Special Issue on Pattern Recognition Techniques in Data Mining

Eva Armengol[a], Dionís Boixader[b], Francisco Grimaldo[c]

[a] *Artificial Intelligence Research Institute (IIIA-CSIC), Campus UAB, E-08193 Bellaterra, Catalonia, Spain.*
[b] *ETS d'Arquitectura del Vallès. Dept. Tecnologia de l'Arquitectura, Universitat Politècnica de Catalunya, 08190 Sant Cugat del Vallès, Catalonia, Spain*
[c] *Departament d'Informàtica. Escola Tècnica Superior d'Enginyeria (ETSE-UV), Universitat de València, Av. de la Universitat, s/n. 46100-Burjassot, Spain*

1. Introduction

The ever-increasing amount of readily available data makes the use of automatic tools unavoidable. Enterprises, public organisations, and a wide variety of customers feel the need to search for hide patterns behind the raw data. Their goal may be very different, enterprises may want to know about the behaviour of potential clients to refine offers and increase benefits. Health related public organisations may be interested in detecting some kind of prevalence and evolution of diseases in order to either prevent or palliate their effects on the population. Organisations involved in education may want to find new technologies better adapted to the capabilities and life style of individuals in order to improve learning. Data Mining is defined as the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The goal of data mining is to extract useful information, mainly patterns among subsets of data, that could be further used for tasks such as prediction or classification.

Data Mining involves many different techniques for pre-processing, analysing and interpreting data. These techniques fall mainly in two fields: Pattern Recognition and Machine Learning. The goal of pattern recognition is the identification of implicit objects and relations, i.e., the extraction of patterns from the input data. These techniques are mainly related with image analysis although this is not the only kind of application. Machine Learning techniques are mainly applied to extract generalised knowledge from data (including images) that will be further used for predictive tasks. Machine learning techniques can be classified according to the

input data: *supervised learning* techniques are used when the input data are labeled with class names, meaning that they can be naturally divided into groups of data; while in *unsupervised learning* techniques input data do not have a class label. This last group of techniques include *clustering* algorithms whose goal is to group data by similarity. Clustering techniques are particularly useful for data mining since generally, class labels for input data are not available.

Pattern recognition and Machine Learning techniques face a new problem when they are used for Data Mining: the huge amount of input data and their high dimensionality tend to make impractical the direct use of such techniques because of computational complexity, memory shortage, or both. This means that some kind of strategies such as dimensionality reduction, feature extraction or distribution have to be used.

The aim of the present special issue is to illustrate the current research trends on Pattern Recognition and Machine Learning motivated by Data Mining goals and problems, and to do so by means of a sample of selected papers which are representative of such trends and techniques. Some papers describe methodologies to improve the classification results of some already known method, while some others propose new approaches for solving a particular problem. The papers are grouped into four areas: classification, clustering, dimensionality reduction and applications.

Given a set of examples with known class labels, the goal of the classification task is to predict the class labels of new unseen examples based on the features of these known examples. An interesting issue in classification is how to deal with unbalanced data sets. This is specially important in data mining where most of the data can belong to a class (think, for instance, in a data base from a city where the majority of records may belong to healthy people but our interest is to analyse people with a given disease). This issue is addressed in the paper by Devi et al., where a method is proposed based on outlier detection, and introduction of redundant and noisy instances for improving the results of the classification task when classes are unbalanced. A common classification method is Support Vector Machines. However, this method is very sensitive with respect to input parameters. To address this issue Tharwat et al., propose an algorithm to optimise the input parameters of a Support Vector Machine classifier. Finally, a key issue in Data Mining is the handling of the huge amount of input data. Using distributed data provides a way of tackling this problem. Two of the papers presented here deal with this particular issue. The one by Limón et al., introduces a windowing strategy to construct induction trees from distributed data; and the one by Fan et al., proposes a method for learning distributed representations of entities, relations and words within entity descriptions. Finally, Bakhtiary et al. propose a new approach to improve the efficiency of convolutional

neural networks for classification tasks.

The goal of clustering algorithms is to group data by similarity. However, it is clear that their results greatly depend on parameters such as the expected number of clusters, the similarity function and the input data. There is a great variety of clustering algorithms depending on the criteria that they use to extract the clusters. A common strategy for clustering in Data Mining problems is the construction of density models which determine the dense regions in the data space and turn them into clusters. In this special issue there are two papers that consider this kind of clustering: the paper by Louichi et al., and the paper by Capdevila et al. The paper by Louichi et al., introduce a new density-based clustering algorithm based on k-NN and exponential spline interpolation, to obtain the structure of the data set. In that way, the algorithm shows different density levels and characterises the clusters of each level according to their density. Capdevila et al., propose an event discovery technique based on the DBSCAN density-based clustering algorithm. The proposed technique models textual content through a probabilistic topic model and considers the Jensen-Shannon distance for the task of neighbourhood identification in the textual dimension. Sometimes, clustering can be formulated as a multi-objective optimisation problem and this is the approach taken by Garcia-Piquer et al. that focus on optimising a multi-objective evolutionary clustering algorithm based on making subsets from the original large dataset, reducing in that way computational and memory costs. A different approach for optimisation is the one taken by Banharnsakun that proposes a new data clustering method where the Artificial Bee Colony algorithm is implemented based on the MapReduce model. Another important aspect is how to interpret the clusters created by an algorithm. Because of the great amount of data, clusters can include a lot of apparently unrelated objects. Sevilla-Villanueva et al., propose a methodology that facilitates the discovery and the understanding of complex patterns inside clusters.

Sometimes dimensionality reduction is considered as a step to be performed to pre-process the input data. However, it is a wide research field by itself. Feature selection and feature extraction methods aim to detect and eliminate redundant information and take only those features considered relevant for the task at hand. This is a key step in data mining where data are complex and lead to high costs of both memory, space and time. The selection of the appropriate features greatly influences the quality of the data used by the algorithms and, thus, the goodness of the final result. There are two papers focused on this issue. López-Iñesta et al., analyse how the representation of data influences the performance of the classification, and propose a new similarity learning method that combines feature extraction and feature expansion. A different approach for dimensionality reduction is the use of associa-

tive memories since they allow the retrieval of data by (part of) their content. The paper by Ramírez-Rubio et al., presents a classification algorithm using associative memories that is an extension of the Alpha-Beta Associative Memory.

An important aspect of Data Mining techniques is to prove their feasibility in real applications. The present special issue includes a set of papers that, while presenting novel algorithms and methodologies, are mainly relevant by both the choice and the handling of their target application domains. Such application domains are quite different in nature, ranging from text to video and time series, and the strategies for dealing with each particular situation differ a great deal from each other. There are three examples of applications on textual documents. Chen presents a distance-based method to assess a weight to a term according to the text. This allows a more accurate comparison of documents for classification and clustering purposes. Tamen et al., propose a multi-classifier for Arabic handwritten recognition. The representation used for input data is based on Chebyshev moments and Contour-based features and the authors explore combinations of several classifiers such as Multilayer Perceptron, Support Vector Machines and Extreme Learning Machine. A different kind of textual analysis is the one introduced in the paper by Bandhakavi et al., where authors propose a model based on a domain specific emotion lexicon that is able to extract effective features for emotion classification. Finally, Pushpalatha and Ananthanarayana propose the representation of multimedia documents by means of trees. Pattern Recognition techniques are traditionally used for image analysis with two different purposes: object detection or diagnose. The paper by Sharif and Hölzel is included in the first group since authors perform an in-depth study about how several kinds of pre-filters fit the object detection task. In addition, authors introduce a new methodology to evaluate pre-filters. The papers by Abdel-Nasser et al., and by Martín et al., can be included in the second group. Abdel-Nasser et al., propose a method to analyse images of breast tumours pursuing the quantification and visualisation of the changes and, thus, it helps physicians to best adjust the treatment of the patient. Martín et al., propose a method for feature selection based on kernel alignment with the ideal kernel in Support Vector Machines. This method is tested on a data base of magneto-encephalogram recordings and the goal is to provide support for the diagnosis of Schizophrenia. Finally, there is the paper by Coniglio et al., that uses video images as input. The paper focus on the problem of detection of people in images and use several state-of-the-art approaches to improve the results of typical people detection methods. After an accurate evaluation of these enhancements, authors propose a segmentation scheme that incorporates colour space change, a weighted combination of mean shape and appearance-based priors, and shape cut clustering. There are also many applications that take as input data

the output of machine sensors and that, most of the times, imply to deal with time series. This is the case of the paper by Das and Ghosh whose goal is to improve the meteorological prediction. Authors propose a multivariate prediction approach based on a variant of Semantic Bayesian Networks.

The editors want to thank the authors and reviewers for their enthusiastic efforts, and the Pattern Recognition Letters' team for their generous and efficient support.