A Statistical Theory for Quantitative Association Rules*

Yonatan Aumann†
Bar-Ilan University
aumann@cs.biu.ac.il

Yehuda Lindell‡
IBM T.J.Watson Research
lindell@us.ibm.com

April 10, 2003

Abstract

Association rules are a key data-mining tool and as such have been well researched. So far, this research has focused predominantly on databases containing categorical data only. However, many real-world databases contain quantitative attributes and current solutions for this case are so far inadequate. In this paper we introduce a new definition of quantitative association rules based on statistical inference theory. Our definition reflects the intuition that the goal of association rules is to find extraordinary and therefore interesting phenomena in databases. We also introduce the concept of sub-rules which can be applied to any type of association rule. Rigorous experimental evaluation on real-world datasets is presented, demonstrating the usefulness and characteristics of rules mined according to our definition.

1 Introduction

Association rules. The goal of data mining is to extract higher level information from an abundance of raw data. Association rules are a key tool used for this purpose. An association rule is a rule of the form $X \Rightarrow Y$, where X and Y are events. The rule states that with a certain probability, called the *confidence* of the rule, when X occurs in the given database so does Y. A well-known application of association rules is in market basket data analysis.

The problem of mining association rules was first introduced by Agrawal et al. in [1], and later broadened in [2]. These papers related to the case of databases consisting of categorical attributes only (a categorical attribute is one which contains discrete, and typically unordered, data). Thus the events X and Y, on both sides of the rule, are the appearance of given *categorical* items. In this case, the aim is to find all rules with confidence and support above user-defined thresholds (*minconf* and *minsup*). Several efficient algorithms for mining categorical association rules have been published (see [2, 9, 13] for just a few examples). In all, the problem of mining categorical association rules has been extensively researched and is well understood, on both the algorithmic and conceptual levels.

*An extended abstract of this work appeared in KDD'99. This article has been revised and accepted for publication in the Journal of Intelligent Information Systems and is scheduled to appear in Vol. 20(3):255–283. Citations to and quotations from this work should reference that publication. If you cite this work, please check that the published form contains precisely the material to which you intend to refer.

†Address: Bar-Ilan University, Department of Computer Science, Ramat Gan, Israel 52900.

‡Work carried out while at Bar-Ilan University. Address: IBM T.J.Watson Research, 19 Skyline Drive, Hawthorne, New York 10532, USA.

Quantitative association rules. In practice, many, if not most, databases contain much quantitative data and are not limited to categorical items only. Unfortunately, the definition of categorical association rules does not translate directly to the quantitative case. It is therefore necessary to provide a definition of association rules for the case of a database containing quantitative attributes. Srikant and Agrawal [12] extended the categorical definition to include quantitative data. The basis for their definition is to map quantitative values into categorical events by considering *intervals* of the numeric values. Thus, each basic event is either a categorical item or a range of numerical values. An example of a rule according to this definition would be:

$$\textsf{sex} = \textsf{female and age} \in [20{,}30] \Rightarrow \textsf{wage} \in [\$5,\ \$10]\ (\textsf{conf. } 85\%)$$

Given this definition, [12] provides an algorithm which approximately finds all rules by employing a discretization technique. In addition, [12] provides an *interest* filter, aimed at reducing the problem that this definition is likely to yield many nested rules.

While the [12] definition and algorithm for quantitative association rules provide strong tools for mining quantitative data, there are also several drawbacks to their approach. First and foremost, the use of a range for describing a distribution of quantitative values can be limited and, at times, misleading. For example, the rule "height $\in$ [100cm,150cm] $\Rightarrow$ age $\in$ [0,14] (70%)" may be true even though few, if any, children under the age of one are 100cm tall (see Section 4.3 for real-world examples). In addition, the [12] definition often results in an exponential blowup of the number of rules, as the right-hand side of any given rule can always be enlarged. Hence, they place an a priori restriction on the maximum support of a rule (*maxsup*), partially solving this problem. Finally, the discretization employed in the mining algorithm results in loss of information. In particular, the algorithm can only *approximate* the best rules (see [12] for details). (We note that although this last drawback is algorithmic in nature, it also arises due to the nature of the definition.)

The aim of this work is to provide a new definition that both solves the above problems and is algorithmically workable.

1.1 A New Definition

The key contribution of this paper is in providing a new definition of quantitative association rules. Our definition is not based on a discretization paradigm that "converts" quantitative data into categorical items. Rather, we directly consider the *distribution* of the continuous data, via standard statistical measures such as mean and variance.

The new definition rests on the following abstraction of the notion of an association rule:

> An association rule indicates an association between a *subset* of the population, described by the left-hand side of the rule, and an extraordinary *behavior* of this subset, described by the right-hand side of rule.

The following is a valid association rule according to this definition:

$$\textsf{sex} = \textsf{female} \Rightarrow \textsf{mean wage} = \$7.90\ \textsf{p/hr (overall mean wage} = \$9.02)$$

This above rule states that the average wage for females is only \$7.90 per hour. This rule is (subjectively) *interesting* because it reveals a group of people earning a significantly lower wage than the average of \$9.02 p/hr (if the average wage was close to \$7.90 p/hr, then the fact that females also earn \$7.90 p/hr would be "uninteresting"). An integral aspect of our definition involves applying statistical tests to confirm the validity of our rules. That is, the above rule is only accepted if a statistical test (in this case, a Z-test) confirms that with high confidence one can infer that the mean wage of the female population is indeed lower than the mean wage of the rest of the population. (We note that the above rule was mined from a real database based on a 1985 census in the USA; see Section 1.3 below for more details.)

We now turn to a more formal explanation of our new definition. Before doing so, it is helpful to backtrack a little and discuss the goal and structure of association rules in general. Association rules are designed to help us discover what we may consider as "interesting" phenomena in databases. This is accomplished by pointing to subsets of transactions containing extraordinary behavior. As we have mentioned, each rule is comprised of a *left-hand side* and a *right-hand side*:

- The left-hand side of a rule is a description of a subset of the population.
- The right-hand side of a rule is a description of extraordinary behavior particular to the population described on the left-hand side.

Thus, the general structure of an association rule is:

$$\textsf{population-subset} \Rightarrow \textsf{extraordinary-behavior}$$

Defining "extraordinary behavior". An essential step in defining meaningful association rules is understanding what constitutes "extraordinary behavior". This involves, first and foremost, understanding what is meant by "behavior". The best way to describe the behavior of some population is by describing its *probability distribution* (or at least, by providing important information about this distribution). Thus, in the above example, we consider the distribution of the wage-attribute for the female population. This distribution (of the wages of female workers) then constitutes "extraordinary behavior" if it is significantly different to the distribution of the wage-attribute for the rest of the population. There are a number of standard measures used for describing a numerical distribution; in this work we focus mainly, although not exclusively, on the mean. We note that in the case of categorical attributes, the (discrete) distribution is described by the probability of each item appearing. Then, extraordinary behavior for a subset would be the appearance of some categorical item with probability significantly higher (or lower) than for the rest of the population. This is essentially the motivation behind the *confidence* and *lift* measures of association rules.

Confirming the validity of mined rules. So far we have discussed the "interestingness" of rules. However, a rule can only be interesting if it is correct. Clearly, the rule is correct with respect to the data in the database being mined (since it is extracted from this database). However, it is only meaningful if we can infer that it is also correct for the general population (where the database being mined is a random sample of this population). Therefore, a central aspect of our definition involves testing the validity of each rule with an appropriate statistical test.

In summary, an association rule under the new definition is a rule of the form:

$$\textbf{population-subset} \Rightarrow \textbf{mean} \textsf{ of values for the subset}$$

where the mean of the subset is significantly different to the mean of its complement in the database (and this is validated by an appropriate statistical test).

We note at this stage that such a definition can only be of interest if it is possible to find such rules efficiently. We indeed present efficient algorithms that demonstrate the feasibility of mining rules of this type. We also present a rigorous field study that shows that rules mined according to our definition are truly of help to end users.

1.2 Extensions

Other statistical measures. We note that a similar definition can be established using any other measure of probability distributions (e.g., variance or median). Thus, our definition actually provides a framework for an entire family of association rules. Although mean-based rules are the most natural, other measures provide important information. The variance of a subset, for example, points to the homogeneity of those included in the subset. We choose to focus on the mean and variance measures, as they are the most commonly used, and tend to provide most of the essential information regarding the distribution.

Sub-rules. In this paper, we also introduce the notion of sub-rules. Loosely speaking, a sub-rule of a (super) rule is one that highlights extraordinary behavior of a subset of the population of the super-rule. Consider, for example, the rule that shows that on average females earn significantly lower wages than their male counterparts. In this case, we can focus our attention to the distribution of wages for females only, and search for subsets of the women population whose wages are significantly higher or lower than \$7.90 p/hr. For example, the rule

$$\mathsf{Sex} = \mathsf{female\ and\ South} = \mathsf{Yes} \Rightarrow \mathsf{mean\ wage} = \mathsf{\$6.30\ p/hr}$$

is a sub-rule of the rule "$\mathsf{sex} = \mathsf{female} \Rightarrow \mathsf{mean\ wage} = \mathsf{\$7.90\ p/hr}$" and shows that although females overall are paid lower wages, in the South of the USA the situation is much worse. A rule in the opposite direction (showing a subset of the female population that earned above \$7.90, even if it was not above the overall average) would also be of interest. We note that the notion of sub-rules is not specific to quantitative association rules and can be applied to the categorical case as well.

1.3 Sample Results

In the coming sections we will present exact definitions, algorithms and an evaluation of our new concept. However, before delving into these details, we first present some sample results, obtained from an actual database.

We applied our algorithms to a database called *Determinants of Wages from the 1985 Current Population Survey in the United States* (the database may be found at [16]). The database contains 534 transactions and 11 attributes (7 categorical and 4 quantitative). Here are some of the rules discovered with 95% statistical confidence. The overall mean wage is \$9.02 per hour.

$$\begin{aligned} \mathsf{Sex} = \mathsf{female} \quad &\Rightarrow \quad \mathsf{mean\ wage} = \mathsf{\$7.90\ p/hr} \\ \mathsf{Sex} = \mathsf{female\ and\ South} = \mathsf{Yes} \quad &\Rightarrow \quad \mathsf{mean\ wage} = \mathsf{\$6.30\ p/hr} \end{aligned}$$

These rules show the dismal state of affairs regarding the wages of working women in the USA (and even more so in the south of the USA) in 1985. The other side of the coin of these two rules is the next rule also found:

$$\mathsf{Sex} = \mathsf{male\ and\ Race} = \mathsf{White} \quad \Rightarrow \quad \mathsf{mean\ wage} = \mathsf{\$10.33\ p/hr}$$

Other rules linked Education (as in years of formal education) to Wage and justify the argument that on average, education improves earning power:

$$\begin{aligned} \mathsf{Education} \in [2,13]\ \mathsf{years} \quad &\Rightarrow \quad \mathsf{mean\ wage} = \mathsf{\$7.52\ p/hr} \\ \mathsf{Education} \in [14,18]\ \mathsf{years} \quad &\Rightarrow \quad \mathsf{mean\ wage} = \mathsf{\$11.64\ p/hr} \end{aligned}$$

Rules may also have multiple attributes on both sides. For example, the following rule shows that being a male union member correlates with higher wages despite having less than average education (the average education is 13 years):

$$\text{Sex} = \text{male and Union=Yes} \quad \Rightarrow \quad \text{mean wage} = \$11.29 \text{ p/hr, mean education} = 12.66 \text{ years}$$

In a different and somewhat unexpected direction, a rule connecting education to age was found. The mean age of the population was 37 years.

$$\text{Education} \in [3,10] \text{ years} \quad \Rightarrow \quad \text{mean age} = 46 \text{ years old}$$

This rule shows us that those with very little schooling are on average far older, a sign of positive progress in society. Usually, most research on this database would be limited to factors affecting a person's wage. Through our data-mining technique we exposed interesting information which we would not initially have thought to look for.

In summary, the *Wages* database provided us with interesting results. However, these results only provide us with a flavor of the rules. A rigorous evaluation of the quality of rules mined according to our definition is provided in Section 4.

1.4 Related Work

Association rules in the quantitative case were first considered in the work of Piatetsky-Shapiro [11]. Srikant and Agrawal [12] provided the first general definition and algorithm for the quantitative case. They provide a definition based on considering intervals of quantitative attributes and then present an algorithm that uses discretization of the quantitative data. A detailed comparison with [12] is provided in Section 4.2.

Fukuda et al. [5] also consider the problem of quantitative association rules, providing efficient algorithms for finding association rules with maximum confidence or support. Using computational geometry and sampling methods, [5] provide highly efficient algorithms for these problems, even for very large databases. However, [5] focuses on the case where the right-hand side of the rule is a categorical attribute. Our work, in contrast, focuses on the case where the right-hand side of the rule is numerical. It should be noted that our notion of a *maximal* association rule (Section 2.1.2) is similar to that of an *optimized support rule* of [5]. Thus, the generalized algorithm for finding *optimized support rules*, as described in [5, Section 4.2], can be adapted to finding maximal quantitative-to-quantitative association rules, as defined in this paper. However, the [5] algorithm does not ensure that the rules are *irreducible* (see Section 2.1.2). We note that for large databases, where sorting may be expensive, our algorithm should be augmented with a sampling and bucketing procedure, as in [5]. Of relevance is also the work of Morimoto et al. [10] on the construction of regression trees. There, the authors provide elegant algorithms for finding optimal region splitting in regression trees. Region splitting finds a subregion with significant difference in mean, though this region is not necessarily irreducible, and the algorithms of [10] are not designed for finding *all* such regions.

Other work on quantitative association rules includes Zhang et al. in [15] who use clustering methods to improve the partitioning (or discretization) of the quantitative attributes in the [12] The case where there are *two* numerical attributes on the right-hand side is considered by Fukuda et al [4, 5] and Yoda et al. [14].

The explora project headed by Kloesgen, see [7], considered searching for statistically interesting phenomena much in the way we do. Although they do not do this in the framework of association rules, there is much common motivation. We note finally that a variation on categorical association rules was introduced by Brin et al. in [3]. Their new definition is based on relating to associations as statistically interesting correlations. Theirs is the first work to integrate statistical inference methods in order to confirm the validity of the mined association rules.

1.5 Outline

In the next section we present formal definitions of two important types of quantitative association rules. We also motivate and define the concept of sub-rules. Next, in Section 3 we present efficient algorithms for some cases of quantitative rules and demonstrate the computational viability of mining for these rules. In Section 4 we present experimental evaluation and analysis of results obtained, including comparisons to [12]. We use a real-world database and present an evaluations by a domain expert who provided us with the database and used the results for his research. We believe that our evaluation is of independent interest as it takes first steps in developing a rigorous methodology for checking whether new data-mining tools are actually helpful to end users. Finally, in Section 5 we discuss the problem of multiple comparison procedures and suggest initial steps to overcome this problem.

2 Definitions

An association rule contains a left-hand side and a right-hand side. In the most general form, the left-hand side of the rule is a description of a subset of the database, while the right-hand side provides a description of an outstanding behavior of this subset. This general structure gives rise to many different concrete rule types, determined by the type of subset used on the left-hand side, and the description used for the right-hand side. In this paper we focus on two specific types, which we found to be most useful in practice, and algorithmically manageable. The two are: Categorical to Quantitative rules with an unlimited number of attributes on each side, and Quantitative to Quantitative where both sides contain a single attribute only. (By "Categorical to Quantitative", we mean that the left-hand side is comprised of categorical attributes whereas the right-hand side is comprise of quantitative attributes.) We also provide a general definition of sub-rules, which, as we shall see, are essential for providing both comprehensive and exact information.

Notations. Let $E = \{e_1, ..., e_m\}$ be the set of attributes (or fields) for a database $\mathcal{D}$ of transactions. Let $E_Q \subseteq E$ be the set of quantitative attributes, $E_C \subseteq E$ the set of categorical attributes, and C the set of all possible categorical values. Each transaction in $\mathcal{D}$ is a set $t = \{\langle e_1, v_1\rangle, \ldots, \langle e_m, v_m\rangle\}$ of m attributes and corresponding values (i.e., for each i, if e_i is categorical then $v_i \in C$, and if e_i is quantitative then $v_i \in \mathbb{R}$.

2.1 Rule Types

2.1.1 Categorical $\Rightarrow$ Quantitative Rules

The first type of rule we consider are rules where the left-hand side is a set of categorical attributes, and the right-hand side is a vector of mean values for some set of quantitative attributes.

The left-hand side. The left-hand side of a rule is a set $X \subseteq E_C \times C$ of categorical attributes and matching categorical values. The set X, which we call *the profile*, defines a subset of the database. For a transaction $t = \{\langle e_1, v_1\rangle, \ldots, \langle e_m, v_m\rangle\}$, we say that t *has profile* X if $X \subseteq t$, i.e. t coincides with X whenever X is defined. We denote the set of transactions with profile X by T_X.

The right-hand side. The right-hand side of a rule consists of a vector of mean values for a set of quantitative attributes, with the mean taken over the transactions which match the profile of the left-hand side. Formally, for a set of quantitative attributes J, and a set of transactions $T \subseteq \mathcal{D}$, we denote by $\mathsf{Mean}_J(T)$ the vector of mean values of the attributes in J for the set T. The right-hand side of the rule is $\mathsf{Mean}_J(T_X)$ for some $J \subseteq E_Q$.

Significance. The rule points to extraordinary behavior if the mean of the attributes in J over the transactions in T_X is significantly different from the rest. We therefore compare the mean in T_X to the mean of the complement $\mathcal{D} - T_X$. Note, however, that although the two means may be numerically different in the database, we may not have sufficient statistical evidence to infer a difference in the real populations. Thus, we use statistical tests to establish the *significance level* of the difference. Specifically, we use the standard Z-test to establish significance of the inequality of the means. We test the hypothesis that the mean of the two subsets are equal (the null hypothesis) against the hypothesis claiming a difference exists. A rule is considered *significant* if the null hypothesis is rejected with confidence above a set threshold (usually set at 95%). Formally, we say that $\mathsf{Mean}_J(T_X)$ is *significantly different from* $\mathsf{Mean}_J(T_Y)$, denoted $\mathsf{Mean}_J(T_X) \not\approx \mathsf{Mean}_J(T_Y)$, if we can statistically infer that for every $e \in J$ the means of attribute e in T_X and T_Y are different. We are now ready to define mean-based categorical to quantitative association rules.

Definition 1 *A (mean-based)* categorical to quantitative association rule *is a rule of the form* $X \Rightarrow \mathsf{Mean}_J(T_X)$*, where* X *is a profile of categorical attributes* ($X \subseteq E_C \times C$)*,* J *is a set of quantitative attributes* ($J \subseteq E_Q$)*, and* $\mathsf{Mean}_J(T_X) \not\approx \mathsf{Mean}_J(\mathcal{D} - T_X)$*.*

Minimum difference. Sometimes, finding a population for which the mean is merely different does not lead to what the user may consider interesting information. If we were to discover, for example, a group of people with life expectancy three days more than the overall population, it may not be of interest to us even if it passes a statistical test. We therefore allow a user-defined minimum difference parameter, denoted *mindif*. In this case we write $\mathsf{Mean}_J(T_X) \not\approx \mathsf{Mean}_J(T_Y)$ if for every $e \in J$ there is statistical evidence for inferring that $|\mathsf{Mean}_e(T_X) - \mathsf{Mean}_e(T_Y)| > mindif$.

We stress that this parameter is *not* introduced for efficiency reasons, as a way to filter out rules without using a statistical test. Rather, it is used in order to allow the user to prevent the system from outputting rules that are of no interest to the him/her (as in the above example of a group of people with average life expectance 3 days more than the overall population). We stress that if the *mindif* parameter is used, then then some statistically sound and significant rules that do not pass the *mindif* test may be eliminated.

Categorical $\Rightarrow$ Quantitative rules based on other distribution measures. The rules defined above provide a tool for discovering extraordinary behavior of the distribution with regards to its mean value. An analogous definition can be provided using any other measure of the distribution (e.g., variance or median). For a given measure M (e.g., $M = \mathsf{Variance}$), an *M-based association rule* is of the form $X \Rightarrow M_J(T_X)$. The rest of the definitions carry over directly from the mean-based rules by changing Mean with M throughout. Naturally, when we find a difference in the given measure we check the *significance* of the difference using an appropriate test for this measure (e.g., the F-test for variance). The algorithm presented in section 3.2 is correct for any measure and we have implemented the algorithm for variance-based rules as well.

2.1.2 Quantitative $\Rightarrow$ Quantitative Rules

Next, we consider rules for which both the left-hand side and the right-hand side are comprised of a *single* quantitative attribute.

The **left-hand side** of the rule is a triplet $X = (e, r_1, r_2)$, where e is a quantitative attribute, and r_1, r_2 are real values, $r_1 \leq r_2$. We call X the *profile*. We say that a transaction t has profile X if the value of t for the attribute e is within the range $[r_1, r_2]$. We denote by T_X the set of transactions in the database with profile X. The **right-hand side** of the rule is a quantitative attribute j ($j \neq e$) together with its mean value $\mathsf{Mean}_j(T_X)$. As before, **significance** is ensured by demanding that $\mathsf{Mean}_j(T_X) \not\approx \mathsf{Mean}_j(\mathcal{D} - T_X)$.

At this stage we would like to simply define a rule as one of the form $X \Rightarrow \mathsf{Mean}_j(T_X)$ where $\mathsf{Mean}_j(T_X) \not\approx \mathsf{Mean}_j(\mathcal{D} - T_X)$. However, not all rules of this type are desirable, as we now show. Consider the following (fictitious) database. Assume that the average weight of the entire population is 80 kilograms.

Age	Weight
...	...
50	80
60	90
70	90
80	90

The following rules may all be deduced from the above database.

$$\begin{array}{l} (1)\ \mathsf{age} \in [60,80] \Rightarrow \mathsf{average} = 90.0 \\ (2)\ \mathsf{age} \in [70,80] \Rightarrow \mathsf{average} = 90.0 \\ (3)\ \mathsf{age} \in [50,80] \Rightarrow \mathsf{average} = 87.5 \end{array}$$

Intuitively, it is clear that the first rule is the most and only desirable rule of the three. The second rule is essentially "contained" in the first rule, adding no new information and is therefore superfluous. Presenting only widest possible rules allows us to provide a succinct description of the extraordinary phenomena.

The third rule is even wider than the first one and also may be significantly different from the overall average. Nevertheless, it is undesirable. This is because the rule does not focus exclusively on the population with the extraordinary behavior. In fact, it is obtained by appending an adjacent, irrelevant region to that of the first rule. Furthermore, the rule is *misleading* because it leads one to think that the phenomenon of being overweight begins at 50 rather than 60. (We note that we deal with value fluctuations *inside* a rule by presenting sub-rules, as will be discussed below.) These examples motivate the following formal definitions.

Irreducible and maximal rules. Consider a rule $X \Rightarrow \mathsf{Mean}_j(T_X)$ with $X = (e, a, b)$, and suppose that $\mathsf{Mean}_j(T_X)$ is above average. We now define the notions *irreducibility* and *maximality*. We provide the definition for above average rules. The definitions for below-average rules are analogous. Intuitively, the rule $X \Rightarrow \mathsf{Mean}_j(T_X)$ is *irreducible* if one cannot cut the interval $[a, b]$ into two adjacent parts and obtain an interval which is not above average. This property ensures that we do not have irrelevant regions appended to the edge of the rule (Rule (3) above is reducible). Formally, we say that the rule is *irreducible* if for any $a < c < b$, setting $Y = (e, a, c)$, and $Z = (e, c, b)$, then both $\mathsf{Mean}_j(T_Y)$ and $\mathsf{Mean}_j(T_Z)$ are above average.

The rule is *maximal* if we cannot enlarge the interval $[a, b]$ either to the right or the left and still remain with an irreducible rule with above average distribution. Formally, the rule is *maximal* if for any $c < a$ (respectively, $c > b$) setting $Y = (e, c, b)$ (respectively, $Y = (e, a, c)$) yields that $\mathsf{Mean}_j(T_Y)$ is either not above average or the interval Y is reducible. Maximal rules are therefore the largest "good" rules and provide the most concise presentation (the second rule in the above example is not maximal).

We now have all the necessary concepts to define quantitative to quantitative rules:

Definition 2 *A (mean-based)* quantitative to quantitative association rule *is a maximal and irreducible rule of the form* $X \Rightarrow \mathsf{Mean}_j(T_X)$ *where* X *is a profile for a single quantitative attribute* $(X \in E_Q \setminus \{j\} \times \mathbb{R} \times \mathbb{R})$*,* j *is a quantitative attribute* $(j \in E_Q)$*, and* $\mathsf{Mean}_j(T_X) \not\approx \mathsf{Mean}_j(\mathcal{D} - T_X)$*.*

Remark. Note that we have used a *range* to define the profile in the rule. This seems to run counter to our argument in the introduction, that range is not a good statistical measure for association rules. However, our claim was only with regards to the right-hand side of the rule (i.e., the behavior) and not for the left-hand side (i.e., the profile). The reason for this difference is as follows. The profile classifies the transactions for which the rule applies. For determining this a range is best, as it provides a clear indication where the rule applies and where not. For example, by saying that a phenomenon occurs for people between the ages of 10 and 20, it is clear to whom it applies. If we were to say that the phenomenon occurs for people of average age 15, then it would not be clear exactly what population this rule is based on and where it applies.

2.2 Sub-Rules

So far, we provided a framework for defining rules, and definitions for two important types of rules. However, not all rules are desired. In presenting association rules we are interested in finding the *key factors* of extraordinary behavior in the database. This issue of sub-rules is applicable to any type of association rules, not just those presented here. The best way to motivate this notion is by example. Therefore, consider the following set of (fictitious) rules, where the overall life expectancy is 70 years:

$$
\begin{array}{ll}
\text{(1) smoker} \Rightarrow & \text{life expectancy} = 60 \\
\text{(2) male and smoker} \Rightarrow & \text{life expectancy} = 60 \\
\text{(3) smoker and wine-drinker} \Rightarrow & \text{life expectancy} = 70
\end{array}
$$

Both the second and third rules are more specific than the first rule and are therefore *contained* in it. Intuitively, it is clear to us that the second rule is undesirable as it introduces a factor that does not contribute to the extraordinary behavior. In fact, it misleads us into thinking that being male is a contributing factor to the below-average life expectancy.

On the other hand, although the life expectancy in the third rule is *exactly average*, this rule is (subjectively) interesting when viewed in light of rule (1). This is because it shows that there is a subset of the smokers (i.e., wine drinkers) whose life expectancy is higher than that of other smokers. Note that when viewed in light of the entire population this rule does not seem to point to extraordinary behavior, but clearly does so when viewed in the context of the smoker population. Therefore, the third rule qualifies as a *sub-rule* of the first rule (as long as the difference passes the statistical tests). We now formalize these ideas.

The intuition behind rule containment is that a rule with profile X is contained in a rule with profile Y, if $T_X \subset T_Y$ (that is, Y "covers" a larger set of transactions including those covered by X). Formally, *rule containment* is defined as follows:

- **Categorical attributes:** Let X and Y be profiles containing categorical attributes only (as in 2.1.1). Then we say that rule $Y \Rightarrow \mathsf{Mean}_J(T_Y)$ contains rule $X \Rightarrow \mathsf{Mean}_J(T_X)$, if $Y \subseteq X$.

- **Quantitative attributes:** Let $X = (e, c, d)$ and $Y = (e, a, b)$, then we say that the rule $Y \Rightarrow \mathsf{Mean}_J(T_Y)$ (as in 2.1.2), contains $X \Rightarrow \mathsf{Mean}_J(T_X)$ if $a \leq c \leq d \leq b$.

We now define basic rules, sub-rules and basic sub-rules. Loosely speaking, a basic rule is a rule that is not contained in any other rule. A sub-rule is a rule that is contained in some super-rule and displays extraordinary behavior with respect to the super-rule. However, there may be many sub-rules of some rule. This leads us to the definition of a basic sub-rule, which is a sub-rule that is not contained in any other sub-rule. A basic sub-rule is the "most comprehensive" sub-rule. The formal definition follows.

Definition 3 (basic rules and sub-rules):

1. *A rule is a* basic rule *if it is not contained in any other rule.*
2. *A rule* $X \Rightarrow \mathsf{Mean}_J(T_X)$ *is a* sub-rule *of* $Y \Rightarrow \mathsf{Mean}_J(T_Y)$ *if:*
   (a) $Y \Rightarrow \mathsf{Mean}_J(T_Y)$ *contains* $X \Rightarrow \mathsf{Mean}_J(T_X)$
   (b) $\mathsf{Mean}_J(T_X) \not\approx \mathsf{Mean}_J(T_Y - T_X)$

A sub-rule $X \Rightarrow \mathsf{Mean}_J(T_X)$ *is a* basic sub-rule *if it is not contained in any other sub-rule of* $Y \Rightarrow \mathsf{Mean}_J(T_Y)$.

Note that a basic sub-rule is a basic rule with regards to the set of transactions T_Y. That is, view the transactions in T_Y as an independent database. Then, any basic rule in the database T_Y is a basic sub-rule of $Y \Rightarrow \mathsf{Mean}_J(T_Y)$.

As we have shown, contained rules which are not sub-rules (i.e., their mean value is not significantly different from that of the super-rule) are undesirable. In the above example, the second rule is contained yet does not have a different mean (condition 2(b) of the definition is not met), whereas the third rule is a sub-rule exactly because of its different mean with respect to the super-rule. We therefore wish to find all basic rules, their basic sub-rules, the basic sub-rules of these sub-rules, and so on.

Definition 4 *We recursively define* desired rules (*those which we wish to obtain*)*:*

1. *Any basic rule is* desired.
2. *Any basic sub-rule of a desired rule is also* desired.

In Section 3 we provide algorithms to find all desired rules.

2.3 Classical Association Rules

Our definition of association rules is actually a generalization of the classical definition of categorical association rules. In the categorical case, the left-hand and right-hand sides are defined by lists of items X and Y. The measures of significance used by [2] are support and confidence. Since the appearance of a set of items is a Bernoulli random variable, the mean of "Y given X" is exactly the confidence of the rule. The rules defined by [3] are defined in the same way with a different significance measure (a statistical χ^2 test is used).

3 Algorithms for Finding Rules

Efficient algorithms for finding quantitative association rules are provided for two types of rules:

1. **Quantitative to Quantitative:** $X \Rightarrow \mathsf{Mean}_J(T_X)$ where both X and J contain a single quantitative attribute only.

2. **Categorical to Quantitative:** $X \Rightarrow M_J(T_X)$ where $X \subseteq E_C$ (only categorical attributes) and $J \subseteq E_Q$ (only quantitative attributes). There is no limit on the number of attributes in X or J. The algorithm is correct for any measure M.

3.1 Quantitative to Quantitative

Our algorithm finds rules between any two given quantitative attributes. The algorithm is applied to every pair, thereby obtaining all rules of this type. The algorithm allows the user to specify a minimum support parameter, though it is not required (in which case the minimum support is set to zero).

Algorithm overview. Let x and y be a pair of quantitative attributes, and suppose we are searching for rules with x on the left-hand side and y on the right-hand side. Note that if we sort the database by attribute x then any contiguous set of transactions for which the average of the y attribute is above or below average constitutes a rule from x to y (provided it passes the necessary statistical-significance test). This is because attribute x is sorted, and therefore any contiguous region defines a range of x values. Thus, it only remains to ensure that the rule is irreducible and maximal. To this end, we use the *Window* algorithm, described below. In a single pass over the sorted array of transactions, the *Window* algorithm finds all maximal irreducible rules from x to y. The algorithm is based on the following simple fact: if the regions $[a, b]$ and $[b, c]$ are both above or below average, then so too is the region $[a, c]$. The search for above and below average regions is symmetrical. We will therefore concentrate only on above-average rules from now on, and describe the *Window-above* procedure. The *Window-below* procedure is analogous. Further note that when we say "above-average" we mean above the overall mean plus *mindif*.

Notations.

- For an array $\mathcal{D}$ of transactions, and indexes i, j, we denote by $\mathcal{D}[i \ldots j]$ the sub-array from $\mathcal{D}[i]$ to $\mathcal{D}[j]$, inclusive.

- For a transaction t and attribute y, we denote by $t.y$ the value of t for the attribute y. Thus, $\mathcal{D}[i].y$ is the y value of the i-th entry of D.

- For $\mathcal{D}, i, j, y$ as above, we denote by $\text{Average}(\mathcal{D}[i].y \ldots \mathcal{D}[j].y)$ the average of the values $\{\mathcal{D}[i].y, \ldots, \mathcal{D}[j].y\}$.

- For $\mathcal{D}, i, j, y$ as above, and constant μ, we denote by $\text{Z-test}_{>}(\mathcal{D}, i, j, y, \mu)$ the procedure which runs a Z-test to check if there is statistical evidence to infer that $\text{Average}(\mathcal{D}[i].y \ldots \mathcal{D}[j].y)$ is greater than μ. Similarly, $\text{Z-test}_{<}(\mathcal{D}, i, j, y, \mu)$ determines whether there is statistical evidence to infer that $\text{Average}(\mathcal{D}[i].y \ldots \mathcal{D}[j].y)$ is less than μ.

Input: An array D of transactions, attributes x and y, and a value *mindif*. D is sorted according to attribute x.
Output: All association rules from x to y

`Window`($\mathcal{D}$, x, y, *mindif*)
    `Window-above`($\mathcal{D}$, x, y, *mindif*)
    `Window-below`($\mathcal{D}$, x, y, *mindif*)

`Window-above`($\mathcal{D}$, x, y, *mindif*)
1 $\mathit{Last} \leftarrow$ Last Index of the array $\mathcal{D}$
2 $\mathit{current} \leftarrow$ first index of the array $\mathcal{D}$
3 $\mu \leftarrow$ Average($\mathcal{D}[\mathit{current}].y \ldots \mathcal{D}[\mathit{Last}].y$) + mindif
4 While ($\mathit{current} \leq \mathit{Last}$)
5 {
6   While ($\mathcal{D}[\mathit{current}].y < \mu$ & $\mathit{current} \leq \mathit{Last}$)   // find above average entry
7         $\mathit{current} \leftarrow \mathit{current} + 1$
8     $a \leftarrow \mathit{current}$; $b \leftarrow a$;   // $A \leftarrow \mathcal{D}[a]; B \leftarrow \emptyset$
9   While (Average($\mathcal{D}[a].y \ldots \mathcal{D}[\mathit{current}].y$) $\geq \mu$ & $\mathit{current} \leq \mathit{Last}$)
10   {
11         $\mathit{current} \leftarrow \mathit{current} + 1$ // enlarge B by one
12   if (Average($\mathcal{D}[b].y \ldots \mathcal{D}[\mathit{current}].y$) $\geq \mu$)   // B is above average
13             $b \leftarrow \mathit{current}$ // add B to A
14   }
15   Run Z-test$_{>}(\mathcal{D}, a, b, y, \mu)$.
16   If Z-test returns 'yes' then
17   {
18   Output the association rule
19             $x \in [\mathcal{D}[a].x, \mathcal{D}[b].x] \Rightarrow$ Average for y is Average($\mathcal{D}[a].y \ldots \mathcal{D}[b].y$))
20   `Window`($\mathcal{D}[a \ldots b]$, x, y,*mindif*)   // recursive call
21   }
22     $\mathit{current} \leftarrow b + 1$ // continue with next value after A
23 }

`Window-below`($\mathcal{D}$, x, y, *mindif*)
    Identical to `Window-above`, except reverse appropriate inequalities.

Figure 1: Window Procedure for finding "Quantitative $\Rightarrow$ Quantitative" Rules

The *Window* procedure. The data-driven *Window* procedure accepts as input a pair of attributes, x and y, an array $\mathcal{D}$ of transactions, and a value for *mindif*. The input array is sorted by the attribute x. We execute a single pass to find all rules from x to y. From now on, all references to above-average, below-average, irreducible and maximal are with regards to the y value of the transactions. We now describe the procedure. A detailed pseudo-code description is presented in Figure 1. First we define μ to be the average of the value y for the entire array $\mathcal{D}$, plus the user defined *mindif* (line 3). The value μ is the threshold for an "above average" rule. (From here and on "above average" shall mean "above μ".) The procedure keeps three pointers/indexes into the array $\mathcal{D}$: a, b and *current*, with $a \leq b \leq \mathit{current}$. These pointers define two adjacent regions (or "windows"): $\mathcal{D}[a \ldots b]$ and $\mathcal{D}[b+1 \ldots \mathit{current}]$, which we denote by A and B, respectively. We shall prove that the algorithm guarantees that:

- A is always an irreducible above-average region.

- B is an adjacent region to A, such that $A \cup B$ is above average.
- B is joined to A if $A \cup B$ is also irreducible.

To begin, we initialize A to the first above-average value in $\mathcal{D}$ and B is empty (lines 6-8). Given A and B, we advance the pointer *current* by one, thus adding a new entry to B (line 11). There are three possibilities at this stage:

1. The resulting B region is above-average. In this case, $A \cup B$ is irreducible, and we join B to A (and empty B) (lines 12-13).
2. The average of $A \cup B$ is above average, but not that of B alone (the test in line 9 succeeds but the test in line 12 fails). In this case, we continue and enlarge B (i.e. advance *current*).
3. The resulting $A \cup B$ region is not above-average (the test in line 9 fails). In this case, A and B will never be joined (we shall prove that $A \cup B$ cannot be part of an irreducible rule). A alone is a potential rule.

When a potential rule is identified we test the statistical significance of the rule (line 15). If the rule passes the test, we output the rule (lines 18-19), and recursively call the Window procedure to search for sub-rules (line 20). After finding all sub-rules, we continue advancing the counter for the first entry after A, and start the procedure again.

Recursive calls to ***Window*****.** Given a rule in a region A we call *Window* recursively to find all sub-rules. The input to the recursive call is the sub-array defined by the region A. Thus, in the recursive call, the region A is treated as the entire database of transactions.

Theorem 5 (Soundness): *All rules in the output of the* Window *algorithm are rules as in Definition 2.*

Proof: We first prove the following invariant: at every stage, A is an irreducible above-average region. Remember that a region is irreducible if any partition of it into two parts results in two above-average regions.

The proof is by induction on construction of A. Initially, A is initialized to a single above-average value, which is trivially irreducible. The size of A is increased by appending B to A, in case B is also above average (line 13). We show that after this occurs, A remains irreducible. Remember that B is joined to A only if B is an above average region. Let A_{new} be the new A region created in line 13, and A_{old} the former A region. Thus, $A_{\text{new}} = A_{\text{old}} \cup B$. There are three possible partitions of A_{new} into two regions. We show that in each partition type, both regions are above average. The three partitions are:

1. Divide A_{new} into A_{old} and B: by the inductive step, A_{old} is above-average and by the condition of the algorithm, so is B.
2. Divide A_{new} into A_1 and $A_2 \cup B$ (where A_1 and A_2 are a partitioning of A_{old}): A_{old} is irreducible and therefore A_1 must be above-average. The same is true for A_2, and by the algorithm so is B. Therefore $A_2 \cup B$ is also above-average.
3. Divide A_{new} into $A_{\text{old}} \cup B_1$ and B_2 (where B_1 and B_2 are a partitioning of B): Assume by contradiction that $A_{\text{old}} \cup B_1$ is not above-average. Then, at a previous stage in the algorithm when the test in line 9 was applied to $A_{\text{old}} \cup B_1$ the test would fail. In this case, the loop

would end, and we would not get to the stage the algorithm is in. Thus, $A_{\text{old}} \cup B_1$ is above-average. Similarly, if B_2 is not above-average then B_1 must be above average (because, by the algorithm, all of B is above-average). Thus, by line 12, A_{old} would be already joined with B_1.

We have therefore proved the invariant. Thus, A is always irreducible. Next, we show that the region A, for rules outputted is also *maximal.* A rule is accepted as potential in line 15. This occurs when the average of the region $A \cup B$ is not above-average or when $current = Last$. Consider the following possible expansion of A:

1. A is expanded past B (this cannot happen if $current = Last$): In this case the resulting region is reducible. This is because $A \cup B$ is not above average. Thus, it can omitted from the region.

2. A is expanded to include $B_1 \subset B$: In this case B_1 must be below average. Otherwise, B_1 would have already been joined to A in a previous stage (lines 12-13). Thus $A \cup B_1$ is reducible.

3. A is expanded "backwards" into a region before A. Denote by P the expansion. For A to be expanded to $P \cup A$, this region must be irreducible. Hence, P is above average and P must be have been identified as an above average region prior to considering A. By arguments (1)-(2) above, if the rule $P \cup A$ was not considered by the algorithm, then extending P forward (to include A) cannot yield an irreducible region. Thus, $P \cup A$ cannot be irreducible above average.

Thus, we have shown that A is irreducible and maximal. By line 15, the rule also passes a statistical test. Thus, the rule outputted in line 19 is indeed a rule as in Definition 2. ■

Theorem 6 (Completeness): *The* Window *algorithm finds all association rules, as in definition 4.*

Proof: We first show that *Window* finds all maximal irreducible rules on the first level (i.e., before any recursive calls). Let A be a maximal irreducible rule. Since A is maximal, the region P preceding (and adjacent to) A must be below average. Furthermore, it belongs to no other maximal irreducible region preceding A. Otherwise, the union of this region with A would also be irreducible, contradicting the maximality of A. The P region is therefore "skipped" in the iteration of lines 6-7 and at line 8, the variable *current* points to the first index of A.

A is irreducible and therefore there is no partition of A into A_1 and A_2 so that A_1 is below average. Therefore, while *current* remains in the A region (and considering $(\mathcal{D}[a].y \ldots \mathcal{D}[current].y)$ as the A_1 region), the test on line 9 always succeeds. That is, the Z-test on line 15 is not executed until *current* points beyond the last index of A. We now claim that b must point to the last index of A when line 15 of the algorithm is reached. This can be seen by noticing that since A is irreducible, for every partitioning of A into A_1 and B, B is above average. Therefore, when *current* points to the last index of A, the check in line 12 must succeed and b is updated to equal *current*. The fact that b cannot point *past* the end of A is covered by the previous theorem proving soundness.

We have so far shown that all basic rules are found. The fact that all basic sub-rules are also found is due to the recursive nature of *Window.* That is, since basic sub-rules are just basic rules with respect to the restricted database of the super-rule, *Window* finds these rules upon the recursive calls. We note that in this scenario, all sub-rules found are clearly basic. We conclude that *Window* finds all desired rules. ■

Complexity analysis. For a given pair of attributes, the time taken for n transactions is $O(n \log n)$ for the sort, plus the complexity of the Window algorithm. The complexity of Window is clearly upper-bound by $O(n)$ times the number of levels of rules (i.e. the number of recursive calls). Since the number of levels is expected to be low (as experience has shown), we effectively maintain linear complexity. Note that the minimum support has no effect on the running time, enabling us to find rules with low support. For k quantitative attributes, the time taken to find all rules of this type is therefore $O(k \cdot n \log n + k^2 \cdot n)$. We note that with very large databases, the sort may take considerably longer as it needs to be executed in secondary memory.

3.2 Finding Rules from Categorical to Quantitative Attributes

For this type of rule, we are looking for rules where the profile is comprised of categorical items only. A set of categorical items is called a *frequent set* if more than *minsup* (a user-defined parameter) transactions support it. Therefore each possible profile is a frequent set and we begin by finding all frequent sets on the categorical attributes only, using known algorithms. Since the major bottleneck in finding frequent sets is the number of attributes, the fact that we find sets only on the categorical attributes is a considerable advantage.

The set of frequent sets contains all potential rule profiles and is the basis for our search. For a given frequent set X, we calculate $M_i(T_X)$ for any measure M and any quantitative attribute i. Then if $M_i(T_X) \not\approx M_i(\mathcal{D} - T_X)$ and the rule $X \Rightarrow M_i(T_X)$ is a basic rule or sub-rule, we add it to the set of results. Notice that the algorithm is the same for *any* distribution measure.

Algorithm outline. The algorithm has three distinct stages:

1. Find all frequent sets of categorical items only, using known algorithms such as Apriori (see [2]).

2. For all quantitative attributes, calculate the distribution measure/s (e.g. mean and variance) for each frequent set, using the hash-tree data structure presented in [2]. One pass over the database is sufficient.

3. Find all basic rules and sub-rules. For every frequent set X and quantitative attribute e, it remains to check if $X \Rightarrow \mathsf{Mean}_e(T_X)$ and $X \Rightarrow Variance_e(T_X)$ are basic rules or sub-rules or neither. We do this by traversing a lattice of the frequent sets while keeping track of containment relations between sets and the sub-rule hierarchy.

Stage 1: Finding all frequent sets. Known algorithms, such as Apriori (see [2]) are used for this stage. Sampling may be used for very large databases.

Stage 2: Calculating the distribution measure. The distribution measure of each quantitative attribute for each frequent set is calculated using the hash-tree data structure presented in [2]. This structure enables us to efficiently reach all frequent sets supported by a given transaction t. For each frequent set supported by t and for every quantitative attribute, the values necessary for calculating the mean and variance are updated. Clearly, a single pass over the database is enough to calculate the distribution of every frequent set.

Stage 3: Finding all basic rules and sub-rules. We first show how this is done for rules with a *single* quantitative attribute in the right-hand side. We then show how to combine the rules together to obtain the general case.

Now that the distribution of every quantitative attribute e for every frequent set X has been calculated, it remains to check if $X \Rightarrow M_e(T_X)$ is a basic rule or sub-rule or neither. In order to do this we build a lattice, $G = (V, E)$, (a directed acyclic graph) of the frequent sets, see Figure 2. Nodes of G are frequent sets and edges are (X, Y) where $X \subseteq Y$ and $|Y| = |X| + 1$. The lattice is made up of levels where level j contains all the frequent sets containing j items (level 0 is the top level and contains the empty frequent set).

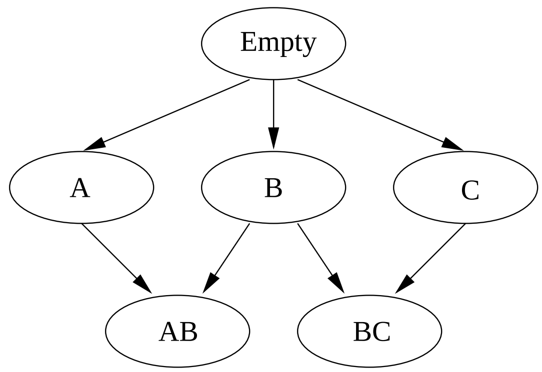

Figure 2: Lattice of frequent sets

A *son* of a frequent set X is a frequent set immediately below it in the lattice (similarly a *father*). Note that any set X is a subset of all its descendants. In each node we store the distribution of the appropriate frequent set (already calculated). The lattice is easily built given all the frequent sets.

We now present an algorithm, which given the above lattice with the calculated distributions, finds all desired rules. The algorithm works by finding candidate sets Y for every frequent set X such that $X \Rightarrow M_i(T_X)$ is potentially a *basic* sub-rule of $Y \Rightarrow M_i(T_Y)$. (Recall, that a basic sub-rule is one that is not contained in any other sub-rule of $Y \Rightarrow M_i(T_Y)$.) Since the frequent set X is identified with the rule $X \Rightarrow M_i(T_X)$, from here on we also refer to X as a rule.

Before describing the algorithm, we explain the necessary conditions for sets Y which can be super-rules of X:

- X is a sub-rule of Y only if $Y \subset X$.
- X is a sub-rule of Y only if Y is a basic rule or sub-rule.
- X is a basic sub-rule of Y only if there is no Z such that Z is a sub-rule of Y and X is contained in Z.

We therefore conclude that we need only consider ancestors of X in the lattice that are basic rules or sub-rules and for which there is no other sub-rule on the *path* from Y to X in the lattice.

In order to identify the candidate super-rules for X, the algorithm maintains a number of sets. We now list these sets and show how they can be derived from the sets of X's already processed fathers in the lattice. This yields an algorithm which traverses the lattice top-down in a breadth-first fashion.

- *Contained(X):* this set is comprised of all the basic rules and sub-rules that contain X, yet *cannot* be super-rules of X. The reason that they cannot be super-rules is because they have sub-rules that contain X and so X cannot be a *basic* sub-rule.

  Let Y be a father of X in the lattice. Since X is contained in Y, any rule belonging to $Contained(Y)$ must also belong to $Contained(X)$. Furthermore, all super-rules of Y also belong to $Contained(X)$.

- *Cand(X):* this set contains all the candidate super-rules for X. That is, $Cand(X)$ contains all the basic rules and sub-rules that are X's ancestors and do *not* appear in $Contained(X)$.

  Let Y be a father of X in the lattice and let $CandReject(Y)$ equal the candidate super-rules of Y who were rejected. Consider now the union of $CandReject(Y)$ for every father Y of X. All of the sets in this union that do *not* appear in $Contained(X)$ are potential candidates. This is because they are rules which do not have sub-rules containing Y and therefore X. In addition, if Y itself is a basic sub-rule, then Y is also a candidate super-rule of X.

- *Sub-Rules(X):* this is the set of all X's basic sub-rules. Therefore, if $Z \in Sub-Rules(X)$ then $Z \Rightarrow M_i(T_Z)$ is a basic sub-rule of $X \Rightarrow M_i(T_X)$. This set is needed for storing the output.

Initialization: Let *top* be the top node of the lattice (the empty frequent set).
$Cand(top) \leftarrow \{top\}$
$Contained(top) \leftarrow \phi$

We execute a breadth-first search of the lattice G (top-down).
For each node X (in BFS-order), execute the following:

1 $Contained(X) \leftarrow \bigcup_{Y \ is \ a \ father \ to \ X} Contained(Y)$
2 $Cand(X) \leftarrow \bigcup_{Y \ is \ a \ father \ to \ X} Cand(Y) - Contained(X)$

3 For every $Z \in Cand(X)$ do
4 If $M_i(T_X) \not\approx M_i(T_Z - T_X)$
5 *Sub-Rules(Z)* $\leftarrow$ *Sub-Rules(Y)* $\cup \{X\}$
6 $Contained(X) \leftarrow Contained(X) \cup \{Z\}$
7 $Cand(X) \leftarrow Cand(X) - \{Z\}$
8 $Cand(X) \leftarrow Cand(X) \cup \{X\}$

Figure 3: Algorithm for finding all "Categorical $\Rightarrow$ Categorical" Rules

The algorithm is presented in Figure 3 and works by iterating over every $Z \in Cand(X)$ and checking whether X is a sub-rule of Z. We now explain the algorithm.

As previously described, $Contained(X)$ equals the union of $Contained(Y)$ and the super-rules of Y, for every father Y of X. Now, during the processing of Y, any super-rule of Y was added to $Contained(Y)$ (see Line 6). Therefore, upon processing X, we have that

$$Contained(X) = \bigcup_{Y \ is \ a \ father \ to \ X} Contained(Y)$$

as defined in Line 1.

Furthermore, in Line 7 any super-rule of Y is *removed* from $Cand(Y)$. Therefore, after the algorithm concludes with Y, the set $Cand(Y)$ is actually the set $CandReject(Y)$ described above. In addition, in Line 8, Y itself is added to the set if it is a sub-rule of any previous rule. Therefore, we have that

$$Cand(X) = \bigcup_{Y\ is\ a\ father\ to\ X} Cand(Y) - Contained(X)$$

as defined in Line 2.

Given that $Cand(X)$ is correctly initialized ($Contained(X)$ is needed for computing $Cand(X)$ and for processing X's sons), in steps 3-5 we iterate over all the candidates Z and check if X is a sub-rule of Z using an appropriate statistical test (line 4). The sets $Cand(X)$ and $Contained(X)$ are updated as needed so that the candidates of X's sons can be calculated later (lines 6-8).

It is clear that all rules and sub-rules are found (completeness) as we traverse the entire lattice keeping record of all potential rules. Furthermore, a contained sub-rule cannot be returned (soundness) as the *Contained* set prevents this from happening.

Expanding to many attributes on the right-hand side. The algorithm above finds all rules and sub-rules where the distribution measure contains a single attribute. We will now show that this exact algorithm solves the general case where the right-hand side contains an unlimited number of attributes. The following trivial lemma is the basis for solving the general case using the solution to the specific case described above.

Lemma 3.1 *Let $X \subseteq E_C$ be a profile containing categorical attributes and $J \subseteq E_Q$ a list of quantitative attributes. Then $X \Rightarrow M_J(T_X)$ if and only if for every $e \in J$, $X \Rightarrow M_e(T_X)$.*

Based on the above lemma, in a post-processing stage, we simply combine all rules found for any given profile (keeping track of the sub-rule hierarchy). We present the results in a tree of rules and sub-rules.

Complexity. As described above, calculating the distribution for each quantitative attribute in each node requires one pass over the database while updating the frequent sets. The complexity of this stage is similar to regular association rules algorithms (while working only on the categorical attributes).[1] In the next stage of the algorithm, we traverse the lattice finding rules and sub-rules. We note that an important advantage of the lattice structure and the fact that we traverse it using BFS (one level at a time), is that we may store it in sections on disk while minimizing disk access.

The algorithm for finding all rules and sub-rules executes k traversals of the lattice, where k is the number of quantitative attributes. We may actually traverse the lattice only once, but we need to store the appropriate sets separately for each attribute. It is easily shown that for m frequent sets, the number of edges in the lattice is less than $m \log m$. Therefore each traversal takes $O(m \log m + m \cdot t)$ where t is the amount of work in each node. For each node we execute two set-unions over sets contained in all its fathers, and one set-difference. Set union and difference are operations linear in the size of the sets (when the sets are kept in sorted order). This is equivalent to saying that we execute two set unions for every edge in the lattice. In the worst case, *Cand* and *Contained* can be of size $O(r)$ where r is the number of rules in the output (r is far smaller than m in practice), and therefore we have an overall complexity of $O(m \log m + k \cdot r \cdot m)$. However, in reality the complexity is far lower.

[1] This is a critical difference to [12] who need to construct frequent sets over all attributes, including the quantitative ones. Furthermore, the discretization of the quantitative data typically results in a significant blowup of the number of frequent sets.

Compound rules. We note that the ideas in sections 3.1 and 3.2 may be combined in order to find rules with profiles containing many categorical and a single quantitative attribute. For every frequent set X, we run *Window* on T_X. We may run *Window* in parallel on each frequent set and efficiently achieve the desired result.

4 Experimental Evaluation

4.1 A Rigorous Evaluation

Measuring success. A major problem confronting data mining researchers is the question of how to measure success. In any evaluation it is necessary to measure both the correctness of the rules obtained and the *interestingness* of those rules to the user. A rule saying that "abortion $\Rightarrow$ female" is certainly true, but is completely uninteresting.

As we have discussed, we deal with the issue of *correctness* by validating each rule with statistical test. On the other hand, evaluating how interesting are the rules obtained more difficult. We, as computer scientists, are certainly not qualified to judge whether or not rules found in a medical database, for example, are of interest. As our goal is to help the user, we believe that the only way to measure success is to ask the end user himself. He is not only the most *objective* judge, but the only one qualified to judge at all. He can inform us on whether or not the type of rule found is helpful, if it revealed information *new* to him and what percentage of the rules found are truly interesting.

Our evaluation. We therefore tested our new definition on a real-world database and had a domain expert perform an in-depth evaluation of the results. The database we mined is from the field of linguistics and was built during a study on the English writing habits of non-native English speakers. The study was conducted by Prof. Joel Walters of the English Department in Bar-Ilan University. Previously Walters had researched the database extensively using standard statistical tools such as SPSS. We presented him with the association rules we discovered, and asked him to evaluate the rules. For each rule, he categorized the rule as: non-interesting, interesting, or very-interesting. Among the interesting and very-interesting rules, he marked if he would have otherwise found them or not (e.g., using SPSS).

Description of the database. The database is based on a study involving essay writing under different conditions. Each transaction in the database contains data extracted from an essay and background information on the author. The data extracted from the essay includes part-of-speech measures (as in percentage of words used which are nouns, adjectives etc.) and lexical measures (relating to the level of words used, the level of variation, originality and many others). The database contains 643 transactions and 42 attributes: 15 categorical and 27 quantitative. With approximately 27,000 entries and 42 different factors (making for many hundreds of possible patterns), the database is large and computerized tools are necessary. (We note that although there are not many transactions in the test database, there are a large number of attributes. This is significant because the number of attributes usually has a far stronger impact on the complexity than the number of transactions.)

The evaluation results. The results of the quantitative evaluation by the domain expert (Walters) are summarized in Table 1. The row labeled "*Otherwise not found*" contains the number of rules that our expert found interesting or very interesting, yet claimed that he would not have found using regular statistical tools.

	Categorical $\Rightarrow$ Quantitative	Quantitative $\Rightarrow$ Quantitative	Overall
Number of Rules:	70	284	354
Not Interesting:	50 or 71%	178 or 63%	228 or 64%
Interesting:	16 or 23%	86 or 30%	102 or 29%
Very Interesting:	4 or 6%	20 or 7%	24 or 7%
Otherwise Not Found:	12 or 17%	80 or 28%	92 or 26%

Table 1: *Interestingness* classification of rules

Overall we see that 36% of rules were interesting or very interesting, 26% of all rules would not have been found using the standard hypothesis checking model. 7% of the rules were graded very interesting and would not have otherwise been found. This is *very* high for an automated tool and the result is critical to the usefulness of the method. Users are unlikely to use tools which provide interesting results hidden amongst endless junk.

Rule complexity. We further looked at the breakdown of interesting rules within the "Categorical $\Rightarrow$ Quantitative" rules and found that more than 50% of rules with a *single* categorical attribute in the profile were graded interesting. On the other hand, those with more than one attribute in the profile were judged not-interesting in 86% of the cases. This is most likely due to the difficulty in understanding complex rules. Furthermore, we found that most rules with more than one attribute in the profile had only a single attribute on the right-hand side. This is in contrast to rules with a single attribute in the profile which generally had many attributes in the distribution measure (frequently between 5 and 10). Our evaluator explained that he learned much from the *combination* of different attributes on the right-hand side (often including both part-of-speech and lexical measures together). This conjunction of varying exceptional behaviors enabled him to compare attributes and develop a more complete picture of the writing patterns of a given group of participants.

In summary, we found that most interesting rules have simple profiles. On the basis of this, it would seem that our algorithms cover most of the interesting cases.

A qualitative evaluation. The strength of our technique can be seen by viewing a number of results judged to be interesting by our evaluator. We present two examples here.

In the study, some participants were given a source text and were asked to base their essay upon it and others were not. We present a suprising rule regarding the effect of these source texts. Our evaluator judged the following rule to be interesting and claimed that he would not have found it using standard statistical tools:

$$\textsf{First Language} = \textsf{Russian AND No Source Text} \Rightarrow \textsf{mean frequency of the word } \text{“\textit{the}”} = 3.9\%$$

The mean frequency of the word "*the*" over all the texts was 6.7% (i.e., on average, 6.7% of the words in a given text were "*the*"). Therefore, this rule tells us that Russian essay writers who were not presented with a source text used the word "*the*" well below average. It is a known fact to Linguists that Russian has no definite article. Therefore, is not surprising to see that Russian writers use the word "*the*" less than the average. However, this was *not* inferred from the database. Rather, we found that only when the participants had no source text to base on, they fell back on their Russian habit of not using a definite article. On the basis of this rule Walters found (using a statistical query) that Russians given a source text used the word "*the*" 8.1% of the time in contrast to the 3.9% result shown in the rule. His conclusion was that essay writing based on source texts should be used for teaching use of the definite article.

This rule is an important example of where relying only on a priori hypotheses is not enough. It cannot be inferred from the database that Russians use the word *the* less and it is highly unlikely that one would guess that the source texts were the key. As a result, something *new* was learned in discovering that having a text to base one's writing on can also improve style. This discovery was of great importance to Walters.

The following example demonstrates the comprehensive picture obtained by rules combining many factors together in one. We will need to first discuss the Linguistic meaning of the attributes involved:

- *English Proficiency:* this is a categorical attribute grouping the subjects by proficiency. The *MA* category refers to students studying a Master's degree who must take a compulsory English course due to their low level.
- *Lexical Originality:* this attribute counts the percentage of words used by a given participant not used by anyone else in the study. Lexical originality is an indicator of richness of vocabulary.
- *Spelling Errors:* percentage of misspelt words.
- *Length of Essay:* the number of words in the essay. Length is a known and accepted indicator of skill, with high proficiency students writing significantly longer essays than those with low proficiency.
- *Source Originality:* for those provided with a source text (see above rule) this attribute counts the percentage of words used by the participant not appearing in the source text. This indicates how willing the writer is to depart from what she has in front of her and use new words.

We are now ready to present the following rule, graded very interesting (and unlikely to be found using regular methods) by our expert:

English Proficiency = MA ⇒

Lexical Originality:	mean = 3.3%	(overall mean = 6.7)
Spelling Errors:	mean = 9.5%	(overall mean = 6.5)
Length of Essay:	mean = 87.8 words	(overall mean = 172.4)
	deviation = 40	(overall deviation = 130)
Source Originality:	mean = 51.8%	(overall mean = 43.5)

The lexical originality of the students is low pointing to the fact that their vocabulary is limited and very similar to each other. We further see that they had many spelling errors and wrote extremely short essays, strengthening our view of their proficiency as low. The source originality measure, however, provides us with a very surprising result in that they were *more* original than the average student. Even though their level is inferior, they were willing to be daring and use words they did not see in front of them. This pattern is the opposite of what we would expect and is therefore very valuable. We note that the low deviation in the lengths of their essays gives evidence that the level of proficiency is quite homogenous in the group. This too is important information and is an example of where a variance-based rule is useful.

We see here that the grouping together of the different interesting distributions provides a concise summary of the exceptional writing habits of those MA students. This information is very important for those constructing an educational plan for these students.

Running time. We ran our tool on a Pentium-Pro with 128Mb RAM. With minimum support 40, the overall time taken on the *Linguistics* database (643 transactions, 42 attributes) was just 10.1 seconds. Of this time, 0.79 seconds was spent on Window (finding rules: 1 Quantitative $\Rightarrow$ 1 Quantitative, not including time to load the data into memory). For a minimum support of 20, the overall time taken was 23.8 seconds, the time spent of Window was 0.81 seconds.

The statistical tests. We found that the use of statistical tests to validate the accuracy of potential rules is crucial. In the *Linguistics* database, 29,959 potential rules were discovered by the algorithm, but were rejected due to lack of statistical confidence (a confidence level of 95%). On the other hand, only 354 rules were accepted. A person may view 350 rules in a reasonable amount of time; however, it is not feasible to go through 30,000 rules. More importantly, this shows that the vast majority of potential rules actually have no statistical basis and thus their validity is highly questionable.

Summary. We found many rules determined to be truly interesting and revealing to the user. Furthermore, a high percentage of these rules were not likely to have been discovered at all without our data mining tool. These two results show that our notion of quantitative associations is truly helpful and provides what we would expect from a data mining tool. Finally, our rules are easily understood and interpreted and due to the statistical tests, provide reliable information.

In all, we believe that the results provide strong indication of the usefulness of our new definition in real world applications. Further experimentation is necessary in order to fully measure the extent of its applicability in the various domains.

4.2 A Comparative Evaluation of [12]

Remember that in [12] a quantitative association rule is defined as a rule $X \Rightarrow Y$ with a certain support and confidence, where X and Y contain categorical items or numerical ranges. Their algorithm is based on mapping the problem to the categorical case by way of discretization, finding all association rules and then filtering superfluous nested rules. We now present an evaluation (rendered by our expert) of the rules generated by [12] and examples of some of the problems. We used the *Linguistics* database as our basis for the evaluation.

A quantitative evaluation. In order to create a fair comparison of [12] rules to ours, we limited the search for rules with one attribute on each side of the rule. This was so we could choose a relatively high maximum support, low minimum support and low K (otherwise we would encounter extreme computational difficulties, especially with a database of 42 attributes). We also wished to limit the number of rules we found to something that could be realistically evaluated. In order to do this, we chose 10 pairs of attributes uniformly at random (with the condition that at least one of the attributes was quantitative) and obtained rules from these pairs only. We then evaluated these rules as a sample of the set of all rules of this type.[2]

After some fine tuning we settled on the following input parameters:

Minimum Support	= 40 or 6%
Maximum Support	= 0.7
Completeness Level (K)	= 9
Interest Level (R)	= 1.5
Minimum Confidence	= 0.6

The minimum support was chosen to be the same as for our algorithm and we chose K=9 in order to reduce the number of rules found.

We obtained 81 rules in the output. As the number of possible pairs of attributes equals 756 (for 15 categorical and 27 quantitative attributes) we expect approximately 6,000 rules from this limited type alone. This is a significant problem to anyone who must review the results (note that with our definition, only 354 rules in total were found for the same minsup).
The results of the evaluation are as follows:

Number of rules:	81 (chosen *randomly*)
Not interesting:	80
Interesting:	1
Very interesting:	0

Overall, 1.2% of the rules were judged to be interesting.

Discussion. We will now present two examples demonstrating some of the conceptual drawbacks to the [12] definition. As we have mentioned, range is a weak measure for describing a distribution. The following rule we found illustrates this point:

$$\textsf{Lexical Variation} \in [27.91,\ 62.02] \Rightarrow \textsf{Lexical Originality} \in [4.27,\ 31.19]\ \textsf{(sup 42\%, conf 68\%)}$$

Lexical variation measures the diversity of vocabulary used by the participants. Lexical originality is a measure of how many words the participant used that no other participants in the study used. The above rule tells us that those with average and below lexical variation (the average is 59.2) have high lexical originality (the average is 6.7, which is included in the interval, but the interval extends far to the right). This rule may seem interesting and surprising but in actuality is very misleading. It is true that 68% of the values are between 4.27 and 31.19 but nearly all of those are below 13! In fact, the following rule was also found:

$$\textsf{Lexical Variation} \in [27.91,\ 62.02] \Rightarrow \textsf{Lexical Originality} \in [2.91,\ 12.54]\ \textsf{(sup 42\%, conf 68\%)}$$

We checked the original data and found that only 39 transactions within the Lexical Variation range, have a Lexical Originality value above 12.54.

Another example of rules found to be misleading due to the distribution measure is the following:

[2]We stress that our comparison is with respect to the raw algorithm of [12] without any post-processing on the output (beyond the filtering techniques suggested there). Indeed, using some additional statistical measures to filter out rules that are output by the [12] algorithm may change the results. Nevertheless, our comparison is instructive.

$$\text{English Proficiency} = \text{Advanced} \Rightarrow \text{Length} \in [91, 214] \text{ (sup 6\%, conf 50\%)}$$
$$\text{English Proficiency} = \text{Fluent} \Rightarrow \text{Length} \in [91, 214] \text{ (sup 25\%, conf 50\%)}$$

The "English Proficiency" categories ranked the student's English skills according to accepted levels. The above pair of rules gives us the feeling that the different proficiencies "Advanced" and "Fluent" are actually very similar regarding their writing ability (as essay length is a known predictor of writing ability). This is a surprising rule and would be very interesting. However, consider the following statistics: the average essay length in the entire study was 172, for "Advanced" was 108 (way below average, i.e. low proficiency) and for "Fluent" was 222 (a strong sign that this group is well skilled). This example strongly confirms our claim that describing a numerical distribution with a range and probability may be very misleading. Not only may we not find interesting behavior, we may be presented with rules which lead us to erroneous conclusions.

5 Multiple Comparison Procedures and Data Mining

A general problem in data mining is that of *multiple hypothesis testing*. The theoretical foundations we rely on when using statistical tests to check a single hypothesis do not directly apply when checking many hypotheses. Regular statistical tests enable us to bound the probability of observing a *given* event by chance, e.g. the probability of a fair coin falling on heads ten times in a row. If this probability is small we deduce that the behavior in not coincidental, but rather reflects a genuine characteristic of the observed phenomena, e.g. we deduce that the coin is not a fair coin after all. However, when we test *many* hypotheses simultaneously, we can no longer make this inference in such a simple way. The reason being that while the probability of passing any single test may be small, the probability of passing one or more of the many tests may still be high. In statistical terms, applications in which multiple hypotheses are analyzed simultaneously are called *multiple comparison procedures*, and this problem arises in most data mining settings.

Rules generated on a random database. We demonstrate this point by presenting association-rule results obtained from a purely random database. Clearly, while association-rules may be found in such a database, they do not reflect "real" associations between the underlying attributes. That is, even if rules are found and pass appropriate statistical tests, these are purely a statistical coincidence. In particular, such rules have no predictive meaning (as the "true" underlying behavior is random, by construction). Nevertheless, such rules may be accepted and the number of accepted rules provides some indication to the significance of rules accepted on a real database. (That is, if many rules are output when an algorithm is run on a random database then this would lead us to doubt the accuracy of the rules output by the same algorithm on a real database.)

In order to study this issue, we built a database of 10 attributes and 10,000 transactions, the entries being uniformly distributed values between 0 and 1. We then ran our algorithm for quantitative association rules for different supports. We emphasize that all rules found were *tested statistically* according to our definition. Table 2 contains the number of rules found for different minimum supports. The growth is clearly exponential as the minimum support gets smaller. (This type of behavior can be analytically explained using a simplistic model of independent windows, and an analysis of the probability that a small window of values is biased.) This simple experiment shows that even using statistical tests, "association rules" are observed even when no information can truly be inferred. We note that this problem applies to the other definitions of association rules no less than it does to ours.

Support	Number of Rules
60	0
50	2
40	13
30	74
20	342
10	3498
5	19987

Table 2: Exponential growth of rules on a *random* database

Obtaining statistical confidence. The phenomenon of "meaningful" information surfacing even when none exists is a major problem in the theory and practice of data mining in general, and association rules, in particular. To overcome this problem, we must establish means to test the validity of multiple hypothesis checking procedures. Jensen and Cohen [6] provide an excellent analysis of a similar problem in the context of Decision Tree Learning, and provide solutions and directions. Here, we focus on association rules, as defined in this paper, and take some steps in the direction of providing a methodology for testing reliability. We note, however, that this subject still needs further research.

In order to be convinced of the correctness of rules found, we must be capable of accurately measuring the probability of error. In other words, we must be able to calculate the expected number of incorrect rules found. We begin by relating to the somewhat simpler issue of how many rules we expect to find if no correlations exist at all. This can be estimated in the following way. Given a database which we wish to mine, we construct a different *random* databases, based on exactly the same distributions. Specifically, we apply independent random permutations to the values of each attribute in the given database. The resulting database is one with no correlations, but with exactly the same distribution for each attribute. We then count the number of rules obtained in this randomized database. This randomization process is repeated a large number of times (say 50) and for each randomized database the number of rules is recorded. With this information, we have the mathematical tools to measure the probability of coincidentally obtaining the number of rules found in the real database. If this probability is small, we may conclude that real correlations do exist in the real database. We stress that this test should be executed on the exact database to be mined, and with the same minimum support.

We ran this experiment on the Linguistics database described in Section 4 with a confidence of 95% on each individual rule, and a minimum support of 40. We ran the algorithm on 50 random versions of the database, and found that the mean number of rules obtained was 16.7 with a variance of 17.7. Furthermore, the distribution was Poisson (as expected). Given the Poisson distribution with parameter of approximately 17, the probability of finding more than 30 rules on a random database is less than 0.001. In practice, we found more than 350 rules, which provides overwhelming statistical evidence that the results are not random.

We might further wish to conclude from the experiment that all but 30 of the rules are "real"; i.e., reflect true associations (with very high probability). Unfortunately, while this is a very plausible conclusion, it has no theoretical basis. This is because the experiment only tests the behavior of the algorithm on random, non-correlated databases. It does not test the algorithm's behavior on databases containing correlations. Hence, the experiment only proves that the database is not random, and that some correlations do exist, but not which correlations and how many. It is theoretically possible (although it seems highly unlikely), that due to dependencies, when few correlations do exist, the algorithm generates many more by coincidence. This is an intricate theoretical question requiring in-depth statistical research and is an important question for future work.

6 Open Questions and Future Work

Generalizing quantitative association rules. We introduced a general framework for quantitative association rules in the form of *"Profile $\Rightarrow$ Significant Distribution"*. However, we provided concrete the definitions and algorithms for rules of two specific types only (categorical-to-quantitative and quantitative-to-quantitative with a single attribute on the left hand side). These two types proved to be very useful and are important categories of rules. However, we would like to see a truly general definition of quantitative association rules. In particular, rules where the left-hand side contains two or more quantitative attribute are of interest. We note that the difficulty here is two-fold. Firstly, on the algorithmic side, it is easily shown that it is not possible to use one-dimensional rules in order to find *all* two-dimensional rules. Secondly, a conceptual difficulty arises in that the union of two overlapping, above-average rules is not necessarily above-average (a problem that does not arise in the one-dimensional case). This may lead to the undesirable property of many overlapping rules. Thus, the notion of maximality must be defined, and it is not clear how to do so. We note that the works of Fukuda et. al [4] and Yoda et. al [14] directly address the issue of rules with two quantitative attributes in the left-hand-side.

Other statistical tests. Another interesting question for future work is that of the effect of the specific statistical tests used. We used the Z-test as it is the most natural mean test, especially as we need not assume anything about the distribution of the values. However, the effects of using other tests and methods, or even a combination of them, is an interesting and important issue for continuing research.

7 Acknowledgments

We would like to thank Ronen Feldman for his invaluable contributions. The in-depth evaluation is due to Joel Walters and we thank him for his great investment of time and effort. Finally, we thank Lawrence Freedman for his enlightening discussions regarding the statistical background necessary for our work.

References

[1] R. Agrawal, T. Imielinski and A. Swami. Mining association rules between sets of items in large databases. *Proc. of the 1993 ACM SIGMOD Intl. Conference on Management of Data*, pp 207-216.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. *Proc. of the 20th Intl. Conference on VLDB*, 1994.

[3] S. Brin, R. Motwani and C. Silverstein. Beyond Market Baskets: Generalizing Association Rules to Correlations. *Proc. of the 1997 ACM SIGMOD Conference on Management of Data*, 1997.

[4] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama. Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms and Visualization. *Proc. of the 1996 ACM SIGMOD Conference on Management of Data*.

[5] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama. Mining Optimized Association Rules for Numeric Attributes. *Journal of Computer and System Sciences*, **58**, 1–12, 1999.

[6] D. Jensen and P. Cohen. Multiple Comparisons in Induction Algorithms. *Journal of Machine Learning* **38**:309-338, 2000.

[7] W. Kloesgen, *Exploration of Simulation Experiments by Discovery.* Proceedings of KDD-94 Workshop, AAAI-94. Further information may be found at the explora homepage: http://orgwis.gmd.de/projects/explora.

[8] Lindgren, Bernard W. *Statistical Theory.* Macmillan Publishing Co., Inc. New York, 1976.

[9] H. Mannila, H. Toivonen and A. I. Verkamo. Efficient Algorithms for Discovering Association Rules. *KDD-94: AAAI Workshop on Knowledge Discovery in Databases*, pp 181-192, 1994.

[10] Y. Morimoto, H. Ishii and S. Morishita. Efficient Construction of Regression Trees with Range and Region Splitting. *Machine Learning*, **45**, 235–259, 2001.

[11] G. Piatetsky-Shapiro. Discovery, Analysis, and Presentation of Strong Rules. In *Knowledge Discovey in Databases*, pp 229-248, 1991.

[12] R. Srikant and R. Agrawal. Mining Quantitative Association Rules in Large Relational Tables. *Proc. of the ACM SIGMOD Conference on Management of Data*, 1996.

[13] H. Toivonen. Sampling Large Databases for Association Rules. *Proc. of the 22nd VLDB Conference*, 1996.

[14] K. Yoda, T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama. Computing Optimized Rectilinear Regions for Association Rules. *Proc. of KDD '97*, August 1997.

[15] Z. Zhang, Y. Lu and B. Zhang. An Effective Partitioning-Combining Algorithm for Discovering Quantitative Association Rules. *Proc. of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1997.

[16] http://lib.stat.cmu.edu/datasets.