Timing:	M19-M20	Study of the current state-of-the-art.
	M21-M24	Analysis, design and implementation of associative classification algorithms over imbalanced data, and their application in the intrusion-detection problem.
	M25-M27	Adaptation of previous rare association rule mining algorithms, extracted from T.B.2, to this field.
	M28-M31	Automation of the detection process for user-based collaborative recommender systems, and design and development of integrated tools to support it.
	M32-M33	Assessment and validation of the proposal.

Achieved Objective: 3.f
Expected result: New GP-based and G3P-based algorithms for associative classification; adapted rare association rule mining algorithm; integrated tool to automate the detection process using the previous proposals.

D – Development of data repositories. Integration of models in KEEL and WEKA

This block of tasks will develop data repositories in order to make it easier to compare the performance of the proposals for each topic; during tasks A and B, the proposals will be integrated in the KEEL and WEKA learning environment.

T.D.1 – Repository datasets for comparison of results in MIL, MLC and RL

Coordinator: E. Gibaja

Participants: A. Cano, J.L. Ávila, E. Gibaja, C. Romero, J. R. Romero, A. Zafra

Description: Development of a repository of datasets regarding to MI, RL and MLC.

Timing:	M22-M24	Repository datasets for MLC and MIL.

M31-M33 Repository datasets for MI-MLC and RL.
M33-M36 Integration and deployment of the datasets.

Achieved Objective: 4.a
Expected result: Datasets about MIL, MLC and RL

T.D.2 – Simulator for Generalized Multiple Instance Problems

Coordinator: A. Zafra

Participants: J.L. Olmo, C. Romero, S. Ventura, A. Zafra

Description: Design of artificial datasets for multiple instance learning considering generalized multiple instance learning. This task will comprise the use of different hypotheses which determine the relationship between the instances in a bag. The simulator tool has to allow the specification of how a specific instance interacts with the rest of the instances that belong to the same bag.

Timing: M01-M03 Generation of artificial dataset for generalized multiple instance learning.
M04-M06 Design and evaluation of algorithms applied on the new data sets.

Achieved Objective: 4.b
Expected result: New artificial generalized multiple instance problems. Design of different hypotheses to work in multiple instance learning to allow generalized relationships between the instances in one pattern or bag to be represented.

T.D.3 – Integration of algorithms in KEEL

Coordinator: S. Ventura

Participants: J.L. Ávila, A. Cano, J.M. Luna, J.L. Olmo, S. Ventura

Description: Integration of KDD algorithms in the KEEL software.

Timing: M25-M27 Study of the algorithms to be included in KEEL.
M27-M30 Analysis, design and implementation of the algorithms.
M31-M36 Integration, test and evaluation of the algorithms developed.

Achieved Objective: 4.c
Expected result: New version of the MIL and MLC modules of the KEEL software, that includes new algorithms developed in class A and B tasks.

T.D.4 – Integration of algorithms in WEKA

Coordinator: E. Gibaja

Participants: J.L. Ávila, A. Cano, E. Gibaja, J.M. Luna, J.L. Olmo, S. Ventura

Description: Integration of GP based KDD into the WEKA software.

Timing: M10-M12 Integration of the algorithms developed in T.A.1, T.A.2 and T.A.4
M22-M24 Integration of the algorithms developed in T.A.3, T.A.5
M31-M36 Integration of the algorithms developed in T.B.1, T.B.2, T.B.3, T.B.4 and T.B.5

Achieved Objective: 4.d
Expected result: New WEKA components for MIL, ARM and other tasks based on the evolutionary algorithms developed.

4.1 CHRONOGRAM MODEL (EXAMPLE)

This chronogram must indicate the persons involved in the project, including those contracted with project funds.
Underline the name of the person responsible of each task.

Tasks	Centre	Persons	First Year (*)	Second Year (*)	Third Year (*)
T.A.1. Development of models for MIL	UCO	A. Zafra A. Cano, E. Gibaja, M. Pechenizkiy, S. Ventura	☒☒☒☒☒☒☒☒☒□□□	□□□□□□□□□□□□	□□□□□□□□□□□□
T.A.2. Development of models for MLC	UCO	E. Gibaja J.L. Ávila, J.M. Luna, S. Ventura, A. Zafra	☒☒☒☒☒☒☒☒☒□□□	□□□□□□□□□□□□	□□□□□□□□□□□□
T.A.3. Development of models for MI-MLC	UCO	S. Ventura J.L. Ávila, A. Cano, E. Gibaja, A. Zafra	□□□□□□□□□☒☒☒	☒☒☒☒☒☒☒☒☒□□□	□□□□□□□□□□□□
T.A.4. Development of models for learning classical association rules	UCO	J. R. Romero J.M. Luna, J. L. Olmo, M. Pechenizkiy, C. Romero	☒☒☒☒☒☒☒☒☒□□□	□□□□□□□□□□□□	□□□□□□□□□□□□
T.A.5. Development of models for learning relational association rules	UCO	J. R. Romero J.M. Luna, J. L. Olmo, C. Romero, S. Ventura	□□□□□□☒☒☒☒☒☒	☒☒☒☒☒☒□□□□□□	□□□□□□□□□□□□
T.B.1. Development of models for learning with imbalanced data in MIL, MLC and MI-MLC	UCO	S. Ventura J.L. Ávila, E. Gibaja, M. Pechenizkiy, A. Zafra	□□□□□□□□□☒☒☒	☒☒☒☒☒☒☒☒☒☒☒☒	□□□□□□□□□□□□
T.B.2. Develpment of models for learning infrequent association rules	UCO	J. R. Romero J.M. Luna, J. L. Olmo, C. Romero, M. Pechenizkiy, S. Ventura	□□□□□□□□□☒☒☒	☒☒☒☒☒☒☒☒☒☒☒☒	□□□□□□□□□□□□
T.B.3. Development of models based on GPGPU for large scale problems	UCO	S. Ventura J.L. Ávila, A. Cano, J.M. Luna, A. Zafra	☒☒☒☒☒☒☒☒☒☒☒☒	☒☒☒☒☒☒☒☒☒☒☒☒	☒☒☒□□□□□□□□□
T.B.4. Feature Selection Methods in MIL, MLC and MI-MLC	UCO	A. Zafra J.L. Ávila, A. Cano, E. Gibaja, M. Pechenizki, S. Ventura	□□□□□□□□□□□□	☒☒☒☒☒☒☒☒☒☒☒☒	☒☒☒□□□□□□□□□

(*) Mark an X inside the corresponding boxes (months)

<table>
<tr><th>Tasks</th><th>Centre</th><th>Persons</th><th>First Year (*)</th><th>Second Year (*)</th><th>Third Year (*)</th></tr>
<tr><td>T.B.5. Instance Selection Methods in MIL, MLC and MI-MLC</td><td>UCO</td><td>E. Gibaja
J.L. Ávila, A. Cano, E. Gibaja, M. Pechenizki, S. Ventura</td><td> | | | | | | | | | | | </td><td>X|X|X|X|X|X|X|X|X|X|X|X</td><td>X|X|X| | | | | | | | | </td></tr>
<tr><td>T.C.1. Multi-instance and relational representations for predicting students’ performance</td><td>UCO</td><td>C. Romero
A. Cano, M. Pechenizkiy, J. R. Romero, A. Zafra</td><td>X|X|X|X|X|X|X|X|X|X|X|X</td><td>X|X|X|X|X|X| | | | | | </td><td> | | | | | | | | | | | </td></tr>
<tr><td>T.C.2. Predicting drop-out students</td><td>UCO</td><td>C. Romero
J.M. Luna, J. L. Olmo, C. Romero, S. Ventura</td><td>X|X|X|X|X|X|X|X|X|X|X|X</td><td> | | | | | | | | | | | </td><td> | | | | | | | | | | | </td></tr>
<tr><td>T.C.3. Learning Objects Categorization</td><td>UCO</td><td>C. Romero
J.L. Ávila, J.L. Olmo, M. Pechenizkiy, J.R. Romero</td><td> | | | | | | | | | | | </td><td> | | | | | | | | | | | </td><td>X|X|X|X|X|X|X|X|X|X|X|X</td></tr>
<tr><td>T.C.4. Learning Personalization Rules in LMS</td><td>UCO</td><td>C. Romero
J.L. Ávila, J. L. Olmo, M. Pechenizkiy, J.R. Romero</td><td> | | | | | | | | | | | </td><td> | | | | | |X|X|X|X|X|X</td><td>X|X|X|X|X|X| | | | | | </td></tr>
<tr><td>T.C.5. Web Page Categorization</td><td>UCO</td><td>A. Zafra
A. Cano, E. Gibaja, J.M. Luna, J.L. Olmo</td><td> | | | | | | | | | | | </td><td> | | | | | |X|X|X|X|X|X</td><td>X|X|X|X|X|X| | | | | | </td></tr>
<tr><td>T.C.6. Intrusion Detection</td><td>UCO</td><td>J.R. Romero
J.M. Luna, J.L. Olmo, S. Ventura</td><td> | | | | | | | | | | | </td><td> | | | | | |X|X|X|X|X|X</td><td>X|X|X|X|X|X| | | | | | </td></tr>
<tr><td>T.D.1. Development of MIL, MLC and MI-MLC data repositories</td><td>UCO</td><td>E. Gibaja
J.L. Ávila, A. Cano, C. Romero, J.R. Romero</td><td> | | | | | | | | | | | </td><td> | | | | | | | | |X|X|X</td><td> | | | | | |X|X|X|X|X|X</td></tr>
</table>

<table>
<tr><th>Tasks</th><th>Centre</th><th>Persons</th><th>First Year (*)</th><th>Second Year (*)</th><th>Third Year (*)</th></tr>
<tr><td>T.D.2. Datasets simulator for MIL</td><td>UCO</td><td>A. Zafra
J.L. Olmo, C. Romero, S. Ventura</td><td>X|X|X|X|X|X| | | | | |</td><td> | | | | | | | | | | | </td><td> | | | | | | | | | | | </td></tr>
<tr><td>T.D.3. Integration of algorithms in KEEL</td><td>UCO</td><td>S. Ventura
J.L. Ávila, A. Cano, J.M. Luna, J.L. Olmo</td><td> | | | | | | | | | | | </td><td> | | | | | | | | | | | </td><td>X|X|X|X|X|X|X|X|X|X|X|X</td></tr>
<tr><td>T.D.4. Integration of algorithms in WEKA</td><td>UCO</td><td>E. Gibaja
J.L. Ávila, A. Cano, J.M. Luna, J.L. Olmo, S. Ventura</td><td> | | | | | | | | |X|X|X</td><td> | | | | | | | | |X|X|X</td><td> | | | | | |X|X|X|X|X|X</td></tr>
</table>

5. BENEFITS DERIVED FROM THE PROJECT, DIFFUSION AND EXPLOITATION OF RESULTS (maximum 1 page)

PROJECT BENEFITS

The most relevant scientific benefit derived from this project is the development of models to resolve the new challenges in the knowledge discovery environment. In fact, as has been mentioned previously, there is still a great amount of work to be done in these research lines, and we hope that our contribution can be of real help to advance progress in this field.

On the other hand, we feel that the application of our models for solving real problems is very worthwhile. The problems proposed in this project have a real intrinsic interest. For example, the model of the web spam is especially attractive for those companies that run web search engines, whose income depends partially on the reliability of their recommendations. In addition, the intruder detection topic is of great commercial interest, as can be seen in the huge investments that companies have made in information technology security. Finally, the prediction of potential student drop-outs and student performance are problems of great social interest. In fact, in all the descriptive handbooks of the four year university degree that has recently been implanted include a section with previsions about graduation, drop-out and expected efficiency rates, as well as a general procedure for evaluating the progress and results of student learning. We hope our contributions can help provide solutions for these problems, and if deemed appropriate, allow us to contact with different social agents to transfer the knowledge that can be of interest, through the development of software applications, assessment, etc.

Regarding the expected scientific-technical contributions, the project should produce at least 6-9 papers for publication in high level international scientific journals. According to the curriculum of project members, we can observe that they have published papers in important journals in fields related to the project.

ADAPTATION TO THE PRIORITIES OF THE CALL

In our opinion, the proposed project totally dovetails with the objectives and priorities of the National Program of Research I+D+I 2008-2011 in the following lines: 8.4. Acción estratégica de Telecomunicaciones y Sociedad de la Información (National Strategy on Telecommunications and Information Society). 01 Tecnologías Informáticas (Information Technologies), section 1.3. Sistemas Inteligentes (Intelligent Systems).

DIFFUSION PLAN

The diffusion of the estimated results will be widespread and achieved in different ways:

- By means of the PhD dissertations performed within the project framework by members of the project. We expect to obtain around 2-3 PhD thesis in the context of this project.
- By attendance at *national and international conferences* related to the project fields. We have two objectives: on the one hand, to show the results obtained and on the other, to contact with other research groups with similar concerns. We will highlight the importance of the Spanish conferences on Evolutionary Algorithms (MAEB) and Data Mining (TAMIDA), as well as the international conferences on Evolutionary Computation, Machine learning and Data Mining.
- *By Publications in journals* related to our topics of interest. Some worthy of note are: Evolutionary Computation, IEEE Trans. on Evolutionary Computation, IEEE Trans. on Systems, Man and Cybernetics (Parts A, B and C), IEEE Trans. on Pattern Analysis and Machine Intelligence, Data Mining and Knowledge Discovery, Int. Journal of Intelligent Systems, Knowledge and Information Systems, Machine Learning, Journal of Machine Learning Research; Pattern Recognition, Pattern Recognition Letters, Data and Knowledge Engineering, Computers and Education or Expert Systems with Applications.
- In Book chapters on international editorials, as we have been publishing during the last few years on the invitation of prestigious authors.
- We have scheduled stays in other research centers and hosted researchers from other universities.

6. BACKGROUND OF THE GROUP
(In the case of a coordinated project the topics 6. and 6.1. must be filled by each partner)
(maximum 2 pages)

The project members are in the research group "Knowledge Discovery and Intelligent Systems" (KDIS – http://www.uco.es/grupos/kdis), whose reference is TIC-222 in the Andalusian Research Plan. This research group started its activity in 2009, working in the field of knowledge discovery with evolutionary algorithms (mainly genetic programming) and its applications, although most of their members started their careers previously in other research groups.

In the last few years, the group has collaborated with Drs. Paul de Bra, Toon Calders and M. Pechenizkiy (Eindhoven University of Technology, Holland). Prof. De Bra is a well-known expert in the field of e-learning and Dr. Toon Calders is an expert in the field of association rule mining. They have worked with us on topics related to the application of knowledge extraction for the improvement of adaptive learning environments. Drs. Pechenizkiy is an expert in the field of data mining and also collaborates with the group in the field of multiple instance learning.

Recently, the group has also collaborated with Dr. Kryzstoff Cios (Virginia Commonwealth University, USA) in the field of multiple instance learning. Also, Drs. Ventura and Zafra organized the workshop "Computational Intelligence in Data Mining and Knowledge Discovery" (held at ISDA 2010) together with Professor Cios.

The team on this project includes researchers with a high level of experience in the field of evolutionary learning, data mining and their applications to educational problems. It is accredited by the following items:

1. *Their publications, which are cited in the curricula of the members of the research team.*
2. *Dr. Ventura has co-edited 2 books in the field of educational data mining (co-edited, among others, with C. Romero and M. Pechenizkiy, members of the project team).*

   - C. Romero and S. Ventura (eds.), Data Mining in e-learning. WIT Press, 2006. Wessex (UK), 2006. ISBN: 1-84564-152-3.
   - C. Romero, S. Ventura, M. Pechenizkiy and R. Baker. *Handbook of Educational Data Mining*. CRC Press, 2010. ISBN:

3. *Drs. Ventura and Romero were guest editors of 1 special issue about Educational Data Mining*

   - C. Romero and S. Ventura (guest editors), Data Mining for Personalized Educational Systems. *User Modeling and User Adapted Interaction Journal,* 2011 (in press).

4. *Doctoral theses directed: two within the period 2006-2010*

   - A. Zafra, Grammar-guided genetic programming models for multiple instance learning. University of Granada (Spain). October, 2009. Advisor: Sebastián Ventura.
   - E. García, Using Data Mining Techniques for the Improvement of E-Learning Courses. University of Cordoba (Spain). December, 2010. Advisor: Cristóbal Romero.

5. *Drs. Ventura and Romero have participated in the organization of international scientific reunions:*

   - Eleventh International Conference on Intelligent Systems Design and Applications (ISDA 2011), Córdoba, 2011. (S. Ventura, general co-chair, C. Romero, program co-chair).
   - Fourth International Conference on Educational Data Mining (EDM 2011), Eindhoven, 2011. (S. Ventura, program co-chair).
   - First International Workshop on "Computational Intelligence in Knowledge Discovery" in Tenth International Conference on Intelligent Systems Design and Applications (ISDA 2010). El Cairo, 2010. (S. Ventura and A. Zafra, co-chairs).
   - Special session on "Knowledge Discovery with Evolutionary Learning" in Fifth International Conference on Hybrid Artificial Intelligent Systems (HAIS 2010). S. Sebastián, 2010. (S. Ventura, E. Gibaja and A. Zafra, co-chairs).
   - Second International Conference on Educational Data Mining, Córdoba, 2009. (S. Ventura and C. Romero, general co-chairs).
   - First International Workshop on Applying Data Mining in e-Learning (ADML'07) as part of the 2nd European Conference on Technology Enhanced Learning (EC-TEL 2007), Crete (Greece), 2007. (C. Romero, workshop co-chair).

6. Development of a tool for knowledge extraction based on evolutionary learning. More information about the KEEL project can be found on the WEB of the project http://www.keel.es. The following reference introduces KEEL:

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

7. Development of a Java class library for evolutionary computation, called JCLEC. This software has been used as computing engine in several KEEL modules. The software has been released as a free software tool, available on the WEB address http://jclec.sourceforge.net. The following reference introduces JCLEC:

- S. Ventura, C. Romero, A. Zafra, J.A. Delgado & C. Hervás. JCLEC: A Java Framework for Evolutionary Computation. *Soft Computing*, 12:4 (2008), 381-392.

8. According to the objectives of the project, we are working on the development of evolutionary algorithm models in different data mining areas. We indicate the objective (number and description of objective) and list the publications.

1.a Multiple instance learning

- A. Zafra, E. Gibaja and S. Ventura. Multi-instance Learning with Multi-Objective Genetic Programming for Web Mining. *Applied Soft Computing*, 11:1 (2011), 93-102.
- A. Zafra and S. Ventura. G3P-MI: A Genetic Programming Algorithm for Multiple Instance Learning. *Information Sciences*, 180:23 (2010), 4496-4513.
- A. Zafra, C. Romero, S. Ventura and E. Herrera-Viedma. Multi-Instance Genetic Programming For Web Index Recommendation. *Expert Systems with Applications*, 36 (2009), 11470-11479.

1.b Multi-label classification

- J. L. Ávila, E. L. Gibaja, A. Zafra and S. Ventura. A Niching Algorithm to Learn Discriminant Functions with Multi-Label Patterns. *Journal of Multiple-Valued Logic and Soft Computing*, 2010 (in press).

1.d. Association rule mining

- J.M. Luna, J.R. Romero, S. Ventura. Design and Behavior Study of a Grammar Guided Genetic Programming Algorithm for Mining Association Rules. *Knowledge and Information Systems*, 2011 (submitted).
- J.M. Luna, J.R. Romero, S. Ventura. G3PARM: A Grammar Guided Programming Algorithm for Mining Association Rules. *IEEE World Congress on* Computational *Intelligence (WCCI 2010)*. Barcelona, Spain. 2010.

2.c Development of high performance models based on GPGPU

- A. Cano, A. Zafra and S. Ventura. Speeding up GP classification algorithms on GPUs. *Soft Computing. Special Issue on Evolutionary Computation Models based on GPU*, 2010 (submitted).

2.b Feature selection methods in multiple instance learning

- A. Zafra, M. Pechenizkiy and S. Ventura. HyDR-MI: A Hybrid Algorithm to Reduce Dimensionality in Multiple Instance Learning. Information Sciences. Information Sciences, 2011 (accepted).
- A. Zafra, M. Pechenizkiy and S. Ventura. Relief-MI: A Feature Selection Method for Multiple Instance Learning. *Neurocomputing* (submitted).

3. Educational data mining

- C. Romero and S. Ventura. Educational Data Mining: A Review of the State-of-the-Art. IEEE Transactions on Systems, Man and Cybernetics. Part C: Applications and Reviews, 40:6 (2010), 601-618.
- C. Romero, S. Ventura, P. de Bra. Using Mobile and Web-based Computerized Test for Evaluating University Students. Computer Applications in Engineering Education, 17:4 (2009), 435-447.
- C. Romero, S. Ventura, A. Zafra and P. De Bra. Applying Web Usage Mining for Personalizing Hyperlinks in Web-based Adaptive Educational Systems. Computers and Education, 53 (2009), 828–840.
- E. García, C. Romero, S. Ventura and C. de Castro. An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. User Modelling and User Adapted Interaction, 19:1-2 (2009), 99-132.
- C. Romero, P. González, S. Ventura, M. J. del Jesus and F. Herrera. Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data. *Expert Systems With Applications*, 36:2 (2009), 1632-1644.
- C. Romero, S. Ventura and E. García. Data Mining in Course Management Systems: MOODLE Case Study and Tutorial. Computers and Education, 51:1 (2008), 368-384.
- C. Romero, S. Ventura. Educational Data Mining: A Survey from 1995 to 2005. *Expert Systems with Applications*, 33:1 (2007), 135-146.

3. Web mining

- J.M. Luna, A. Ramírez, J.R. Romero, S. Ventura. An Intruder Detection Approach Based on Infrequent Rating Pattern Mining. In Proceedings of the 10th Intl. Conference on Intelligent Systems, Design and Applications (ISDA). El Cairo, Egypt. 2010.
- A. Zafra, E. Gibaja and S. Ventura. Multi-instance Learning with Multi-Objective Genetic Programming for Web Mining. Applied Soft Computing, 11:1 (2011), 93-102.
- A. Zafra, C. Romero, S. Ventura and E. Herrera-Viedma. Multi-Instance Genetic Programming for Web Index Recommendation. Expert Systems with Applications, 36 (2009), 11470-11479.

Also, the members of the group have published articles in several international conferences among which we can cite the European Conference on Machine Learning, the Genetic and Evolutionary Computation Conference (GECCO), the IEEE Conference on Computational Intelligence (WCCI) or the International Conference on User Modelling, Adaptation and Personalization (UMAP).

6.2 PUBLIC AND PRIVATE GRANTED PROJECTS AND CONTRACTS OF THE RESEARCH GROUP
Indicate the project and contract grants during the last 5 years (2005-2009) (national, regional or international)
Include the grants for projects under evaluation

Title of the project or contract	Relationship with this proposal (1)	Principal Investigator	Budget	Funding agency and project reference	Project period (2)
			EUROS		
KEEL-CTNC: Evolutionary Multiple Instance Learning, Product Unit Neural Networks, Educational Data Mining and Web Mining	1	S. VENTURA	112530.00	Spanish Ministry of Science and Technology TIN2008-06681-C03-03/TIN	01/Jan/2009 31/Dec/2011 C
Aplicación de Técnicas de Extracción de Conocimiento en los Sistemas Educativos (ATECSE)	1	S. VENTURA	172743.68	Andalusian Research Plan. P08-TIC-03720	14/Jan/2009 13/Jan/2012 C
Second International Conference on Educational Data Mining	1	S. VENTURA	9000.00	Spanish Ministry of Science and Technology TIN2008-04060-E/TIN	01/Mar/2009 30/Nov/2009 C
Second International Conference on Educational Data Mining	1	S. VENTURA	5040.00	Andalusian Research Plan. IAC08-III-4185	01/Jul/2009 03/Jul/2009 C
Research group “Knowledge Discovery and Intelligent Systems” (KDIS)	2	S. VENTURA		Andalusian Research Plan Research Group TIC-2010	01/Jan/2010 31/Dec/2010 S

(1) Write 0, 1, 2 or 3 according to: 0 = Similar project; 1 = Very related; 2 = Low related; 3 = Unrelated.
(2) Write C or S if the project has been funded or it is under evaluation, respectively.

7. TRAINING CAPACITY OF THE PROJECT AND THE GROUP
(In the case of Coordinated Projects this issue must be filled by each partner)

The capability of our research group to correctly form doctoral students is fully guaranteed if we consider that:

- 6 of the 10 members in the project are PhD. Also, 2 of the PhD students have their theses at a very advanced stage:

    - José Luis Ávila, supervised by Drs. Sebastian Ventura and Eva Gibaja.
    - José María Luna, supervised by Drs. Sebastián Ventura and José Raúl Romero.

  We expect them to read their dissertations during or at the end of 2011.
- In recent years, three doctoral theses have been directed by members of the research group (for a detailed list of PhD theses, see http://www.uco.es/grupos/kdis). The quality of doctoral dissertations presented in the framework of the group is accredited by the publications described in the curricula of the research members, some of them developed in collaboration with national and international contacts maintained during their doctoral work.

- All PhD members on the team participate in the PhD program "Engineering and technology" of the University of Cordoba (research line "Machine Learning and Data Mining"), and Dr. Ventura also participates in the PhD program "Soft Computing and Intelligent Systems" (Department of Computer Sciences and Artificial Intelligence, University of Granada).

At this moment, according to the number of PhD and non-PhD members in the project, and taking into account the aforementioned aspects, we think that the presence of PhD students is positive, and guarantees her/his formation with respect to reaching a future PhD degree.