Association Rules and Statistics

Martine Cadot
University of Henri Poincaré/LORIA, Nancy, France

Jean-Baptiste Maj
LORIA/INRIA, France

Tarek Ziadé
NUXEO, France

INTRODUCTION

A manager would like to have a dashboard of his company without manipulating data. Usually, statistics have solved this challenge, but nowadays, data have changed (Jensen, 1992); their size has increased, and they are badly structured (Han & Kamber, 2001). A recent method—data mining—has been developed to analyze this type of data (Piatetski-Shapiro, 2000). A specific method of data mining, which fits the goal of the manager, is the extraction of association rules (Hand, Mannila & Smyth, 2001). This extraction is a part of attribute-oriented induction (Guyon & Elisseeff, 2003).

The aim of this paper is to compare both types of extracted knowledge: association rules and results of statistics.

BACKGROUND

Statistics have been used by people who want to extract knowledge from data for one century (Freeman, 1997). Statistics can describe, summarize and represent the data. In this paper data are structured in tables, where lines are called objects, subjects or transactions and columns are called variables, properties or attributes. For a specific variable, the value of an object can have different types: quantitative, ordinal, qualitative or binary. Furthermore, statistics tell if an effect is significant or not. They are called inferential statistics.

Data mining (Srikant, 2001) has been developed to precede a huge amount of data, which is the result of progress in digital data acquisition, storage technology, and computational power. The association rules, which are produced by data-mining methods, express links on database attributes. The knowledge brought by the association rules is shared in two different parts. The first describes general links, and the second finds specific links (knowledge nuggets) (Fabris & Freitas, 1999; Padmanabhan & Tuzhilin, 2000). In this article, only the first part is discussed and compared to statistics. Furthermore, in this article, only data structured in tables are used for association rules.

MAIN THRUST

The problem differs with the number of variables. In the sequel, problems with two, three, or more variables are discussed.

Two Variables

The link between two variables (A and B) depends on the coding. The outcome of statistics is better when data are quantitative. A current model is linear regression. For instance, the salary (S) of a worker can be expressed by the following equation:

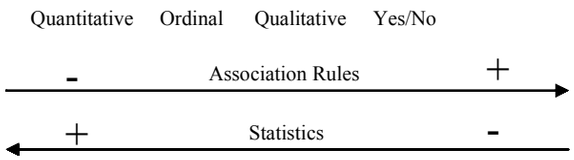
$$S = 100\,Y + 20000 + \varepsilon \qquad (1)$$

where Y is the number of years in the company, and ε is a random number. This model means that the salary of a newcomer in the company is $20,000 and increases by $100 per year.

The association rule for this model is: $Y \rightarrow S$. This means that there are a few senior workers with a small paycheck. For this, the variables are translated into binary variables. Y is not the number of years, but the property has seniority, which is not quantitative but of type Yes/No. The same transformation is applied to the salary S, which becomes the property "has a big salary."

Therefore, these two methods both provide the link between the two variables and have their own instruments for measuring the quality of the link. For statistics, there are the tests of regression model (Baillargeon, 1996), and for association rules, there are measures like support, confidence, and so forth (Kodratoff, 2001). But, depending on the type of data, one model is more appropriate than the other (Figure 1).

Figure 1. Coding and analysis methods



Three Variables

If a third variable E, the experience of the worker, is integrated, the equation (1) becomes:

$$S = 100\,Y + 2000\,E + 19000 + \varepsilon \qquad (2)$$

E is the property "has experience." If E=1, a new experienced worker gets a salary of $21,000, and if E=0, a new non-experienced worker gets a salary of $19,000. The increase of the salary, as a function of seniority (Y), is the same in both cases of experience.

$$S = 50\,Y + 1500\,E + 50\,E\,'\,Y + 19500 + \varepsilon \qquad (3)$$

Now, if E=1, a new experienced worker gets a salary of $21,000, and if E=0, a new non-experienced worker gets a salary of $19,500. The increase of the salary, as a function of seniority (Y), is $50 higher for experienced workers. These regression models belong to a linear model of statistics (Prum, 1996), where, in the equation (3), the third variable has a particular effect on the link between Y and S, called *interaction* (Winer, Brown & Michels, 1991).

The association rules for this model are:

- Y→S, E→S for the equation (2)
- Y→S, E→S, YE→S for the equation (3)

The statistical test of the regression model allows to choose with or without interaction (2) or (3). For the association rules, it is necessary to prune the set of three rules, because their measures do not give the choice between a model of two rules and a model of three rules (Zaki, 2000; Zhu, 1998).

More Variables

With more variables, it is difficult to use statistical models to test the link between variables (Megiddo & Srikant, 1998). However, there are still some ways to group variables: clustering, factor analysis, and taxonomy (Govaert, 2003). But the complex links between variables, like interactions, are not given by these models and decrease the quality of the results.

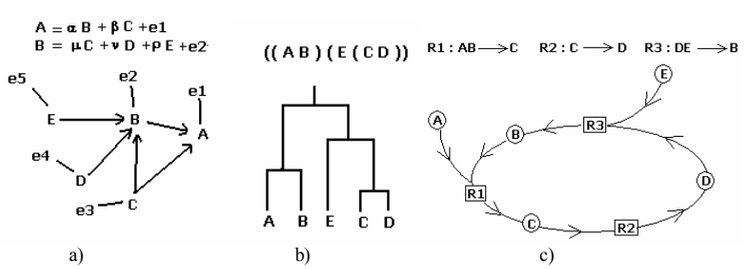
Comparison

Table 1 briefly compares statistics with the association rules. Two types of statistics are described: by tests and by taxonomy. Statistical tests are applied to a small amount of variables and the taxonomy to a great amount of

Table 1. Comparison between statistics and association rules

	Statistics		Data Mining
	Tests	Taxonomy	Association rules
Decision	Tests (+)	Threshold defined (-)	Threshold defined (-)
Level of Knowledge	Low (-)	High and simple (+)	High and complex (+)
No. of Variables	Small (-)	High (+)	Small and high (+)
Complex Link	Yes (-)	No (+)	No (-)

Figure 2. (a) Regression equations; (b) taxonomy; (c) association rules



a) b) c)