Chapter 10
Machine Learning in Studying the Organism's Functional State of Clinically Healthy Individuals Depending on Their Immune Reactivity

Tatiana V. Sambukova
Military Medical Academy, Russian Federation

ABSTRACT

The work is devoted to the decision of two interconnected key problems of Data Mining: discretization of numerical attributes, and inferring pattern recognition rules (decision rules) from training set of examples with the use of machine learning methods. The method of discretization is based on a learning procedure of extracting attribute values' intervals the bounds of which are chosen in such a manner that the distributions of attribute's values inside of these intervals should differ in the most possible degree for two classes of samples given by an expert. The number of intervals is defined to be not more than 3. The application of interval data analysis allowed more fully than by traditional statistical methods of comparing distributions of data sets to describe the functional state of persons in healthy condition depending on the absence or presence in their life of the episodes of secondary deficiency of their immunity system. The interval data analysis gives the possibility (1) to make the procedure of discretization to be clear and controlled by an expert, (2) to evaluate the information gain index of attributes with respect to the distinguishing of given classes of persons before any machine learning procedure (3) to decrease crucially the machine learning computational complexity.

DOI: 10.4018/978-1-4666-1900-5.ch010

INTRODUCTION

Machine Learning is a particular direction in Data Mining related to extracting conceptual knowledge from data in the form of logical and association dependencies or links "object ↔ classes", "object ↔ properties", "properties ↔ classes", "properties ↔ properties", "classes ↔ classes", and "objects ↔ objects. Initial data for machine learning processes must be conceptualized, i.e., attribute values must be represented by the use of nominal, integer, or Boolean meaningful scales (discrete numerical or nominal attributes (features)). Data reduction of quantitative attributes means representing value domain as a set of meaningful discrete intervals. It is performed by dividing the values of a continuous attribute into a small number of intervals (or, equivalently, a set of cut-points) where each interval is mapped to a discrete symbol. Therefore discretization involves two decisions, on the number of intervals and the placement of interval boundaries.

Thus discretization is a key pre-processing step of the machine learning tasks. It offers some cognitive benefits as well as computational ones and improves performance of knowledge discovery process. The boundaries of informative diapasons of an attribute's values are very important conceptual knowledge. However the raw data discretization is frequently integrated with constructing decision rules or trees like it is done in "on-line" discretization (Bruha, Kockova; 1994) or in Naпve Bayes Classificators (Friedman et al., 1998).

The "on-line" discretization, in contrast to "off-line" discretization, is performed by a number of machine learning algorithms for inferring decision trees or decision rules from examples (Fayyad & Irani, 1992, 1993; Perner, & Trautzsch, 1998). In these on-line approaches, the process of extracting the cut-off points in attribute ranges remains hidden from the experts.

Discretization may be used for different purposes. For example, it involves a variable (feature) selection method that can significantly improve the performance of classification algorithms used in the analysis of high-dimensional biomedical data (Liu, & Setiono, 1997; Lustgarten et al., 2008a). In (Abraham et al., 2009), a hybrid feature selection algorithm CHIWSS is described that helps in achieving dimensionality reduction by removing irrelevant data. This leads to increasing the learning accuracy and improving result comprehensibility.

For discretization, some learning algorithms are used (for instance, the Naive Bayes method (Lustgarten et al., 2008b; Abraham et al., 2009)). Furthermore, discretization itself may be viewed as a discovery of knowledge procedure revealing critical attribute values in a continuous domain.

A number of approaches have been suggested for attribute discretization. The methods of discretization restricted to single continuous attribute are called local, while methods that simultaneously convert all continuous attributes are called global. The global discretization methods are usually based on cluster analysis. In (Chmielewski, & Grzymala-Busse, 1996), a method of transforming any local discretization method into a global one is presented.

Two main distinct categories of Discretization methods are: unsupervised methods, which do not use any information of the target variable (disease state, for example), and supervised methods, which do it (Dougherty, et al., 1995). Some well-known unsupervised discretization algorithms are the following ones: the equal-width discretization (EWD), equal frequency discretization (EFD) (Jiang et al., 2009), Minimum Descriptive Length (MDL) discretization (Rissanen, 1987), and entropy based heuristic discretization (Fayyad, & Irani, 1992; Chiu et al., 1991).

Quantitative attributes are usually discretized in Naive-Bayes learning (Boullé, 2004). Under establishing simple conditions, the discretization is equivalent to using the true probability density function during Naпve-Bayes learning. Two efficient unsupervised discretization methods based on Naive-Bayes learning are proposed: the