Data Mining as Generalization: A Formal Model

Ernestina Menasalvas[1] and Anita Wasilewska[2]

[1] Departamento de Lenguajes y Sistemas Informaticos Facultad de Informatica, U.P.M, Madrid, Spain ernes@fi.upm.es
[2] Department of Computer Science, State University of New York, Stony Brook, NY, USA anita@cs.sunysb.edu

Summary. The model we present here formalizes the definition of Data Mining as the process of information generalization. In the model the Data Mining algorithms are defined as generalization operators. We show that only three generalizations operators: classification operator, clustering operator, and association operator are needed to express all Data Mining algorithms for classification, clustering, and association, respectively. The framework of the model allows to describe formally the hybrid systems; combination of classifiers into multi-classifiers, and combination of clustering with classification.

We use our framework to show that classification, clustering and association analysis fall into three different generalization categories.

1 Introduction

During the past several years researches and practitioners in Data Mining and Machine Learning have created a number of complex systems that combine Data Mining or Machine Learning algorithms. The most common, successful and longstanding ([9]) is combination of classifiers. For example, a protein secondary structure prediction system **SSPro** ([1, 2]), combines 11 bidirectional recurrent neural networks. The final predictions are obtained averaging the network outputs for each residue. Another system, named **Prof** ([17]) combines (in multiple stages) different types of neural network classifiers with linear discrimination methods ([20]).

The most recent system for protein secondary structure prediction presented in [21] uses an approach of *stacked generalization*([**?**]). It builds layers of classifiers such that each layer is used to combine the predictions of the classifiers of its preceding layer. A single classifier at the top-most level outputs the ultimate prediction. Predictive accuracy of [21] system outperforms six of the best secondary structure predictors by about 2%.

Natural questions arise: which methods can be combined and which can not, and if one combines them how to interpret the results in a correct and consistent manner.

In attempt to address these questions we build a Formal Model in which basic Data Mining and Machine Learning algorithms and methods can be defined and discussed.

2 The General Model Framework

One of the main goals of Data Mining is to provide a comprehensible description of information we extract from the data bases. The description comes in different forms. In case of classification problems it might be a set of characteristic or discriminant rules, it might be a decision tree or a neural network with fixed set of weights. In case of association analysis it is a set of associations, or association rules (with accuracy parameters). In case of cluster analysis it is a set of clusters, each of which has its own description and a cluster name (class name) that can be written in a form of a set of discriminant rules. In case of approximate classification by the Rough Set analysis ([19],[22], [25], [26], [27]) it is usually a set of discriminant or characteristic rules (with or without accuracy parameters) or set of decision tables.

Of course any data table, or database can be always re-written as a set of rules, or some descriptive formulas by re-writing each record tuple (in attribute=value convention) as $a_1 = v_1 \cap a_2 = v_2 .... \cap a_n = v_n$ or as $a_1 = v_1 \cap a_2 = v_2 .... \cap a_n = v_n \Rightarrow c = c_k$, where a_i, v_i are attributes and their values defining the record and c is the classification (class) attribute with corresponding value c_i, if applicable. But by doing so we just choose another form of knowledge representation or syntactic form of the record (data base) and there is no generalization involved. The cardinality of the set of descriptions might be less then the number of records (some records might have the same descriptions once the key attribute is removed) but it is still far from our goal to have a short, comprehensible description of our information. In data mining methods we also look for shortening the length of descriptions. A database with for example 100 attributes would naturally "produce" descriptions of uncomprehensible length. The "goodness" of data mining algorithm is being judged, in between, by these two factors: number of descriptions and their length. If these two conditions are met we say that we generalized our knowledge from the initial database, that we "mined" a more comprehensive, more general information. If we for example would reduce a 100,000 descriptions (records of the database) of size 500 (number of attributes of the initial database) to 5 descriptions of size 4 we surely would say that we generalized our information, but the question would arise of the quality of such generalization and hence the validity of our method.

The formal model we present here addresses all problems described above. It also provides framework for a formal definitions of intuitive notion of gener-