# Semantic Annotation of Classification Data for KDD Support Services

Claudia Diamantini, Domenico Potena, and Maurizio Panti

Dipartimento di Ingegneria Informatica, Gestionale e dell'Automazione,
Università Politecnica delle Marche - via Brecce Bianche, 60131 Ancona, Italy
{diamanti,d.potena,panti}@diiga.univpm.it

**Abstract.** In order to guide the user in the correct and effective use of tools, a KDD support system should posses a knowledge on the mapping between certain characteristics of data (and the Data Mining task) and the tools which could be profitably used. In this work, we want to dwell upon this problem in the framework of the classification data mining task, by showing an approach to extract some knowledge about the data from the form of the decision border.

## 1 Introduction

In order to guide the user in the correct and effective use of Knowledge Discovery in Databases (KDD) tools, a KDD support system should posses a knowledge on the mapping between certain characteristics of data (and the Data Mining task) and the tools which could be profitably used [4]. In this work, we want to dwell upon this problem in the framework of the classification data mining task, which reads: "find the (unknown) class an instance belongs to, out of $k$ predefined classes". In the literature, the problem is tackled from either a knowledge representation [1] or knowledge induction [5] perspective. However, these approaches do not exploit the information about the decision border, even if it is known that the characteristics of the decision border define the type of classification problem and a relationship exists between the characteristics of the decision border and the performance of different classification architectures.

## 2 An Approach to Decision Boundary Characterization

We propose a method to derive an analytical description of (an approximation of) the decision border. The method resorts to the nearest neighbor Vector Quantizer Architecture with Euclidean distance (VQA), trained by the BVQ algorithm [3] to induce a piecewise linear approximation of the true decision border. The geometrical characteristics of the VQA allowed us to develop an algorithm to extract the analytical Voronoi description in n-dimensional spaces, hence the equations of hyperplanes and vertices of the decision regions.

Elaborating upon the analytical definition of decision borders, we are able to extract any kind of geometrical characteristics of decision regions, for instance

M. Bouzeghoub et al. (Eds.): ICSNW 2004, LNCS 3226, pp. 313–315, 2004.

describing both (1) topological and (2) geometrical properties. The former focuses on properties such as the Connected/Disconnected and the Open/Close properties of the regions. The latter refines the regions categorization describing features like: the surface area, volume, principal components ratio, convexity, volume/surface ratio, position in $\mathcal{R}^n$, and so on.

## 3 Applications

Geometrical characteristics of decision region can be associated to classification data as a sort of semantic annotation of the classification problem. This semantic description can be exploited to build intelligent services supporting the users in each step of the classification KDD process design, from the understanding and pre-processing of the input data, to the understanding of the discovered model.

In the framework of distributed KDD support systems [4], the semantic annotation of data can be exploited by linking it to information about the data mining techniques used to induce the model and their performances. To represent such information, suitable description languages and ontologies can be used, like the DAMON ontology [2]. Then, *Meta Learning Services* can be developed, to search, organize and elaborate such information, to find similarities between datasets on the basis of their decision borders, and correlations between groups of similar datasets and the performance of different data mining techniques. From the relationship between data characteristics and algorithms performances the Meta Learning Service can then extract meta-information about the learning process and exploit it to support the user, by suggesting the set of algorithms, or the typical parameter setting for a given algorithm, that have demonstrated the best performances on data similar in characteristics to a given dataset.

## 4 Conclusions

The paper presents a method to associate decision border information to classification data. This can be considered as a domain knowledge about the classification task, and we discuss how this information can be exploited in the framework of distributed KDD support systems. An important issue related with the method is the accuracy of border approximation needed for different applications. Qualitatively, it turns out for instance that an accurate model of the border is not needed to understand and to pre-process the data, while it is needed to build up a valid knowledge base registry. We are forming a set of experiments to test cost and quality of the method for the different applications.

## References

1. Appice, A., Ceci, M. and Malerba, D. KDB2000: An Integrated Knowledge Discovery Tool. In *Data Mining III*, volume 6. WIT Press, 2002.
2. Cannataro, M. and Comito, C. A Data Mining Ontology for Grid Programming. In *Proc. 1st Work. on Semantics in Peer-to-Peer and Grid Computing*, 2003.

3. Diamantini, C. and Spalvieri, A. Quantizing for Minimum Average Misclassification Risk. *IEEE Trans. on Neural Networks*, 9(1):174–182, Jan. 1998.
4. Diamantini, C., Panti, M. and Potena, D. Services for knowledge discovery in databases. In *Int. Symp. of S.Caterina on Challanges in the Internet and Interdisciplinary Research*, volume 1, Jan. 29 - Feb. 1 2004.
5. Vilalta, R. and Drissi, Y. A perspective view and survey of meta-learning. *Artif. Intell. Rev.*, 18(2):77–95, 2002.