Minimum Description Length Adaptive Bayesian Mining

Diego Liberati
Italian National Research Council, Italy

INTRODUCTION

In everyday life, it often turns out that one has to face a huge amount of data, often not completely homogeneous, often without an immediate grasp of an underlying simple structure. Many records, each instantiating many variables are usually collected with the help of several tools.

Given the opportunity to have so many records on several possible variables, one of the typical goals one has in mind is to classify subjects on the basis of a hopefully reduced meaningful subset of the measured variables.

The complexity of the problem makes it natural to resort to automatic classification procedures (Duda and Hart, 1973) (Hand et al., 2001). Then, a further questions could arise, like trying to infer a synthetic mathematical and/or logical model, able to capture the most important relations between the most influencing variables, while pruning (O'Connel 1974) the not relevant ones. Such interrelated aspects will be the focus of the present contribution.

In the First Edition of this encyclopedia we already introduced three techniques dealing with such problems in a pair of articles (Liberati, 2005) (Liberati et al., 2005). Their rationale is briefly recalled in the following background section in order to introduce the kind of problems also faced by the different approach described in the present article, which will instead resort to the Adaptive Bayesian Networks implemented by Yarmus (2003) on a commercial wide spread data base tool like Oracle.

Focus of the present article will thus be the use of Adaptive Bayesian Networks are in order to unsupervisedly learn a classifier direcly form data, whose minimal set of features is derived through the classical Minimun Description Lenght (Barron and Rissanen, 1998) popular in information theory.

Reference will be again made to the same popular micro-arrays data set also used in (Liberati et al., 2005), not just to have a common benchmark useful to compare results and discuss complementary advantages of the various procedures, but also because of the increasing relevance of the bioinformatics field itself.

BACKGROUND

The introduced tasks of selecting salient variables, identifying their relationships from data and infer a logical and/or dynamical model of interaction may be sequentially accomplished with various degrees of success in a variety of ways.

In order to reduce the dimensionality of the problem, thus simplifying both the computation and the subsequent understanding of the solution, the critical problem of selecting the most relevant variables must be solved.

The very simple approach to resort to cascading a Divisive Partitioning of data orthogonal to the Principal Directions – PDDP - (Boley 1998) and *k-means*, already proven to be a good way to initialize k-means (Savaresi and Booley, 2004) and to be successful in the context of analyzing the logs of an important telecom provider (Garatti et al., 2004), was presented in (Liberati et al., 2005) with reference to a paradigmatic case of micro-arrays data in bioinformatics

A more sophisticated possible approach is to resort to a rule induction method, like the one described as Hamming Clustering in Muselli and Liberati (2000). Such a strategy also offers the advantage to extract underlying rules, implying conjunctions and/or disjunctions between the identified salient variables. Thus, a first idea of their even non-linear relations is provided as a first step to design a representative model, whose variables will be the selected ones. Such an approach has been shown (Muselli and Liberati, 2002) to be not less powerful over several benchmarks than the popular decision tree developed by Quinlan (1994). Then, a possible approach to blindly build a simple linear approximating model is to resort to piece-wise affine (PWA) identification of hybrid systems (Ferrari-Trecate

et al., 2003). The cascading of such two last approaches has been proposed in (Liberati, 2005).

Here just a few more approaches are recalled among the most popular ones, to whom the ones used either here or in either (Liberati, 2005) or (Liberati et al., 2005) are in some way comparable. For a widest bibliography, one could refer to both edition of this Encyclopedia, and in particular to the bibliography cited in the referenced papers. Among the simplest approaches, principal components (MacQueen, 1967) (Golub and van Loan, 1996) help to order the variables from the most relevant to the least one, but only under a linear possibly homogeneous framework. Partial least squares do allow to extend to non-linear models, provided that one has prior information on the structure of the involved non-linearity; in fact, the regression equation needs to be written before identifying its parameters. Clustering (Kaufman and Rousseeuw, 1990) (Jain and Dubes, 1998) (Jain at al., 1999) is instead able to operate even in an unsupervised way, without the a priori correct classification of a training set, but even fuzzy (Karayiannis and Bezdek 1997) and supervised (Setnes, 2000) approaches have been explored. Neural networks are known to learn the embedded rules, but their possibility to make rules explicit (Taha & Ghosh, 1999) or to underline the salient variables is only indirect. Support Vector Machines (Vapnik, 1998) are a very simple and popular general purpose approach whose theoretic foundation makes it worth in many applications.

MAIN THRUST OF THE CHAPTER

Adaptive Bayesian Networks

A learning strategy searching for a trade-off between a high predictive accuracy of the classifier and a low cardinality of the selected feature subset may be derived according to the central hypothesis that a good feature subset should contain features that are highly correlated with the class to be predicted, yet uncorrelated with each other.

Based on information theory, the Minimum Description Length (MDL) principle (Barron and Rissanen, 1998) provides the statement that the best theory to infer from training data is the one that minimizes both the length (i.e. the complexity) of the theory itself and the length of the data encoded with respect to it. In particular, MDL can thus be employed as a criteria to judge the quality of a classification model.

The motivation underlying the MDL method is to find a compact encoding of the training data. To this end, the MDL measure introduced in Friedman et al. (1997) can be adopted, weighting how many bits one do need to encode the specific model (i.e. its length), and how many bits are needed to describe the data based on the probability distribution associated to the model.

This approach can be applied to address the problem of feature selection, by considering each feature alone as a simple predictive model of the target class. As described in (Kononenko1995), each feature can be ranked according to its description length, that reflects the strength of its correlation with the target. In this context, the MDL measure is given by Yarmus (2003), again weighting the encoding length, where one have one sub-model for each value of the feature, with the number of bits needed to describe the data, based on the probability distribution of the target value associated to each sub-model.

However, once all features have been ordered by rank, no a priori criterion is available to choose the cut-off point beyond which features can be discarded. To circumvent this drawback, one can start with building a classifier on the set of the n-top ranked features. Then, a new feature is sequentially added to this set, and a new classifier is built, until no improvement in accuracy is achieved.

Our approach takes into account two different classifiers derived from Bayesian Networks, i.e. the Naïve Bayes (NB) and the Adaptive Bayesian Network (ABN).

NB is a very simple Bayesian network consisting of a special node (i.e. the target class) that is parent of all other nodes (i.e. the features or attributes) that are assumed to be conditionally independent, given the value of the class. The NB network can be "quantified" against a training dataset of pre-classified instances, i.e. we can compute the probability associated to a specific value of each attribute, given the value of the class label. Then, any new instance can be easily classified making use of the Bayes rule. Despite its strong independence assumption is clearly unrealistic in several application domains, NB has been shown to be competitive with more complex state-of-the-art classifiers (Friedman et al., 1997) (Keogh and Pazzani., 2002) (Cheng and Greiner,1999).