TUTORIAL

Machine learning and association rules

Petr Berka & Jan Rauch

(Abstract)

Statistics and data mining (machine learning) have common aims in discovering structure in data. So some people (mainly statisticians) consider data mining as a part of statistics while others (mainly people from machine learning area) consider statistics a part of machine learning. Indeed, methods like regression analysis, discriminant analysis or clustering can be found in both areas. But there are also significant differences between these areas: statistical analysis is usually “hypothesis driven“– we start with the initial hypothesis, then we collect the data, and then we perform the experiments, whereas data mining and machine learning are “data driven” – we start with the available data and try to find “something” interesting.

There are a number of different types of machine learning algorithms that can be used for the (exploratory) data analysis: decision trees, decision rules, association rules, neural networks, SVM, bayesian classifiers. Some of them are based on principles taken from the area of artificial intelligence, some of them use approaches and methods borrowed from statistics. Examples of useful combination of logic and statistic approaches to data analysis are so called association rules.

The term association rule was coined by R. Agrawal in the early 90th in relation to so called market basket analysis. In this analysis, transaction data recorded by point-of-sale (POS) systems in supermarkets are analyzed in order to understand the purchase behavior of groups of customers, and use it to increase sales, and for cross-selling, store design, discount plans and promotions. Anyway, the term association rules can be understood in more general way as representation of knowledge about a relation between two Boolean attributes (created e.g. as conjunctions attribute-value pairs) that are supported by the analyzed data.

This more general understanding of association rules is the basics of the GUHA method, an exploratory data method the goes back to mid 60th. The GUHA method uses observational calculi – special logical calculi formulas of which correspond to suitable statements on observed data. The statements are based on statistical approaches, i.e. estimates of various parameters or statistical hypothesis tests are used. Such statements make possible to accept statistical conclusions on the basis of observed data. The association rule – a formula of observational predicate calculus consists of two derived predicates and of a generalized quantifier.

There are generalized quantifiers corresponding both to simple statements on observational data and to various statistical hypothesis tests, e.g. to Chi-square test. In this way, a general relation of two derived predicates in a finite {0,1} – data matrix can be expressed as an association rule – a formula of an observational calculus. There are both practically useful and theoretically interesting results on observational calculi. The results concern various classes of association rules, deduction rules in observational calculi, dealing with missing information, definability of association rules in classical predicate calculi, etc. The results can be understood as the logic of association rules.

The GUHA method is realized by GUHA-procedures. A GUHA-procedure is a computer program, the input of which consists of the analyzed data and a simple definition of a large set of relevant patterns. The GUHA procedure automatically generates each particular pattern and tests if it is true for the analyzed data. The output consists of all prime patterns. The pattern is prime if it is true for the analyzed data and does not immediately follow from the other more simple output patterns.

The most important GUHA procedure is the ASSOC procedure. It mines for association rules – formulas of observational calculi. The ASSOC procedure does not use the well known apriori algorithm. Its implementation is based on representation of analyzed data by suitable strings of bits. The most used implementation of the GUHA procedure ASSOC is the 4ft-Miner procedure that is a part of the LISp-Miner system. The 4ft-Miner procedure has fine tools for defining sets of relevant association rules to be generated and tested. The 4ft-Miner procedure mines also for conditional association rules. This way a large variety of practically important tasks can be solved.

Tentative content:

The tutorial will start with reviewing the similarities and differences between statistics, machine learning and data mining. Then we will have a closer look on the knowledge discovery process as described by the CRISP-DM methodology. Here we will concentrate on various types of machine learning algorithms used for the modeling step and on the statistical approaches and methods used in these algorithms. Main attention will be paid to different types of association rules. We will introduce basic principles of the GUHA method that combines logical and statistical approaches to association rules mining, we will discuss the used observational calculi and the logic of association rules and its applications. We will also show how these principles have been implemented in the LISp-Miner system and how this system can be used for solving real machine learning and data mining tasks.