Machine Learning Through Data Mining

Diego Liberati
Italian National Research Council, Italy

INTRODUCTION

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

Given the opportunity to have so many data not easy to correlate by the human reader, but probably hiding interesting properties, 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 worthwhile to resort to automatic classification procedures.

Then, the question arises of reconstructing a synthetic mathematical model, capturing the most important relations between variables, in order to both discriminate classes of subjects and possibly also infer rules of behaviours that could help identify their habits.

Such interrelated aspects will be the focus of the present contribution. The data mining procedures that will be introduced in order to infer properties hidden in the data are in fact so powerful that care should be put in their capability to unveil regularities that the owner of the data would not want to let the processing tool discover, like for instance, in some cases the customer habits investigated via the usual smart card used in commerce with the apparent reward of discounting.

Four main general purpose approaches will be briefly discussed in the present article, underlying the cost effectiveness of each one.

In order to reduce the dimensionality of the problem, simplifying both the computation and the subsequent understanding of the solution, the critical issues of selecting the most salient variables must be addressed. This step may already be sensitive, pointing to the very core of the information to look at.

A very simple approach is to resort to cascading a divisive partitioning of data orthogonal to the principal directions (PDDP) (Boley, 1998) already proven to be successful in the context of analyzing micro-arrays data (Garatti, Bittanti, Liberati, & Maffezzoli, 2007).

A more sophisticated possible approach is to resort to a rule induction method, like the one described in Muselli and Liberati (2000). Such a strategy also offers the advantage to extract underlying rules, implying conjunctions or disjunctions between the identified salient variables. Thus, a first idea of their even nonlinear 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 & Liberati, 2002) to be not less powerful over several benchmarks than the popular decision tree developed by Quinlan (1994). An alternative in this sense can be represented by Adaptive Bayesian networks (Yarmus, 2003) whose advantage is also to be available on a commercial wide spread data base tool like Oracle.

Dynamics may matter. A possible approach to blindly build a simple linear approximating model is thus to resort to piece-wise affine (PWA) identification (Ferrari-Trecate, Muselli, Liberati, & Morari, 2003).

The joint use of (some of) such four approaches briefly described in this article, starting from data without known priors about their relationships, will allow to reduce dimensionality without significant loss in information, then to infer logical relationships, and, finally, to identify a simple input-output model of the involved process that also could be used for controlling purposes, even those potentially sensitive to ethical and security issues.

BACKGROUND

The introduced tasks of selecting salient variables, identifying their relationships from data, and classifying possible intruders may be sequentially accomplished with various degrees of success in a variety of ways:

- Principal components order the variables from the most salient to the least one, but only under a linear framework.
- Partial least squares do allow to extend to nonlinear models, provided that one has prior information on the structure of the involved nonlinearity; in fact, the regression equation needs to be written before identifying its parameters.
- Clustering may operate even in an unsupervised way without the a priori correct classification of a training set (Boley, 1998).
- Neural networks are known to learn the embedded rules with the indirect possibility (Taha & Ghosh,

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

1999) to make rules explicit or to underline the salient variables.
- Decision trees (Quinlan, 1994) are a popular framework providing a satisfactory answer to the recalled needs.

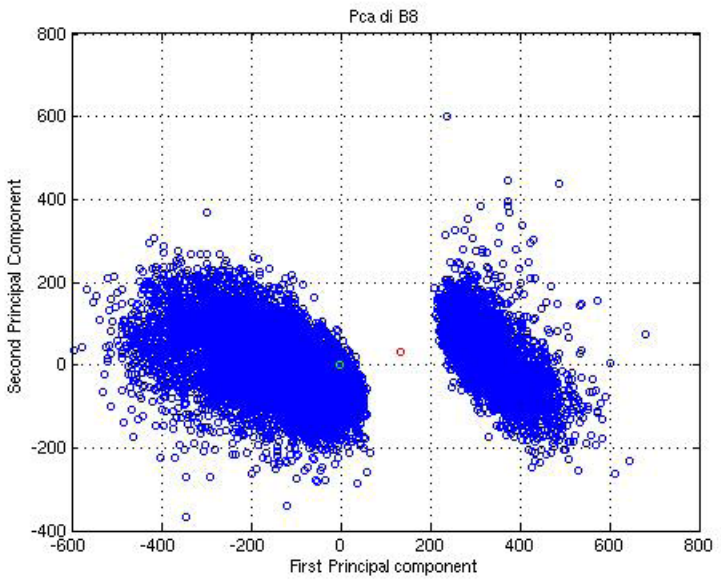
RECENT DEVELOPMENTS

Unsupervised Clustering

In this contribution, we will firstly resort to a quite recently developed unsupervised clustering approach, the Principal Direction Divisive Partitioning (PDDP) algorithm, proposed in Boley (1998). According to the analysis provided in Savaresi and Boley (2004), PDDP is able to provide a significant improvement of the performances of a classical k-means approach (Hand, Mannila, & Smyth, 2001; MacQueen, 1967), when PDDP is used to initialize the k-means clustering procedure. The approach taken herein may be summarized in the following three steps, the second of which is the core of the method, while the first one constitutes a preprocessing phase useful to ease the following tasks, and the third one is a postprocessing step designed to focus back on the original variables.

1. A principal component analysis defines a hierarchy in the transformed orthogonal variables according the principal directions of the data set. Principal Component Analysis (O'Connel, 1974; Hand et al., 2001) is a multivariate analysis designed to select the linear combinations of variables with higher intersubject covariances. Such combinations are the most useful for classification. More precisely, it returns a new set of orthogonal coordinates of the data space, where such coordinates are ordered in decreasing order of intersubject covariance.
2. The unsupervised clustering is performed by cascading a noniterative technique, the PDDP, (Booley, 1998) based upon singular value decomposition (Golub & van Loan, 1996), and the iterative centroid-based divisive algorithm k-means (MacQueen, 1967). Such a cascade, with the clusters obtained via PDDP used to initialize k-means centroids, is shown to achieve best performances in terms of both quality of the partition and computational effort (Savaresi & Boley, 2004). The whole dataset is thus bisected into two clusters, with the objective of maximizing the distance between the two clusters and, at the same time, minimizing the distance among the data points lying in the same clusters. The classification is achieved without using a priori information on the user (unsupervised learning), thus automatically highlighting the user belonging to a (possibly unknown) user class.
3. By analyzing the obtained results, the number of variables needed for the clustering may be reduced by pruning all the original variables that are not needed in order to define the final partitioning hyperplane, so that the classification eventually is based on a few variables only.

Figure 1. Clustering according to principal components



5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/machine-learning-through-data-mining/13931

Related Content

Credit Card Users' Data Mining
André de Carvalho, Antonio P. Braga and Tersea Ludermir (2005). *Encyclopedia of Information Science and Technology, First Edition (pp. 603-605).*
www.irma-international.org/chapter/credit-card-users-data-mining/14305

Designing for Service-Oriented Computing
Bill Vassiliadis (2007). *Journal of Cases on Information Technology (pp. 36-53).*
www.irma-international.org/article/designing-service-oriented-computing/3193

NARA: A Digitization Case Study
Kristen Cissne (2014). *Cases on Electronic Records and Resource Management Implementation in Diverse Environments (pp. 306-317).*
www.irma-international.org/chapter/nara-digitization-case-study/82656

Risk Management of ERP Projects in Manufacturing SMEs
Päivi Iskanius (2010). *Information Resources Management Journal (pp. 60-75).*
www.irma-international.org/article/risk-management-erp-projects-manufacturing/43721

The Expert's Opinion
Karen D. Walker (1995). *Information Resources Management Journal (pp. 35-36).*
www.irma-international.org/article/expert-opinion/51005