OTTO-VON-GUERICKE-UNIVERSITÄT MAGDEBURG

Analyzing the Similarity of Association Rules over Time

— Studienarbeit —

vorgelegt von

Stephan Kauschka

Stephan.Kauschka@st.ovgu.de

22. Dezember 2009

Abstract

Many companies nowadays collect huge amount of data which have to be analyzed in order to gain crucial advantages in highly competitive market environments. Association rule change mining has been suggested as one technique to cover this task. The amount of the resulting association rules however is usually too vast to be investigated manually. This thesis reviews the use of distance measures in order to help human experts to find interesting rules and suggests own measures for this task. It especially focuses on the use of rule histories for those distance measures. The use of the interchange format PMML to incorporate the distance measures is investigated. The results finally were implemented in a software called IDEAL.

Kurzfassung

Viele Unternehmen sammeln heutzutage riesige Mengen an Daten welche analysiert werden müssen um lebenswichtige Vorteile in hoch kompetitiven Marktumgebungen zu erhalten. Assoziationsregeln und ihre zeitliche Änderung wurden als Werkzeuge zur Bewältigung dieser Aufgabe vorgeschlagen. Die Menge der resultierenden Assoziationsregeln ist jedoch gewöhnlich zu riesig um sie einzeln per Hand zu analysieren. Diese Arbeit bespricht die Verwendung von Abstandsmaßen um menschlichen Experten das Finden interessanter Regeln zu erleichtern. Außerdem werden eigene Maße für diese Aufgabe vorgestellt, welche hauptsächlich die Entwicklungsgeschichte der Regeln betrachten. Weiterhin werden die Möglichkeiten des Austauschformats PMML zur Einbeziehung der Abstandsmaße untersucht. Die Ergebnisse wurden dann letztlich in einer Software namens IDEAL implementiert.

Contents

Chapter 1

Introduction

Caused by technological advances, corporate databases nowadays have grown to unexpected levels. They are expanding from petabyte to exabyte sizes and their growth is just not stopping as more and more data is collected. The sheer amount of data is inconceivable and impossible to understand by human means. But as the world is changing faster as ever before the timely analysis of current developments and the proper response to those has become a key survival strategy for market based corporates. To tackle this tremendous task data mining has become an invaluable tool in investigating huge amounts of data in order to find new information. Information that may mean the decisive advantage over other companies. One of the most important, because most versatile, data mining techniques are association rules. Initially they have been introduced by [Agrawal et al., 1993] in the domain of market basket analysis. The goal was to find all those items which frequently occur together in transactions to form rules that predict such co-occurrences. That is, associations among data items were traditionally used to understand the static structure of a database. These days it has been understood that they are also capable of describing its dynamic behaviour ([Böttcher, 2005]). In the context of association rule change mining rules can be used to predict the future, as mentioned before a key ability in todays markets. The problem with association rules is that, while they are intended to help in understanding huge amounts of data, their own amounts are usually also too big to be manually investigated by human specialists. To find those rules that can potentially be exploited is a difficult task. The focus of this work lies within the challenge to help human specialists in detecting interesting rules. While interestingness values and distance measures have been applied separately in the past, here it has been investigated how they can be combined to assist human experts. The premise is that when an interesting rule has been found using interestingness values, other relevant rules might be found using distance measures. Such distance measures may be based on the structure of association rules, that is they discover rules which cover similar topics. But since this approach does not reveal related rules that cover different portions of a database it is suggested to use distance measures that are based on the history of an association rules. Histories represent the development of certain rule measures and comparing those developments might yield hidden relations among association rules.

This thesis aims at developing such distance measures and use them to find hidden relations among association rules. It also investigates how association rules can be stored together with their distance measures in an open interchange format called PMML. The results were then implemented in an existing software prototype called IDEAL with the intention of guiding its users to find interesting association rules in an efficient manner.

The thesis is structured as follows:

Chapter 2 gives the necessary background information on association rules and association rule change mining. It also gives an overview of the interestingness measures that have been used in the IDEAL tool.

Chapter 3 introduces the notion of a metric and explains the correlation coefficient. It then presents related approaches for defining distance measures on association rules and finally introduces those distance measures that have been implemented for the IDEAL tool.

Chapter 4 reviews the open interchange format PMML and presents how this format has been adopted by [Reichelt and Winkelmann, 2008] to meet the requirements of IDEAL. Afterwards it is investigated how it can be further adopted to be used with distance measures.

Chapter 5 reviews the IDEAL tool and the changes that have been implemented in order to help its users in finding interesting association rules.

Chapter 6 summarizes the results of this thesis and gives some possible starting points for future research.

Chapter 2

Association Rules

2.1 Preliminaries

Initially association rule mining was introduced in the field of market basket analysis [Agrawal et al., 1993]. The goal was to find interesting frequent co-occurrences between items of sales transactions like: "If charcoal and lighter are purchased together, then in 80% of the time also matches will be bought.". These co-occurrences are called association rules with the given percentage being referred to as the confidence of the rule. One of the earliest applications for those rules was to help solving problems like the arrangement of items in a store in order to increase their sales volume. Nowadays association rules still help in making business decisions but their application field isn't constrained to sales transactions any more - they can be applied to every relational database in order to find interesting correlations between attributes.

Formally ([Agrawal and Srikant, 1994]), association rules are defined over a set of transactions $\mathcal{D}$. Let $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ be a set of items, then each subset $\mathcal{X} \subseteq \mathcal{I}$ is called an itemset. All transactions $\mathcal{T} \in \mathcal{D}$ are subsets of $\mathcal{I}$ and therefore itemsets. It is said a transaction $\mathcal{T}$ supports an itemset $\mathcal{X}$ if $\mathcal{X} \subseteq \mathcal{T}$.

An association rule is an implication of the form $\mathcal{X} \Rightarrow \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are itemsets with $\mathcal{X} \subset \mathcal{I}$, $\mathcal{Y} \subset \mathcal{I}$ and $\mathcal{X} \cap \mathcal{Y} = \emptyset$. $\mathcal{X}$ is called antecedent and $\mathcal{Y}$ consequent of the rule. A rule $r : \mathcal{X} \Rightarrow \mathcal{Y}$ is called a generalization of a rule $r' : \mathcal{X}' \Rightarrow \mathcal{Y}$ if $\mathcal{X} \subset \mathcal{X}'$ ([Li et al., 2004]). This is denoted by $r \succ r'$. Reversely, r' is called a specialization of r. A special case sufficient for most applications are association rules with just one item in their consequent. Those rules can always be created by splitting a rule with multiple items in its consequent into several rules with the same antecedent but just one item in their consequent. This case also satisfies the requirements of the IDEAL tool covered in chapter 5 and hence will be considered in the following. Rules with just one item in their antecedent will be written as $\mathcal{X} \Rightarrow y$, with $\mathcal{X} \subset \mathcal{I}$ and $y \in \mathcal{I}$.

The previously introduced measure called confidence is the ratio of all transactions that support the antecedent and consequent of a rule to all transactions that support the antecedent of the rule. It can be interpreted as the conditional probability $P(\mathcal{Y} \mid \mathcal{X})$, the probability of the consequent when the antecedent holds. Therefore confidence is a measure for the reliability of a rule:

$$\mathit{conf}(r : \mathcal{X} \Rightarrow \mathcal{Y}) = \frac{|\{\mathcal{T} \in \mathcal{D} \mid \mathcal{X} \cup \mathcal{Y} \subseteq \mathcal{T}\}|}{|\{\mathcal{T} \in \mathcal{D} \mid \mathcal{X} \subseteq \mathcal{T}\}|} \tag{2.1}$$

The significance of a rule regarding the underlying set of transactions $\mathcal{D}$ is measured by its support value, which is defined as the fraction of transactions satisfying the rule. Since the following definition does not take the structure of an association rule into account, it is also applicable to itemsets in general and can be interpreted as $P(\mathcal{X}\mathcal{Y})$ for a rule or as $P(\mathcal{X})$ for a single itemset:

$$\mathit{supp}(r : \mathcal{X} \Rightarrow \mathcal{Y}) = \frac{|\{\mathcal{T} \in \mathcal{D} \mid \mathcal{X} \cup \mathcal{Y} \subseteq \mathcal{T}\}|}{|\mathcal{D}|} \tag{2.2}$$

Another, more intuitive [Böttcher, 2005] measure for the significance of a rule is the antecedent support, sometimes referred to as the coverage of a rule. It measures how often a rule is actually applicable:

$$\mathit{asupp}(r : \mathcal{X} \Rightarrow \mathcal{Y}) = \frac{|\{\mathcal{T} \in \mathcal{D} \mid \mathcal{X} \subseteq \mathcal{T}\}|}{|\mathcal{D}|} \tag{2.3}$$

To further measure the interestingness of a rule, [Brin et al., 1997] introduced a frequently used measure called lift. It compares the co-occurrences of $\mathcal{X}$ and $\mathcal{Y}$ to their expected co-occurrences in case of statistical independence:

$$\mathit{lift}(r : \mathcal{X} \Rightarrow \mathcal{Y}) = \frac{\mathit{supp}(\mathcal{X}\mathcal{Y})}{\mathit{supp}(\mathcal{X}) \cdot \mathit{supp}(\mathcal{Y})} \tag{2.4}$$

A measure equal to one implies statistical independence between $\mathcal{X}$ and $\mathcal{Y}$ while a value unequal to one measures the deviation from independence.

There are several algorithms for mining association rules with many variations and extensions, but most of them are based on the same principles, which shall be outlined briefly. Some of the most prominent algorithms are Apriori [Agrawal and Srikant, 1994], Eclat [Zaki, 2000] and FP-growth [Han et al., 2004], with Apriori currently being the only one implemented for the IDEAL tool. In general, mining association rules might yield every possible proper subset[1] of items in $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ as an antecedent of a rule with any other of the remaining items as its consequent, resulting in a huge amount of generated association rules. The size of the generated rule set $\mathcal{R}$ of rules with one item in the consequent is given by:

$$|\mathcal{R}| = \sum_{i=1}^{n} \binom{n}{i} \cdot (n - i) \tag{2.5}$$

$|\mathcal{R}|$ grows exponentially with $|\mathcal{I}|$. So even a small database with just 20 different

[1] Rules with $\mathcal{I}$ as their antecedent have $\emptyset$ as their consequent and are discarded in this theoretical considerations due to the focus on rules with a consequent size of one.

items would yield 10.485.740 association rules which is infeasible to explore by a human analyst, let alone the fact that the vast amount of rules would be of no interest at all. To decrease the size of the rule set, lower thresholds for support $supp_{min}$ and confidence $conf_{min}$ have been introduced. The task of any rule mining algorithm is now to find all association rules that satisfy those thresholds. This can be accomplished in a two step approach:

1. Find all sets of items whose support is above $supp_{min}$. Such itemsets are called large or frequent itemsets and can be obtained by traversing all possible subsets of $\mathcal{I}$. Due to the downward closure property of the support measure, that is the support of a superset is always lower than or equal to the support of the original set, all supersets of sets that don't satisfy $supp_{min}$ can be discarded. So starting with sets of size one, all supersets are investigated until one of them doesn't satisfy $supp_{min}$.

2. Now the large or frequent itemsets are used to generate the association rules that satisfy $conf_{min}$. This is done by examining all subsets of all frequent itemsets. Each item of those subsets is one after another considered as the consequent and the remaining items as the antecedent of the rule. If the resulting association rule satisfies $conf_{min}$ it is transferred into the result set.

While the resulting association rules satisfy $supp_{min}$ and $conf_{min}$, their number is usually still too big to be analysed by a human expert. To further decrease the size of the association rule set several approaches like constrained mining an pruning have been introduced [Böttcher, 2005]. Techniques that are independent of the underlying transaction set $\mathcal{D}$ are referred to as constrained mining while dependent techniques are considered as pruning. In the case of constrained mining, for example only association rules with certain sizes for their antecedent and consequent might be relevant. Additionally only rules that support certain items or rules whose items satisfy specific conditions could be returned for further analysis. A third aspect when defining constraints are aggregate functions. For instance the summed costs of all items in a rule should exceed a certain threshold in order to be interesting to the analyst.

Pruning however, dismisses rules based on measures like interestingness, significance or redundancy. Although those measures are related to the rule quantity problem they also improve the rule quality, since they discard rules deemed uninteresting to the user. Obviously the definition of the interestingness of a rule is highly application dependent and therefore plenty of related measures do exist. A brief overview as well as further references can be found in [Böttcher, 2005]. Interestingness measures defined in the context of rule change mining and used by the IDEAL tool will be presented in section 2.3.

2.2 Association Rule Change Mining

Traditionally, association rule mining was conducted under the assumption that discovered rules were invariable, that is their properties did not change over time. This assumption however barely ever holds in real world applications. Additionally, this approach only allows for reactive decision making, since there is no indicator for the future behaviour of an association rule. Temporal aspects therefore might contain valuable information which have to be exploited by a company in order to be competitive. For this purpose the standard approach for association rule mining has been extended to incorporate the temporal dimension and consequently to detect and analyse changes in the behaviour of association rules, which ultimately allows for active decision making.

Formally [Böttcher, 2005], the set of transactions $\mathcal{D}$ is now a timestamped dataset, i.e. every transaction is associated with a date or time value. The minimal timespan that covers all transactions is denoted by $[t_0, t_k]$. To track changes of association rules, $[t_0, t_k]$ is partitioned in $k > 1$ non-overlapping periods $T_i := [t_{i-1}, t_i]\,, 1 \leq i \leq k$. The periods do not have to be equal in length but should be chosen such that the resulting pairwise disjoint sets $\mathcal{D}_i$ of transactions that occurred during the corresponding period T_i have a size of $|\mathcal{D}_i| \gg 1$. The set of all periods is denoted by $\hat{T} := \{T_1, \ldots, T_k\}$.

After the partitioning, for every set $\mathcal{D}_i$ the rule mining algorithm will be applied, which yields for each period T_i a set of association rules $\mathcal{R}\left(\mathcal{D}_i\right)$. Since association rules are now related to time periods, properties like confidence, support, antecedent support and lift also get a temporal dimension. Therefore their definitions are extended in a straightforward way:

$$\mathit{conf}(r : \mathcal{X} \Rightarrow \mathcal{Y} \mid T_i) = \frac{|\{\mathcal{T} \in \mathcal{D}_i \mid \mathcal{X} \cup \mathcal{Y} \subseteq \mathcal{T}\}|}{|\{\mathcal{T} \in \mathcal{D}_i \mid \mathcal{X} \subseteq \mathcal{T}\}|} \tag{2.6}$$

$$\mathit{supp}(r : \mathcal{X} \Rightarrow \mathcal{Y} \mid T_i) = \frac{|\{\mathcal{T} \in \mathcal{D}_i \mid \mathcal{X} \cup \mathcal{Y} \subseteq \mathcal{T}\}|}{|\mathcal{D}_i|} \tag{2.7}$$

$$\mathit{asupp}(r : \mathcal{X} \Rightarrow \mathcal{Y} \mid T_i) = \frac{|\{\mathcal{T} \in \mathcal{D}_i \mid \mathcal{X} \subseteq \mathcal{T}\}|}{|\mathcal{D}_i|} \tag{2.8}$$

$$\mathit{lift}(r : \mathcal{X} \Rightarrow \mathcal{Y} \mid T_i) = \frac{\mathit{supp}(\mathcal{X}\mathcal{Y} \mid T_i)}{\mathit{supp}(\mathcal{X} \mid T_i) \cdot \mathit{supp}(\mathcal{Y} \mid T_i)} \tag{2.9}$$

It is possible that for some rules the values for support and confidence do not satisfy supp_{min} and conf_{min} in each period T_i. Those rules will be discarded in the following[2] and the remaining are summarised in the compound rule set denoted by:

$$\hat{\mathcal{R}}\left(\mathcal{D}\right) = \bigcap_{i=1}^{k} \mathcal{R}\left(\mathcal{D}_i\right) \tag{2.10}$$

Because each rule measure consists now of k values, those are summarized in ordered sequences called the history of each respective measure. Support and confidence histories

[2] They could actually be kept, but then missing values for support and confidence would have to be determined in another database scan, a potentially very time consuming task.

are defined next:

$$H_{supp}(r) := (supp(r \mid T_1), \ldots, supp(r \mid T_k)) \tag{2.11}$$

$$H_{conf}(r) := (conf(r \mid T_1), \ldots, conf(r \mid T_k)) \tag{2.12}$$

All other histories are defined accordingly.

The actual task of rule change mining is now to detect patterns in those rule histories. Patterns are global regularities in the values of a rule history, like stabilities, trends or cyclic behaviour. Although detection techniques for those patterns are outside the focus of this work[3], it shall be stated that they usually are applied before any pruning of the association rules has taken place. This is caused by the fact that although rules might be uninteresting at first glance, their histories could actually be quite informative. If pruning had been applied before, the history would have been lost and the user would remain unaware of this. However, without pruning the rule quantity problem translates directly into a pattern quantity problem. Since the pattern detection is applied to a huge amount of rules, it is likely that also a lot of patterns will be detected.

2.3 Interestingness Measures For Change Patterns

To alleviate the pattern quantity problem, related interestingness measures have been introduced. Their purpose is to help the user to find those rules and patterns that are relevant to him. The definition of such interestingness measures depends on specific patterns a rule measure shows. For example the mean value for a history that shows stability as a pattern is quite informative, since it gives the level around which a rule measure lies. For a history that shows an up- or downward trend however, the mean is quite useless, since it just estimates the value the history has in its middle and doesn't give any additional information. In the following subsections those measures will be outlined that are currently used by the IDEAL tool. All measures are defined for association rules that show an up- or downward trend. To simplify the following notations, the history of an arbitrary rule measure will be written as:

$$H := (v_1, \ldots, v_k) \tag{2.13}$$

2.3.1 Clarity

The clarity measure is defined as the certainty that a detected trend actually exists. Therefore it gains its maximum value for an upward trend if each value of a history is greater than his direct predecessor and respectively for a downward trend if each value is smaller than its predecessor. To calculate the measure the absolute value of the

[3] A detailed presentation can be found in [Böttcher, 2005]

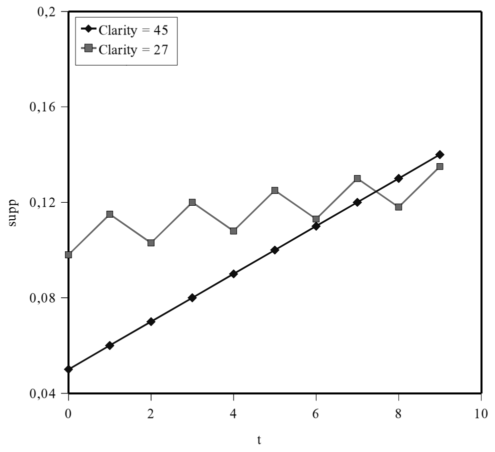


Figure 2.1: Two support histories with upward trend and their clarity values

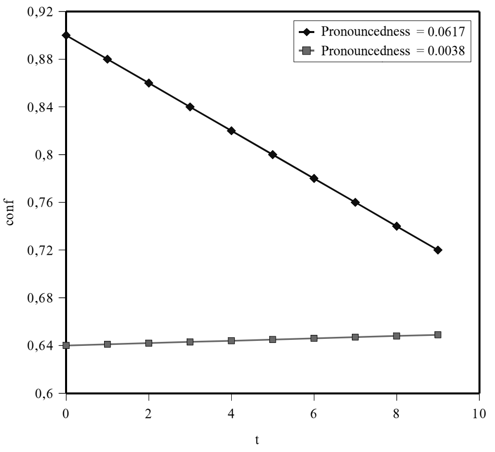


Figure 2.2: Two confidence histories and their pronouncedness values

Mann-Kendall test statistic [Mann, 1945] is used:

$$clarity(H) = \left| \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} sgn(v_j - v_i) \right| \tag{2.14}$$

The function sgn is defined as follows:

$$sgn(v_j - v_i) = \begin{cases} 1 & : \quad v_j - v_i > 0, \\ 0 & : \quad v_j - v_i = 0, \\ -1 & : \quad v_j - v_i < 0 \end{cases} \tag{2.15}$$

The clarity measure compares each value to all its successors. When an upward trend is prevalent then most of the successor values should be larger. Each time such a value is smaller the clarity is decreased. Likewise for downward trends most of the successors should be smaller, else the measure is decreased. For a history of length k the maximum clarity value is $\frac{k \cdot (k-1)}{2}$. If no trend exists the clarity is approximately 0. Figure 2.1 shows two fictitious support histories, one with a maximum clarity value (clarity = 45) and one whose trend is not that obvious (clarity = 27).

2.3.2 Pronouncedness

The pronouncedness measure aims to quantify the deviation of a trend from stability and therefore is a measure for the strength of a trend. Stability is defined as the mean line of a trend, so for the measure to be meaningful, all trends have to be compared to the same mean line. This is achieved by scaling each history, which is conducted by

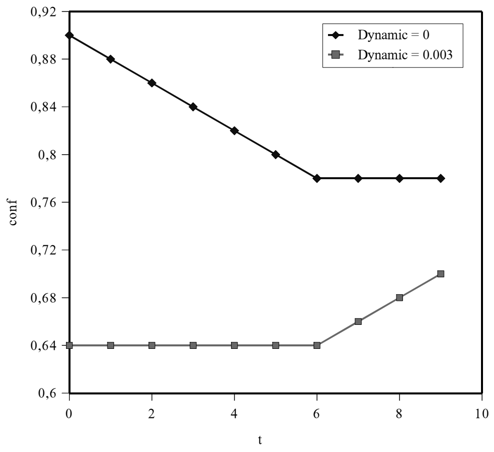


Figure 2.3: Two confidence histories and their dynamic values for $k' = 4$

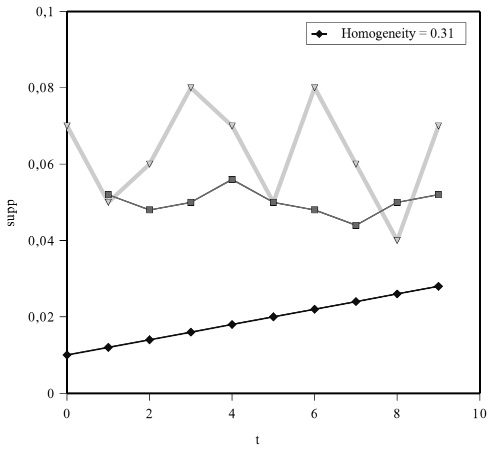


Figure 2.4: Three support histories and their corresponding homogeneity value

recalculating their values as follows:

$$v'_i = \frac{v_i}{\sum_{i=1}^{k} v_i} \tag{2.16}$$

This scaling preserves the relative changes between the values of a history but results in a common mean value of $\frac{1}{k}$ against which all scaled history values are compared for the pronouncedness measure:

$$\mathit{pronouncedness}(H) = \sum_{i=1}^{k} \left| v'_i - \frac{1}{k} \right| \tag{2.17}$$

The more a history value deviates from the mean value the greater the pronouncedness measure becomes. To illustrate the meaning of this measure figure 2.2 shows two fictitious confidence histories which both have a maximum clarity value but whose pronouncedness significantly differs.

2.3.3 Dynamic

The dynamic measure's purpose is to cover the recent development of a trend, that is it gives the rate of incline or decline for the last k' values of a history. The idea behind this measure is that rules, whose histories are currently evolving, might be of more interest than rules that are currently stable. On the other hand, rules that showed a trend in the past but now are stabilizing, can be identified by low dynamic values. Both cases are illustrated in figure 2.3. To make the dynamic values of different histories comparable to each other, those histories have to be scaled the same way as for the pronouncedness measure. For the resulting scaled history values a linear regression line is fitted and its slope m is used to calculate the dynamic measure:

$$dynamic(H) = |m| \tag{2.18}$$

High dynamic values indicate rapid changes in the recent history of a rule measure while values close to 0 indicate stability. Since the last k' values of a history might behave exactly the same way as their predecessors, the dynamic measure alone might not be of great interest. But by comparison with other measures like clarity or pronouncedness, possibly important changes in behaviour can be identified.

2.3.4 Homogeneity

The homogeneity measure compares rule histories to the histories of more general rules and detects differing behaviour. For that purpose for a rule r all rules r', that have exactly one item less in their antecedent, are considered. The actual calculation is divided into two steps. First the deviations of the history of rule r, denoted by $H(r)$, to all other histories $H(r')$ are determined and then those deviations are aggregated into a single value.

To gain the deviation of one history from another both histories have again to be scaled like for the previous interestingness measures. Let the scaled histories of two rules r and r' be denoted as $H(r) = (v'_1, \ldots, v'_k)$ and $H(r') = (w'_1, \ldots, w'_k)$. The deviation of those histories is then the sum of all their value differences:

$$deviation(H(r), H(r')) = \sum_{i=1}^{k} \left| v'_i - w'_i \right| \tag{2.19}$$

The set $\mathrm{dev}(r) = \{deviation(H(r), H(r')) : r' \succ r\}$ summarizes the resulting deviations which are aggregated to gain the homogeneity measure:

$$homogeneity(H) = agg(dev(r)) \tag{2.20}$$

Which operation is ultimately used for the aggregation is application dependent. Possible examples are the minimum, the maximum or the average of all deviations. The IDEAL software uses the average. The higher the homogeneity value the bigger the differences between the histories are. Figure 2.4 shows the support histories of three association rules of which the upper two are more general than the third. The lower rule might be potentially interesting because its behaviour differs from its more general rules in that it shows a clear trend while the other are rather stable.

Chapter 3

Distance Measures

One of the main tasks of the IDEAL tool, besides the mining of association rules, is to help the user in finding those rules he is actually interested in. While the already used techniques of pruning and ranking of association rules are very important and successful, experience shows that the amount of rules is often still overwhelmingly large, too large to be investigated manually rule by rule. In order to alleviate this problem the use of distance measures has been suggested. With those measures a user could be guided from interesting rules he is already aware of, to similar rules, which potentially might also be interesting to him. Using such distance measures could lead to the discovery of relations between rules that were not known before. The actual definition of this similarity can either be based on the structure of a rule, its database coverage or its behaviour, so a plethora of different measures are conceivable, which opens a big area of research in terms of their usefulness and interrelationships.

3.1 Preliminaries

3.1.1 Metrics

The definition of distance measures in general is not restricted by any rules and is completely domain specific. Usually a distance has a minimum value of 0 and growths as the compared objects are getting more and more dissimilar. To make the analysis of such distance measures easier the notion of a metric has been introduced. For a distance measure to be a metric three conditions have to be fulfilled.

Isolation: The distance between two objects A and B is zero only if the two objects are identical and vice versa.

$$d(A,B) = 0 \Leftrightarrow A = B \tag{3.1}$$

Symmetry: The distance between two objects A and B is independent from the point of view, that is whether $d(A,B)$ or $d(B,A)$ is considered.

$$d(A,B) = d(B,A) \tag{3.2}$$

Triangular Inequality: The distance between two objects A and C is always smaller or equal to the length of a path that visits an object B in-between.

$$d(A,B) + d(B,C) \geq d(A,C) \tag{3.3}$$

From those conditions follows that the distance between two objects is always greater or equal to 0. One of the best known examples for a metric is the euclidean distance. If a distance measure does not obey the isolation condition, that is the distance between to objects can be 0 even if those are not the same objects, but the other two, it is called a pseudo metric.

3.1.2 Correlation Coefficient

When investigating two random variables it is of interest how both vary together, that is of what kind their relationship is. An often used measure for this relationship is the covariance, which is defined as follows ([Montgomery and Runger, 2003]):

$$\sigma_{XY} = E\left[(X - \mu_X)(Y - \mu_Y)\right] \tag{3.4}$$

In this equation X and Y are the random variables and μ_X and μ_Y are their respective expected values. The function E yields the expected value of a function of two random variables, in this case the expected value of $(X - \mu_X) \cdot (Y - \mu_Y)$. If the values of X and Y are plotted against each other and tend to fall along a line of positive slope, then σ_{XY} is positive. If they form a line with negative slope σ_{XY} is negative. Therefore covariance is a measure of linear relationship between the random variables.

Another measure for the relationship between two random variables, that is often easier to interpret ([Montgomery and Runger, 2003]), is the correlation:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{3.5}$$

It is defined as the covariance of the two random variables divided by the product of their standard deviations. For any X and Y it can be shown that $-1 \leq \rho_{XY} \leq +1$ ([Montgomery and Runger, 2003]). The correlation scales the covariance by the standard deviation of each variable. Thus it is a dimensionless quantity that can be used to measure the linear relationships between two random variables. Since later on the population of random variables, that will be represented by rule histories, is not known, the covariances and standard deviations will be replaced by their sample estimates. This yields the sample correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \tag{3.6}$$

For r also holds $-1 \leq r \leq +1$. If the samples of X and Y show a positive linear relationship, r will be close to 1 and if they show a negative linear relationship, r will be close to -1. If both have no linear relationship r will be close to 0. That however does not mean that there is no relationship at all, since r only measures linear relations.

3.2 Related Approaches

The following section gives a review of related approaches defining distance measures between association rules. Their relevance to this work and some of their drawbacks will be discussed briefly.

3.2.1 Simple Support Based Distance

In the context of association rule mining [Toivonen et al., 1995] introduced their approach to alleviate the rule quantity problem through pruning and grouping techniques. Their pruning technique is based on the creation of so called rule covers which are related to the pruning of redundant rules in [Böttcher, 2005]. In short, rules are discarded when more general rules exist which cover the same database entries. Since this approach is not related to their attempt to define a useful distance measures it won't be discussed in more detail here.

However, after pruning their initial rule sets, [Toivonen et al., 1995] still faced a large amount of remaining rules. To better understand them their approach was to cluster them. To this end they introduced the following simple distance measure between association rules with the same consequent. Let $r1 : \mathcal{X} \Rightarrow \mathcal{Z}$ and $r2 : \mathcal{Y} \Rightarrow \mathcal{Z}$ be two association rules. Their distance is then defined as:

$$d(\mathcal{X} \Rightarrow \mathcal{Z}, \mathcal{Y} \Rightarrow \mathcal{Z}) = supp(\mathcal{X}\mathcal{Z}) + supp(\mathcal{Y}\mathcal{Z}) - 2 \cdot supp(\mathcal{X}\mathcal{Y}\mathcal{Z}) \tag{3.7}$$

The distance gives the percentage of tuples in the database which are not covered by both rules. While this approach is very intuitive, it is restricted in that it is only applicable for rules with the same consequent.

3.2.2 Conditional Market-Basket Probability Distance

Based on the previous approach, [Gupta et al., 1999] introduced another distance measure using the support values of two association rules. They claim that one of the problems of the distance measure of [Toivonen et al., 1995] is, that it grows as the database grows. Obviously this is true and it might constitute a problem since distances between two rules would not be static but evolve over time - a property which clearly impedes the analysis of association rules and their relationships. Another detriment is that due to the focus on support values, rules with a high support will on average tend to have higher distances to everybody else. To counter those problems [Gupta et al., 1999] introduced the following distance measure:

$$d_{i,j} = 1 - \frac{|m(BS_i, BS_j)|}{|m(BS_i)| + |m(BS_j)| - |m(BS_i, BS_j)|} \tag{3.8}$$

The sets BS_i and BS_j are the itemsets for the association rules i and j and $m(X)$ is the set of all transactions supporting the itemset X. Rules that do not cover the same transactions will have a maximum distance of 1 and rules that cover exactly the same transactions will have a distance of 0. While the intention of investigating the amount of commonly covered transactions is the same as with [Toivonen et al., 1995], both aforementioned problems are solved.

3.2.3 Extended Support Based Distance

Another distance measure inspired by the approach of [Toivonen et al., 1995] was introduced by [Sahar, 2002]. Let $r1 : \mathcal{A} \Rightarrow \mathcal{B}$ and $r2 : \mathcal{C} \Rightarrow \mathcal{D}$ be two association rules. Their distance is given by:

$$\begin{aligned} d(\mathcal{A} \Rightarrow \mathcal{B}, \mathcal{C} \Rightarrow \mathcal{D}) = {} & (1 + \mathit{diff}_{supp}(\mathcal{A}, \mathcal{C})) \cdot \frac{\|\mathcal{A} \otimes \mathcal{C}\|}{\|\mathcal{A} \cup \mathcal{C}\|} \cdot \gamma_1 + \\ & (1 + \mathit{diff}_{supp}(\mathcal{B}, \mathcal{D})) \cdot \frac{\|\mathcal{B} \otimes \mathcal{D}\|}{\|\mathcal{B} \cup \mathcal{D}\|} \cdot \gamma_2 + \\ & (1 + \mathit{diff}_{supp}(\mathcal{A} \cup \mathcal{B}, \mathcal{C} \cup \mathcal{D})) \cdot \frac{\|(\mathcal{A} \cup \mathcal{B}) \otimes (\mathcal{C} \cup \mathcal{D})\|}{\|\mathcal{A} \cup \mathcal{B} \cup \mathcal{C} \cup \mathcal{D}\|} \cdot \gamma_3 \end{aligned} \tag{3.9}$$

Where diff_{supp} is the distance measure of [Toivonen et al., 1995] generalized for arbitrary association rules and the $\otimes$ operator is the exclusive or of two sets given by:

$$\mathcal{A} \otimes \mathcal{B} = (\mathcal{A} \setminus \mathcal{B}) \cup (\mathcal{B} \setminus \mathcal{A}) \tag{3.10}$$

This distance measure compares the antecedents, the consequents and the complete rules with each other and calculates for all of those three a separate distance value. This is emphasized by dividing equation 3.9 in three sections, each dealing with one of the aforementioned parts of the rules. The single distances are then summed to gain the complete distance value. The single distance values in turn are also divided into three parts. The diff_{supp} values measure the amount of tuples from the database which are not covered by both of the given itemsets. The fraction part on the other hand does not compare the tuples which are covered but the amount of items in which the itemsets differ. Finally the last parts, denoted by the γ values, can be considered as weight factors. They are chosen in a way that itemsets, from which one is more general than the other, are preferred, that is their distance is decreased. For more details see [Sahar, 2002].

Obviously, this measure extends the approach of [Toivonen et al., 1995] in that it not only distinguishes between the single parts of a rule and their similarities to each other, but also takes other aspects like the similarity of itemsets into account. The downside is that this measure is not as easy to evaluate, since it doesn't satisfy all conditions of a metric anymore, although it maintains the symmetry property ([Sahar, 2002]).

3.2.4 Tightness

For the purpose of clustering association rules [Natarajan and Shekar, 2008] introduced a measure called tightness which quantifies the strength of binding between the items of an association rule. The idea is that certain items in an application domain might get bound together because they are so strongly related that they often occur together in transactions. This tightness of binding is not covered by traditional measures like support or confidence. Support on the one hand does not consider transactions that contain only some of the bound items and confidence only describes the predictive ability of a rule.

Let $x_1 x_2 \ldots x_m \rightarrow x_{m+1} \ldots x_n$ be an association rule and the x_i, $i = 1 \ldots n$ its respective items. Sx_i denotes the support of item x_i. Support values for the most and least frequent items of an association rule r are given by $S_{r_{max}} = max(Sx_i, \ldots, Sx_n)$ and $S_{r_{min}} = min(Sx_i, \ldots, Sx_n)$. An increase in support of a rule S_r tightens the binding between its constituent items, therefore an increase in S_r also increases the tightness measure. On the other hand consider the expression $\left(\frac{S_{r_{max}}+S_{r_{min}}}{2}\right) - S_r$. It roughly estimates the presence of items of r in other transactions that are not covered by r ([Natarajan and Shekar, 2008]). That is if this value increases, items of r occur more often separately in different transactions and therefore the tightness between those items decreases. Combining those two effects yields the tightness measure:

$$T = \frac{S_r}{\left(\frac{S_{r_{max}}+S_{r_{min}}}{2}\right) - S_r} \tag{3.11}$$

T reaches its maximum (i.e. ∞) when $S_{r_{max}} = S_{r_{min}} = S_r$. Based on the notion of tightness the following distance measure has been introduced:

$$d(r_i, r_j) = \frac{\left|T_{r_i} - T_{r_j}\right|}{T_{r_i} + T_{r_j}} \tag{3.12}$$

As [Natarajan and Shekar, 2008] have shown in the context of market basket analysis, this measure is able to discover similar purchasing behaviour in different item domains. That is, it groups association rules across several item domains. Since this characteristic could not be interpreted in a meaningful manner for the IDEAL tool it has not been used for it. It remains a task for future research to evaluate its usefulness for the IDEAL tool.

3.3 Measures introduced for IDEAL

3.3.1 Binary Distance

When it comes to looking for similar association rules, the most intuitive approach is to look at rules that are thematically linked. That is if a user has found an interesting rule, in order to further investigate it he will be looking for rules that cover the same items. The binary distance yields such a syntactical view on association rules and has been inspired by [Jorge, 2004].

Let $r : \mathcal{X} \Rightarrow \mathcal{Y}$ and $r' : \mathcal{X}' \Rightarrow \mathcal{Y}'$ be two association rules. The binary distance between those rules is the percentage of items that are only present in one of the association rules:

$$d(r, r') = 1 - \frac{|(\mathcal{X} \cup \mathcal{Y}) \cap (\mathcal{X}' \cup \mathcal{Y}')|}{|\mathcal{X} \cup \mathcal{Y} \cup \mathcal{X}' \cup \mathcal{Y}'|} \tag{3.13}$$

Rules that are comprised of the same items will have a distance of 0 while syntactically completely unrelated rules have a distance of 1. The first case might occur for rules which differ in their consequent. For example while the rules $\mathcal{X} \Rightarrow \mathcal{Y}$ and $\mathcal{Y} \Rightarrow \mathcal{X}$ have the same support values, their confidences might heavily differ. Such cases can easily be discovered using the binary distance. Another issue within the IDEAL tool is the presence of association rules that have no antecedent. Those rules basically consist of just one item and might cause problems for other syntax based distances that differ between antecedent and consequent of a rule. The binary distance however is robust for such cases. Consider the following examples. For three rules $r_1 : \{a, b\} \Rightarrow \{z\}$, $r_2 : \Rightarrow \{z\}$ and $r_3 : \{a, b, c\} \Rightarrow \{z\}$ the intuition would be that r_3 is more similar to r_1 than r_2 to r_1 because r_3, although it has more items than r_2, differs in less. The binary distance measure yields exactly this result: $d(r_1, r_2) = \frac{2}{3} > \frac{1}{4} = d(r_1, r_3)$. Another possible case is given by $r_1 : \{a, b\} \Rightarrow \{z\}$, $r_2 : \Rightarrow \{z\}$ and $r_3 : \{c, d\} \Rightarrow \{z\}$. This time the intuition would be that r_2 is more similar to r_1 than r_3 to r_1. Again the distance measure yields exactly this result: $d(r_1, r_3) = \frac{4}{5} > \frac{2}{3} = d(r_1, r_2)$.

The binary distance is a pseudo metric. **Proof:**

Isolation: Let $r : \mathcal{X} \Rightarrow \mathcal{Y}$ and $r' : \mathcal{Y} \Rightarrow \mathcal{X}$ be two association rules. Then obviously $d(r, r') = 0$ although $r \neq r'$, which violates the isolation criterion.

Symmetry: $r : \{x_1, x_2 \ldots x_n\} \Rightarrow \{y_1, y_2 \ldots y_m\}$ and $r' : \left\{x'_1, x'_2 \ldots x'_{n'}\right\} \Rightarrow \left\{y'_1, y'_2 \ldots y'_{m'}\right\}$ are two association rules:

$$\begin{aligned} d(r, r') &= 1 - \frac{\left|(\{x_1, x_2 \ldots x_n\} \cup \{y_1, y_2 \ldots y_m\}) \cap (\left\{x'_1, x'_2 \ldots x'_{n'}\right\} \cup \left\{y'_1, y'_2 \ldots y'_{m'}\right\})\right|}{\left|\{x_1, x_2 \ldots x_n\} \cup \{y_1, y_2 \ldots y_m\} \cup \left\{x'_1, x'_2 \ldots x'_{n'}\right\} \cup \left\{y'_1, y'_2 \ldots y'_{m'}\right\}\right|} \\ &= 1 - \frac{\left|(\left\{x'_1, x'_2 \ldots x'_{n'}\right\} \cup \left\{y'_1, y'_2 \ldots y'_{m'}\right\}) \cap (\{x_1, x_2 \ldots x_n\} \cup \{y_1, y_2 \ldots y_m\})\right|}{\left|\left\{x'_1, x'_2 \ldots x'_{n'}\right\} \cup \left\{y'_1, y'_2 \ldots y'_{m'}\right\} \cup \{x_1, x_2 \ldots x_n\} \cup \{y_1, y_2 \ldots y_m\}\right|} \\ &= d(r', r) \end{aligned}$$

Triangular Inequality: For reasons of brevity the proof of the triangular inequality will be omitted here but can be found in [Levandowsky and Winter, 1971].

3.3.2 Derivativeness based Distance Measures

The idea behind the next distance measures is to not look at the structure or syntax of a rule but to investigate the behaviour to find similarities. As noted before the behaviour

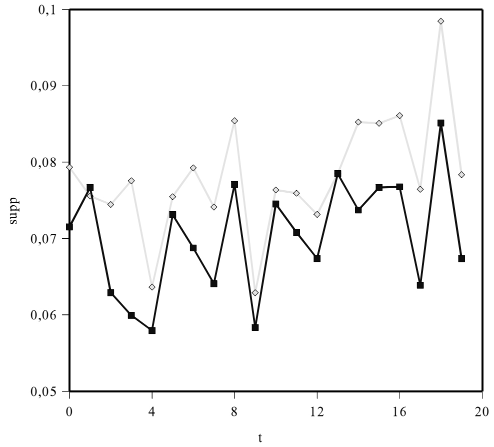


Figure 3.1: Two support histories with a linear derivativeness measure of 0.196

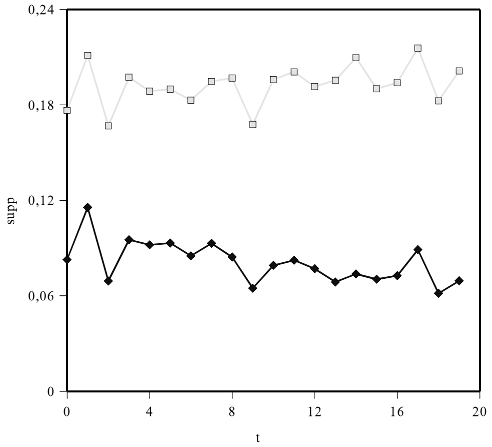


Figure 3.2: Two support histories with a relative change derivativeness of 0.045

of a rule is described by the histories of rule measures like support, confidence and lift. Rules that are linked with each other might show similar behaviour and therefore their histories might be correlated. To test for correlation the correlation coefficient is applied to two rule histories $H_1 : (v_1 \dots v_n)$ and $H_2 : (w_1 \dots w_n)$. The values of the histories are interpreted as the sample of random variables. The linear derivativeness is then given by:

$$d(H_1, H_2) = 1 - \left| \frac{\sum_{i=1}^{n} \left(v_i - \overline{H_1}\right) \left(w_i - \overline{H_2}\right)}{\sqrt{\sum_{i=1}^{n} \left(v_i - \overline{H_1}\right)^2} \sqrt{\sum_{i=1}^{n} \left(w_i - \overline{H_2}\right)^2}} \right| \tag{3.14}$$

This measure estimates how well a rule history can be derived from the history of another rule in terms of a linear relationship. Figure 3.1 shows two support histories with a linear derivativeness measure of 0.196. The linear derivativeness is neither a metric nor a pseudo metric. Isolation does not hold because $r : \mathcal{X} \Rightarrow \mathcal{Y}$ and $r' : \mathcal{Y} \Rightarrow \mathcal{X}$ show exactly the same support history although they are not the same rules. Additionally, despite the fact that the correlation coefficient is a symmetric measure and therefore symmetry for linear derivativeness can be proven, it has been experimentally shown that the triangular inequality does not hold for this measure.

Inspired by the test for derivative rules in [Böttcher, 2005] another test for derivativeness has been devised. For the relative change derivativeness not the histories per se are investigated for linear relationship but their relative changes. To this extend for each rule history $H : (v_1 \dots v_n)$ of size n, a history of size $n - 1$ comprised of its relative changes is computed. Each member of this new history can be calculated by:

$$v'_i = \frac{v_{i+1}}{v_i}, \ i = 1 \dots n - 1 \tag{3.15}$$

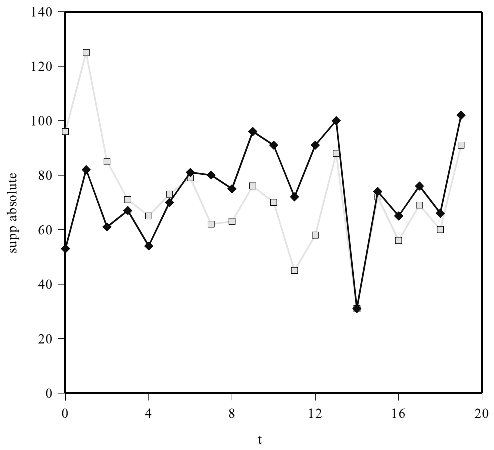


Figure 3.3: Two absolute support histories with an absolute change derivativeness of 0.081

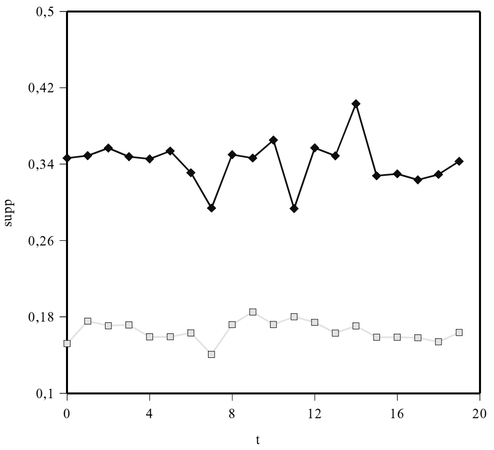


Figure 3.4: Two support histories with a deviation of 0.047 and a linear derivativeness of 0.643

To gain the relative change derivativeness for two rule histories the linear derivativeness of the transformed histories has to be calculated. Figure 3.2 shows two support histories with a relative change derivativeness measure of 0.045 but a linear derivativeness measure of 0.543. Again, the relative change derivativeness is neither a metric nor a pseudo metric because of the same arguments as for the linear derivativeness.

A third derivativeness measure examines the absolute changes of two rule histories. To this extend for each rule history $H : (v_1 \ldots v_n)$ of size n, a history of size $n-1$ comprised of its absolute changes is determined. The members of this new history are calculated by:

$$v'_i = v_{i+1} - v_i,\ i = 1 \ldots n-1 \tag{3.16}$$

Then, to gain the absolute change derivativeness for two rule histories the linear derivativeness of the transformed histories has to be calculated. This measure is intended to be used for histories of absolute values like absolute support values rather than relative value histories. For those histories absolute change correlations are more probable and easier to interpret. Figure 3.3 shows two absolute support histories with an absolute change derivativeness measure of 0.081 but a linear derivativeness measure of 0.59. As for the previous derivativeness measures this is neither a metric nor a pseudo metric because of the same arguments as for the linear derivativeness.

Experiments have shown that all three derivativeness measures yield significantly different rankings of association rules in terms of their distance to each other. However the specific relations between those measures and their exact interpretation regarding association rules remain a topic of future research. Furthermore might distance measures, which are looking for non linear correlations, lead to other interesting discoveries.

3.3.3 Scaled Deviation

The scaled deviation is another distance measure whose purpose is to analyze the behaviour of association rules. But this time it is not looking for linear relationship but for the history that matches another history with the least deviation. In order to make the deviation values comparable to each other, every history H: $(v_1 \ldots v_n)$ has to be scaled as explained in section 2.3:

$$v_i' = \frac{v_i}{\sum_{i=1}^{k} v_i} \tag{3.17}$$

The scaled deviation is now the sum of all value differences of the scaled histories:

$$d(H_1, H_2) = \sum_{i=1}^{n} \left| v_i' - w_i' \right| \tag{3.18}$$

The more similar two association rules are the smaller the distance will be. On the other hand the distance grows with the length of a history, which has to be considered when comparing the distance values. Figure 3.4 shows two support histories with a scaled deviation of 0.047 but a linear derivativeness measure of 0.643. The scaled deviation distance is a pseudo metric. Isolation does not hold because different rules can have the same histories. Symmetry obviously holds and the triangular inequality has been proven to hold for the Minkowski distance with $p = 1$ in [Kolmogorov and Fomin, 1999]. Again, experiments have shown that the scaled deviation yields different rankings of association rules in terms of their distance to each other than the linear derivativeness. The relations between those two measures also remains a topic of future research.

Chapter 4

PMML

The Predictive Model Markup Language is an XML based language that has been developed to provide a standardized format for describing statistical and data mining models [DMG, 2009a]. It can be used by a wide variety of applications as a means of either permanent storage or easy and intuitive sharing of data mining models between PMML compliant applications. Thus PMML enhances compatibility and interoperability of data mining tools. The first version of the standard was published in 1997 by the National Center for Data Mining and as of June 2009 the language has reached version 4.0 and is further developed by the Data Mining Group consortium.

The PMML version initially used by the association rule change mining tool IDEAL was 3.1. However at the time the further development, which is covered by this paper, began, the newest available version was 3.2. Therefore the tool itself and all concerned XML files, like the PMML schema description, have been adapted to comply to that version. The following sections will describe the overall structure of PMML 3.2 documents, focussing on the association rule model, as well as the extensions that have been made on order to incorporate rule change mining information.

4.1 The Predictive Model Markup Language 3.2

Every PMML document is an XML document with a root element of type `PMML` [DMG, 2009b]. This `PMML` element always contains the child elements `Header` and `DataDictionary` and may contain the additional elements `MiningBuildTask`, `TransformationDictionary` and one or several data mining models. A document without such a model can be used to carry the initial metadata that is available before an actual data mining model is computed. The models supported by version 3.2 are `AssociationModel`, `ClusteringModel`, `GeneralRegressionModel`, `MiningModel`, `NaiveBayesModel`, `NeuralNetwork`, `RegressionModel`, `RuleSetModel`, `SequenceModel`, `SupportVectorMachineModel`, `TextModel` and `TreeModel`. The general structure of a PMML document along with the multiplicities of its elements is given in figure 4.1.

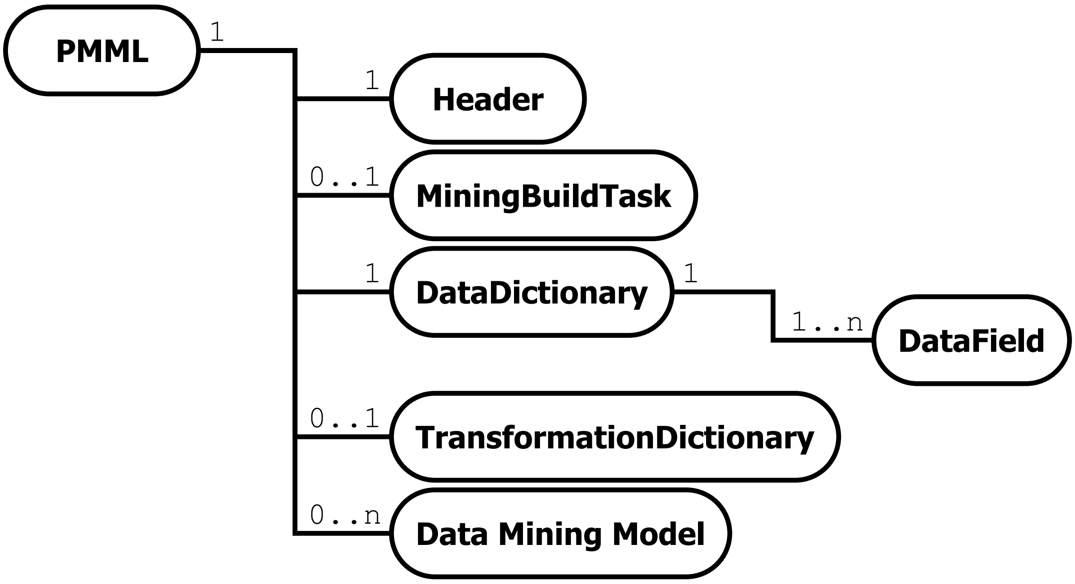


Figure 4.1: General structure of a PMML document

Header: The header contains a human readable description and copyright information for the contained data mining models. Additionally information about the application that created the model, like name and version, as well as a creation timestamp may be specified.

MiningBuildTask: The mining build task element contains the configuration of the training run that produced the model instance stored in the PMML document.

DataDictionary: The data dictionary contains definitions for data fields that are used in the data mining models. It specifies their types and value ranges.

TransformationDictionary: The transformation dictionary stores information about necessary transformations of data fields mentioned in the data dictionary, that have to be applied in order to use them for a data mining model. Those transformations could be comprised of discretizations, normalizations or aggregations.

The following is a simple PMML example that consists of just a header and a data dictionary containing the data fields that belong to the example introduced in chapter 2.

```
<?xml version="1.0" encoding="UTF-8"?>
<PMML xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="3.2"
 xsi:schemaLocation="http://www.dmg.org/PMML-3_2 file:/D:/pmml-3-2.xsd"
 xmlns="http://www.dmg.org/PMML-3_2">
    <Header copyright="BT Group plc">
        <Application name="IdealPrototype" version="1.1"/>
    </Header>
    <DataDictionary numberOfFields="20">
        <DataField dataType="string" name="Date" optype="categorical">
            <Value property="valid" value="Jan 2009"/>
            <Value property="valid" value="Feb 2009"/>
            <Value property="valid" value="Mar 2009"/>
        </DataField>
        <DataField dataType="string" name="Customer Type"
         optype="categorical">
            <Value property="valid" value="Residential"/>
            <Value property="valid" value="Business"/>
        </DataField>
        <DataField dataType="string" name="Location"
         optype="categorical">
            <Value property="valid" value="London"/>
            <Value property="valid" value="Ipswich"/>
            <Value property="valid" value="Norwich"/>
            <Value property="valid" value="Peterborough"/>
            <Value property="valid" value="Luton"/>
            <Value property="valid" value="Chelmsford"/>
        </DataField>
    </DataDictionary>
    <!-- ..any of the aforementioned models could be added here..  -->
</PMML>
```

Since association rules are the subject of this paper, only the corresponding model shall be explained next. Other models are defined completely different according to their special requirements and will be ignored further on. An association model describes association rules which might be obtained by a rule mining algorithm an its general structure is shown in figure 4.2.

AssociationModel: The association model element describes the general characteristics of the stored association rules via its attributes. Among those characteristics are the number of transactions from which the association rules were derived, that is the amount of database tuples that were analysed, as well as the maximum and average number of items in all transactions. Further attributes are the minimum support and confidence values satisfied by all stored rules and the maximum number of items of those rules. Finally

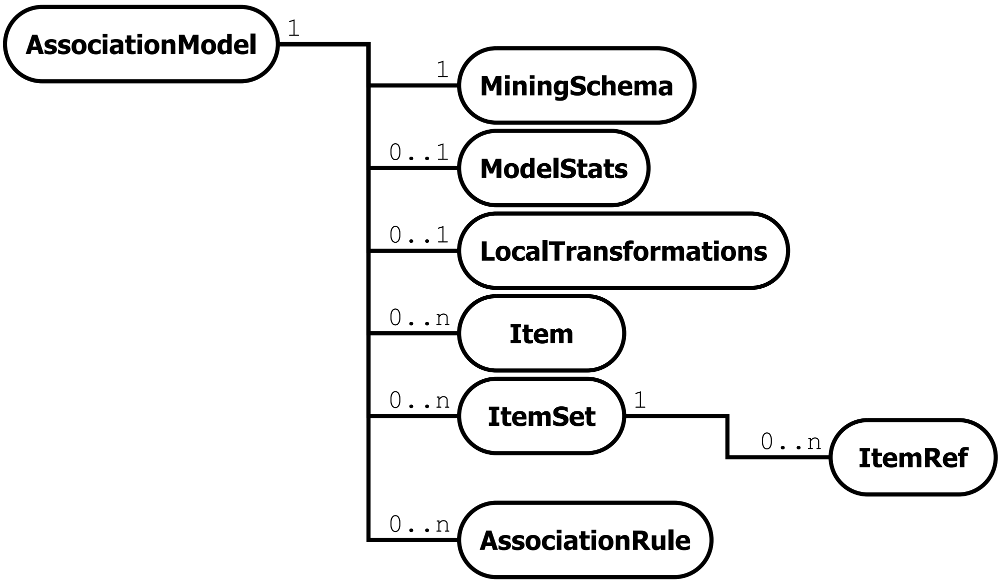


Figure 4.2: General structure of the association model element

the amount of items, item sets and association rules contained in the model are recorded.

MiningSchema: Every data mining model contains a mining schema. It defines the data fields used in each particular model and is a subset of the fields given in the data dictionary. The mining schema provides model specific information about the fields, like missing value treatment strategies, but its main purpose is to list those fields that a user has to provide in order to apply the model.

ModelStats: The model statistics element may contain general statistical information of either the complete model or the data fields mentioned in the mining schema.

LocalTransformations: This element stores information about transformations which have to be applied to the data fields given in the mining schema in order to use them for the current mining model. Unlike the transformations specified in the transformation dictionary, these are not required for other models stored in the PMML document.

Item: Each item contained in an association rule is stored in a single item element. It's main attributes are a unique identifier and the unique value the item has.

Itemset: All item sets that occur as either the antecedent or the consequent of an association rule are recorded by an item set element. Item sets are characterized by a unique identifier, the support value for the set and the number of items contained in the set. For each of those items the item set element has a child element called item reference, which contains an identifier referencing one item element.

AssociationRule: Finally all association rules are stored by association rule elements. Those elements contain references to the item sets for their antecedents and consequents as well as a unique identifier. Furthermore support, confidence and lift values are stored for each association rule.

An association model showcasing the simple example of chapter 2 is given next:

```
<AssociationModel algorithmName="Apriori" functionName="associationRules"
 lengthLimit="6" minimumConfidence="20" minimumSupport="5" modelName=
 "rule model 1" numberOfItems="6" numberOfItemsets="4" numberOfRules="2"
 numberOfTransactions="20">
    <MiningSchema>
        <MiningField name="Customer Type" usageType="active"/>
        <MiningField name="Location" usageType="active"/>
    </MiningSchema>
    <Item id="0" value="London"/>
    <Item id="1" value="Business"/>
    <Item id="2" value="Ipswich"/>
    <Item id="3" value="Residential"/>
    <Itemset id="0" numberOfItems="1" support="0.405">
        <ItemRef itemRef="0"/>
    </Itemset>
    <Itemset id="1" numberOfItems="1" support="0.304">
        <ItemRef itemRef="1"/>
    </Itemset>
    <Itemset id="2" numberOfItems="1" support="0.1">
        <ItemRef itemRef="2"/>
    </Itemset>
    <Itemset id="3" numberOfItems="1" support="0.696">
        <ItemRef itemRef="3"/>
    </Itemset>
    <AssociationRule id="0" antecedent="0" consequent="1" support="0.45"
     confidence="0.76" lift="2.5"/>
    <AssociationRule id="1" antecedent="2" consequent="3" support="0.05"
     confidence="0.9" lift="1.29"/>
</AssociationModel>
```

4.2 Extensions for Association Rule Change Mining

Since PMML 3.1, as well as 3.2, did not provide a model for time series data, several extensions and alterations had to be made in order to satisfy the requirements of the rule change mining scenario [Reichelt and Winkelmann, 2008]. The extensions are shown in figure 4.3 and were implemented through adaptions of the schema file that defines the PMML standard.

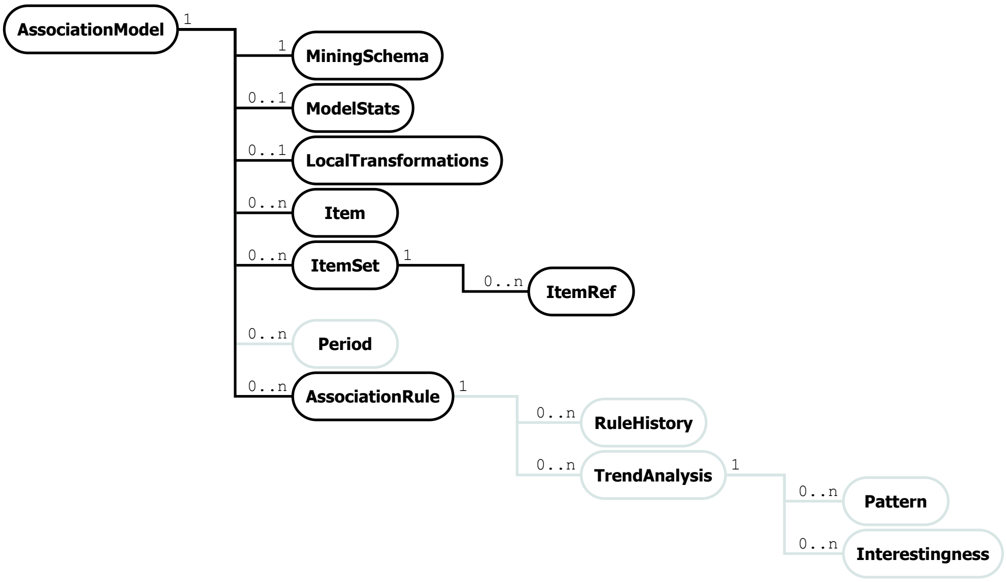


Figure 4.3: Extensions of the association model

Period: When association rules are analysed over time the temporal dimension becomes part of the association model and has to be incorporated into the PMML document. This is achieved by period elements which represent the temporal partitioning of the source data. Every period is characterized by a start date and an end date and is referred to by a unique identifier.

RuleHistory: When mining rule change information, measures like support, confidence or lift are no longer single values characterizing an association rule but become histories consisting of several values. Those are measured for each defined period and stored in a rule history element referring to the corresponding period element via its identifier. Measurements that might be recorded as attributes of a rule history element include support, confidence, lift, the absolute support value, the support of the antecedent and the absolute support of the antecedent.

TrendAnalysis: After trend analysis has been applied on the rule histories, information of all detected trends are stored in a trend analysis element, whereas each rule measure has its own such element. The specific trend characteristics are given in the child elements of type pattern and interestingness.

Pattern: This element stores the type of detected trends, e.g. whether they are stable or show upward or downward trend.

Interestingness: This element stores interestingness measures for detected trends by specifying the name of the measure and its value.

An example for an association rule whose measures have been taken for three different periods is given next:

```
<Period startDate="2009-06-01" endDate="2005-06-07" id="0"/>
<Period startDate="2009-06-08" endDate="2005-06-14" id="1"/>
<Period startDate="2009-06-15" endDate="2005-06-21" id="2"/>
<AssociationRule antecedent="0" consequent="1" id="0">
    <RuleHistory pid="0" conf="0.6519" lift="101.656" sabs="103"
     sabsB="158" sup="0.113" supB="0.174"/>
    <RuleHistory pid="1" conf="0.675" lift="107.175" sabs="108"
     sabsB="160" sup="0.115" supB="0.15795"/>
    <RuleHistory pid="2"conf="0.69697" lift="112.439" sabs="92"
     sabsB="132" sup="0.114" supB="0.12607"/>
    <TrendAnalysis measure="Support">
        <Pattern type="Trend-Stable"/>
        <Interestingness type="MEAN" value="0.114"/>
    </TrendAnalysis>
</AssociationRule>
```

Additional minor changes that haven been conducted include the removal of the attribute support for the element item set, because like association rules, item sets are no longer characterized by a single value. Furthermore an attribute with the name attribute has been added to the element item, to satisfy the fact that an item is defined by its data table attribute, like Location, and value, like Ipswich [1].

4.3 Extensions for Association Rule Comparisons

In order to contain information obtained by association rule comparisons, further extensions to the association model had to be implemented. Comparison results available through the IDEAL tool, and therefore covered by the extensions, include rule history redundancies and rule distance measures. For both, new XML elements, containing the necessary data, were introduced, as shown in figure 4.4.

Rule history redundancy information was added as a new optional child element of the trend analysis element which in turn was renamed history analysis. The renaming was done because redundancy is not an intrinsic property of trends and the introduction of an additional rule measure related element was deemed unnecessary and eventually confusing. Of course the renaming also implies a slight semantic change for the remaining child elements of type pattern and interestingness. Previously both were strictly

[1] Those two conclusions were drawn by the author, since they were not documented in [Reichelt and Winkelmann, 2008].

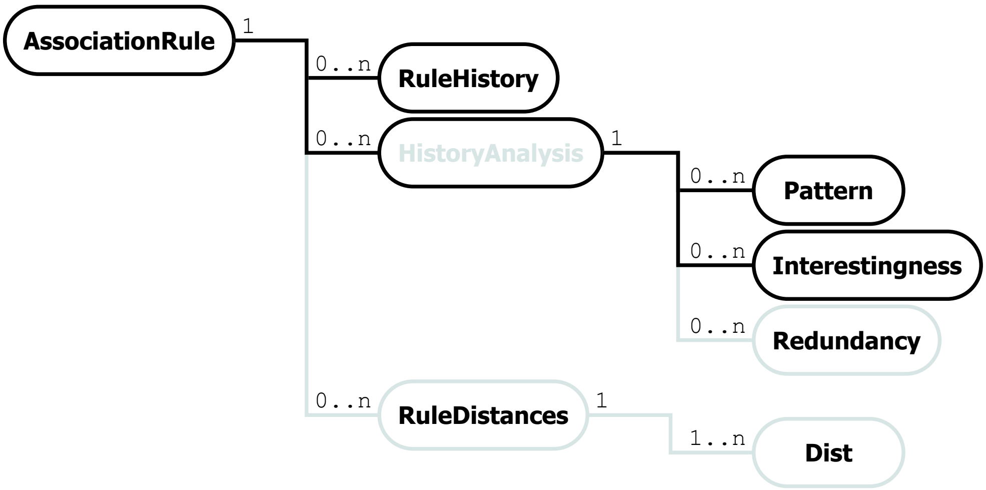


Figure 4.4: Further extensions of the association model

related to trends whereas now they are considered properties of rule histories in general. Pattern, for example, might now include any irregular shape of a history and even more important, interestingness values can now be assigned to every rule history, regardless of a prevalent trend. Both semantic changes are intuitive and reasonable. The actual redundancy information then is just a reference to the association rule to whose history the redundancy applies. When there are more than one such rule, for every of those rules one redundancy element hast to be added.

Rule distances are incorporated as optional child elements of the association rule element. For every rule distance measure that shall be stored, a new rule distances element has to be added, which carries the name of the measure as an attribute. The single distance values, according to the distance measure, are then given as child elements of type `dist`, from which at least one has to be specified. Each distance element states the value of the distance and the identifier of the association rule to which the distance applies. The distance elements have to be given in ascending order without omissions, that is, the least distant association rule has always to be specified and after that the second closest, then the third closest and so on. That way, if k of those distances are given, they are equivalent with the k least distant respectively most similar rules, according to the distance measure specified in the rule distances element. It has to be noted though, that storing the distance values this way, several, if not all might be given twice. This can be avoided when using other structures, for instance a triangular matrix, where row and column number correspond to association rule identifier and the value of row i and column j is the distance between the rule with identifier i and the rule with the identifier j. Such a structure however, yields several problems. When being dense, that is a all possible distances are given, a lot of hard disc space is required. Consider an association model with 5000 rules, then 12.502.500 distance values had to be given in a triangular matrix. For simplicity further assume each distance value consists of an integer which requires 1 Byte storage space. In total 12MB for one distance measure would be required which is a very optimistic calculation. In practice distance measures are real valued and a measure with just 3 decimal digits (e.g. 0.123) needs at least 5 Byte storage space [2] resulting in 60MB, again for just one distance measure. In the end model sizes of several hundred megabytes would be the outcome. That, while still feasible, would quite frankly not be reasonable, since PMML is intended to enable easy sharing of data mining models - a feature that would be hampered by such large amounts of data. Let alone the fact that most distance values given might be to no interest to the user at all, since normally he is only interested in the k least distant rules. Another possibility would be a not densely populated matrix for storing only those k least distant values. While this approach, compared to the actually used one, even reduces the amount of total stored data [3], it increases the complexity of the structure and therefore decreases comprehensibility of the model, which is not desirable. Additionally no ordering is provided, a property that would have to be reestablished through additional computing. So eventually the chosen structure is a trade off between reasonable storage size and model comprehensibility.

A small example storing the three least distant rules for one distance measure is given next.

```
<AssociationRule antecedent="0" consequent="1" id="0">
    <RuleHistory pid="0" conf="0.6519" lift="101.656" sabs="103"
     sabsB="158" sup="0.113" supB="0.174"/>
    <RuleHistory pid="1" conf="0.675" lift="107.175" sabs="108"
     sabsB="160" sup="0.115" supB="0.15795"/>
    <RuleHistory pid="2"conf="0.69697" lift="112.439" sabs="92"
     sabsB="132" sup="0.114" supB="0.12607"/>
    <HistoryAnalysis measure="Support">
        <Pattern type="Trend-Stable"/>
        <Interestingness type="MEAN" value="0.114"/>
        <Redundancy rid="1"/>
    </HistoryAnalysis>
    <RuleDistances measure="EuclideanDistance">
        <Dist d="0" rid="1"/>
        <Dist d="0.2" rid="2"/>
        <Dist d="0.5" rid="3"/>
    </RuleDistances>
</AssociationRule>
```

4.4 The Predictive Model Markup Language 4.0

In June 2009 version 4.0 of the Predictive Model Markup Language was released, which incorporated several changes and additions that made some of the previously introduced

[2] assuming UTF-8 encoding

[3] because every value is just given once

extensions obsolete. Due to time constraints they could not be integrated in the current version of IDEAL. None the less, those parts that are of interest to the association rule change mining process shall be discussed briefly.

The most notable addition of the updated version is the new time series model, whose general structure is shown in figure 4.5.

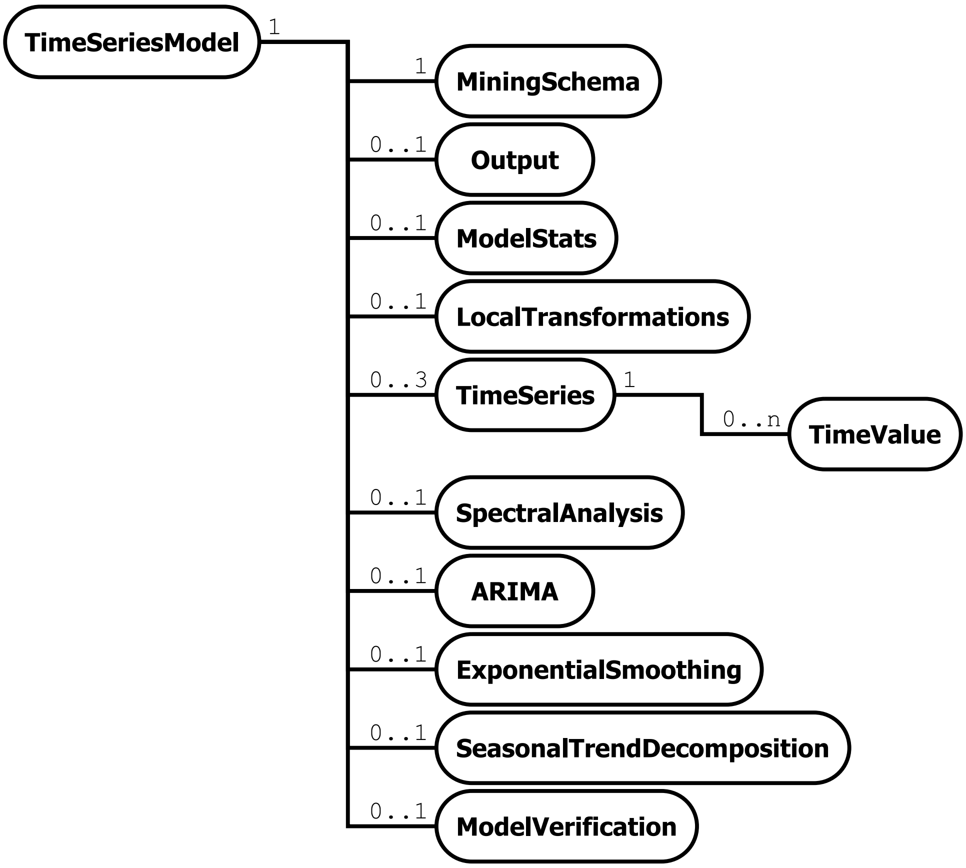


Figure 4.5: General structure of the time series model

This model can be used to incorporate some of the time series data that was previously contained in the association rule elements. Thereto the histories of each association rule have to represented by their own time series model, which, for some undocumented reason, can only store three of them through their time series elements [4]. This amount though, can easily be adapted to fit the number of rule measure histories that shall be stored for each association rule. All resulting models have to be assigned to their corresponding association rules. The PMML standard provides a model name attribute that could be used for that purpose, by specifying the identifier for the corresponding

[4] This might be explained by the definition of three different values for the time series element' s usage attribute, but this is pure speculation by the author.

association rule. Furthermore each time series element needs a reference to the rule measure history it represents. Since the time series element does not provide a fitting attribute, this has to be added. Two attributes that do exist in the standard are start time and end time which can be used for the start and end dates of the analysed time frame. Each value of a history has then to be stored in a time value element, which is a child element of the time series element. In addition to a value attribute this element also has a time attribute. Since data points in rule measure histories are characterized by a time period and not a single point in time, this attribute is not sufficient for the rule change mining scenario and has to be complemented by at least one additional time related attribute. Further history specific information covered previously include trend pattern, interestingness measures and redundancies. While trends are represented in the new time series model, they are related to exponential smoothing and therefore can not intuitively be used in the sense they were introduced here. Hence extensions for those kind of trends as well as for interestingness values and redundancies have to be implemented in order to use the time series model in a association rule change mining context.

Clearly, the time series model was not designed with association rule histories in mind, thus a lot of changes have to be implemented to satisfy their requirements. Eventually it is up to future research to decide whether to use the capabilities introduced in PMML 4.0 or to keep the extensions explained in section 4.2 and 4.3.

Chapter 5

IDEAL

The framework and techniques introduced in [Böttcher, 2005] led to the development of a research prototype for intelligent data exploration and learning - IDEAL. Its purpose was to continuously analyze changes in corporate data without the need of user interaction to detect emerging problems and interesting developments that could be exploited. To achieve this goal association rule change mining was used to analyze the current behaviour of the data. The first prototype has been redesigned and further extended by [Reichelt and Winkelmann, 2008]. The following sections will show the general work flow of the IDEAL tool and outline additional extensions that have been implemented to assist the user in analyzing the data. For detailed information on the implementation see [Reichelt and Winkelmann, 2008].

5.1 General Work Flow of the IDEAL Tool

The basic functions of the IDEAL tool are divided into three layers - the Mining Layer, the Detection Layer and the Evaluation Layer ([Böttcher, 2005]). The task of the Mining Layer is to discover association rules and to store and manage their histories. To this extend the Mining Layer takes the data from a timestamped database and for user defined periods determines all association rules that satisfy given minimum support and confidence values. The discovered rules and their histories are then either stored in another database or are converted into an interchange format like PMML. The Detection Layer analyzes the discovered association rules and tests whether their histories show any change patterns like trends or stability. To enhance the reliability of the detection optionally noise reduction techniques may be applied to the histories. Finally the Evaluation Layer takes those histories that show change patterns and checks whether redundancies can be identified among them. Rules with redundant histories may either be marked as such or be immediately discarded due to the rule quantity problem. The remaining rules are then tested for their interestingness, whereas the measures introduced in section 2.3 are used. An additional task that has been assigned to this layer is the calculation of a user defined subset of distance measures introduced in section 3.3.

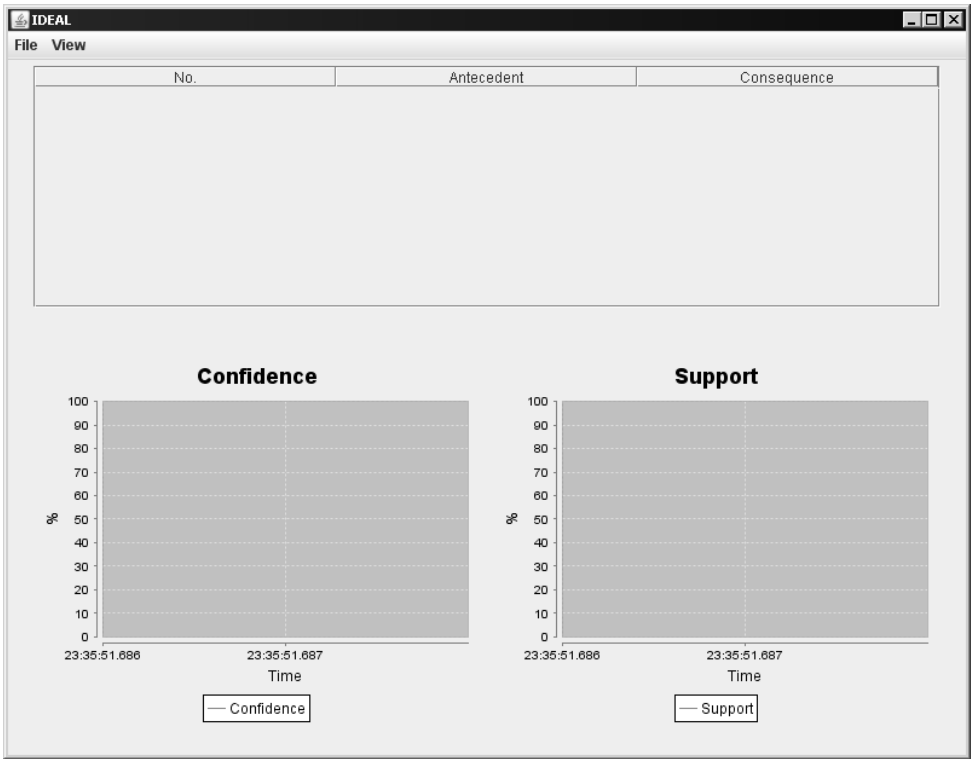
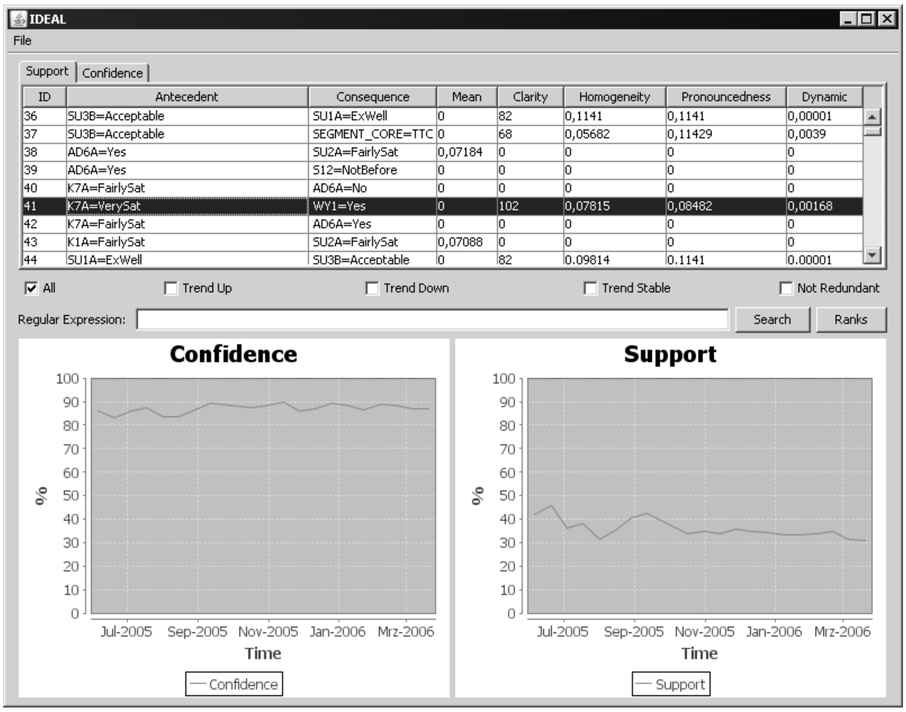


Figure 5.1: The IDEAL tool after the redesign of [Reichelt and Winkelmann, 2008]

5.2 Functions of the IDEAL Standalone Version

The standalone version of IDEAL was implemented in Java and is a completely autonomous software with its own SQL based database access and its own GUI. Any timestamped relational database can be used for association rule mining just by specifying the column that holds the time data. Association rule mining can be influenced by declaring minimum and maximum support values as well as a minimum confidence. Additionally the minimum and maximum size for an association rule can be specified. The tool has been designed to support several association rule mining algorithms but as of now only Apriori was implemented. Further parameters are the size for the histories which have to be established and the length of each period in which a history value is determined. After specifying the parameters the tool follows the work flow explained in the previous section and presents the results. The GUI can be seen in figure 5.1. The mined rules will be shown in a table giving the items the rules are comprised of. By clicking on a rule its support and confidence histories will be shown in two diagrams below the table. Via the View menu it is possible to filter the rules inside the table to show only those that show a certain change pattern in their support or confidence histories. The mined rules then can be saved in a PMML file as specified in section 4.2. Previously saved rules can be loaded and further investigated without mining them again.

5.3 Extensions For The Analysis of Association Rules



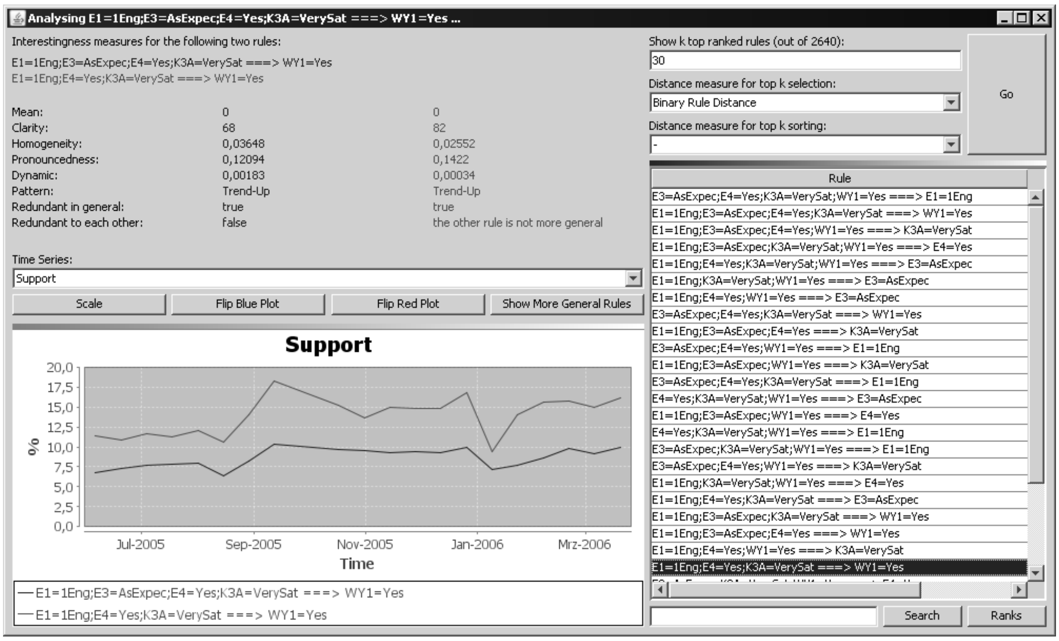
ID	Antecedent	Consequence	Mean	Clarity	Homogeneity	Pronouncedness	Dynamic
36	SU3B=Acceptable	SU1A=ExWell	0	82	0,1141	0,1141	0,00001
37	SU3B=Acceptable	SEGMENT_CORE=TTC	0	68	0,05682	0,11429	0,0039
38	AD6A=Yes	SU2A=FairlySat	0,07184	0	0	0	0
39	AD6A=Yes	S12=NotBefore	0	0	0	0	0
40	K7A=FairlySat	AD6A=No	0	0	0	0	0
41	K7A=VerySat	WY1=Yes	0	102	0,07815	0,08482	0,00168
42	K7A=FairlySat	AD6A=Yes	0	0	0	0	0
43	K1A=FairlySat	SU2A=FairlySat	0,07088	0	0	0	0
44	SU1A=ExWell	SU3B=Acceptable	0	82	0.09814	0.1141	0.00001

☑ All ☐ Trend Up ☐ Trend Down ☐ Trend Stable ☐ Not Redundant

Regular Expression:

Figure 5.2: The extended IDEAL tool

While the first implementation was a first step in helping the user to deal with the massive amount of association rules that are gained through association rule mining, it was hardly more than a pure presentation of the rules which didn't allow much for further investigations. To overcome this detriment the extensions that will be detailed in the following have been implemented. The new GUI can be seen in figure 5.2. The first difference that comes to mind is that the table now also shows the interestingness values. Since it is assumed that the user, for a first orientation, will focus on support and confidence histories, the interestingness values for only those two are shown here and can be selected using the sliders above the table. The rules can be sorted for one or several of their interestingness values and through the rank button, instead of the values, the position of a rule in a ranking according to the interestingness values will be shown. This allows even inexperienced users to judge those values. The next difference is that the filter for rules that show a change pattern were moved below the table and now have the form of check boxes. This way the filter can be combined any way the user likes, for example to find rules that show several change patterns. An additional filter is also available to show only those rules that are not redundant. Beneath those filters an input field has been added, which allows to search the table for specific rules using regular expressions. All this additions have been made to assist the user in finding a rule he is interested in. When he has found one such rule he now can even further investigate it by double clicking on it, which will open a new window shown in figure 5.3.



Analysing E1=1Eng;E3=AsExpec;E4=Yes;K3A=VerySat ===> WY1=Yes ...

Interestingness measures for the following two rules:

E1=1Eng;E3=AsExpec;E4=Yes;K3A=VerySat ===> WY1=Yes
E1=1Eng;E4=Yes;K3A=VerySat ===> WY1=Yes

Mean:	0	0
Clarity:	68	82
Homogeneity:	0,03648	0,02552
Pronouncedness:	0,12094	0,1422
Dynamic:	0,00183	0,00034
Pattern:	Trend-Up	Trend-Up
Redundant in general:	true	true
Redundant to each other:	false	the other rule is not more general

Time Series: Support

Show k top ranked rules (out of 2640): 30

Distance measure for top k selection: Binary Rule Distance

Distance measure for top k sorting: -

Rule
E3=AsExpec;E4=Yes;K3A=VerySat;WY1=Yes ===> E1=1Eng
E1=1Eng;E3=AsExpec;E4=Yes;K3A=VerySat ===> WY1=Yes
E1=1Eng;E3=AsExpec;E4=Yes;WY1=Yes ===> K3A=VerySat
E1=1Eng;E3=AsExpec;K3A=VerySat;WY1=Yes ===> E4=Yes
E1=1Eng;E4=Yes;K3A=VerySat;WY1=Yes ===> E3=AsExpec
E1=1Eng;K3A=VerySat;WY1=Yes ===> E3=AsExpec
E1=1Eng;E4=Yes;WY1=Yes ===> E3=AsExpec
E3=AsExpec;E4=Yes;K3A=VerySat ===> WY1=Yes
E1=1Eng;E3=AsExpec;E4=Yes ===> K3A=VerySat
E3=AsExpec;E4=Yes;WY1=Yes ===> E1=1Eng
E1=1Eng;E3=AsExpec;WY1=Yes ===> K3A=VerySat
E3=AsExpec;E4=Yes;K3A=VerySat ===> E1=1Eng
E4=Yes;K3A=VerySat;WY1=Yes ===> E3=AsExpec
E1=1Eng;E3=AsExpec;WY1=Yes ===> E4=Yes
E4=Yes;K3A=VerySat;WY1=Yes ===> E1=1Eng
E3=AsExpec;K3A=VerySat;WY1=Yes ===> E1=1Eng
E3=AsExpec;E4=Yes;WY1=Yes ===> K3A=VerySat
E1=1Eng;K3A=VerySat;WY1=Yes ===> E4=Yes
E1=1Eng;E4=Yes;K3A=VerySat ===> E3=AsExpec
E1=1Eng;E3=AsExpec;K3A=VerySat ===> WY1=Yes
E1=1Eng;E3=AsExpec;E4=Yes ===> WY1=Yes
E1=1Eng;E4=Yes;WY1=Yes ===> K3A=VerySat
E1=1Eng;E4=Yes;K3A=VerySat ===> WY1=Yes

Figure 5.3: The investigation window for an association rule

This window is divided into three parts. The right part contains another table which shows for the just double clicked rule the k most similar rules according to the chosen distance measure. Those rules can then be sorted by another distance measure to look for possible different orderings. As with the previous window, the table in this window can be searched with regular expressions and instead of the distance value the corresponding ranks can be shown. The upper left part of the window shows details for the investigated rule and the currently chosen of the rules on the right, like interestingness values, change patterns and redundancies. On the lower left part the histories of those two rules can be viewed and examined by scaling or flipping them. There also is a button for showing the more general rules. This is useful when a rule has an interesting homogeneity value, because then the reason for this can be determined. Should a user find another rule he is interested in, double clicking on it opens another investigation window. The last feature in this window is the fact that the diagram and table are dockable, that is if one of the two is too small to show all information they can be removed from the window in their own subwindow for further investigation.

Other new features of the tool are hidden in the settings dialog. Here, for example, database tables can be specified which give decoded descriptions for cryptic item names. That is if the items of the rules are not named in a human readable manner, by specifying the descriptions mouse rollovers on one of the tables result in the display of the explanations.

As outlined in section 4.3, PMML files can store information on distance measures. A new option allows to control this storage by specifying the amount of most similar rules that shall be stored.

Chapter 6

Conclusion

6.1 Summary

To analyze large amounts of data association rules are one versatile tool. However, the rule quantity problem hampers the discovery of interesting and relevant rules. The use of distance measures has been motivated as one possibility to alleviate this task. To this extend the binary distance has been presented to capture structural similarities of association rules. To find hidden relations among those rules three derivativeness measures have been introduced. They compare the rule histories and check for linear dependencies. Another distance measure, the scaled deviation, has a similar purpose in that it defines dissimilarity as the actual deviation two rule histories have. Further on the PMML standard has been presented together with proposed extensions for the inclusion of distance measures between association rules. Those extensions as well as a plethora of other features then have been implemented in the IDEAL tool. The goal was to support the user in finding relevant rules. While this goal has been achieved, a lot of future work still remains, as outlined in the next section.

6.2 Future Work

In this thesis the introduction of distance measures for the purpose of assistance in finding relevant rules has been motivated. The question however, why the three derivativeness distances in some cases yield significantly different rankings still has to be answered.
As in section 4.4 explained, the release of PMML version 4.0 raises several questions as how the new time series model can be used to convey history data. This also has to be investigated in the future.

One of the goals of the IDEAL tool was to somehow automatically guide the user to rules he is interested in. Although the investigation possibilities have clearly been expanded, the search for interesting rules still requires a lot of work. While the proposed distance measures were implemented in IDEAL, their representation as rankings inside a table can be tedious to analyze. A graphical representation using clustering techniques might further ease the task of examining the association rules. How any kind of automation can be achieved, remains a topic for future research.

Bibliography

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., and Swami, A. N. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., United States.

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases*, pages 487–499, Santiago de Chile, Chile.

[Brin et al., 1997] Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 255–264, Tucson, Arizona, USA.

[Böttcher, 2005] Böttcher, M. (2005). Discovering interesting temporal changes in association rules. Master's thesis, Otto-von-Guericke-Universität Magdeburg.

[DMG, 2009a] DMG (2009a). Frequently asked questions about pmml. The Data Mining Group Consortium Project Homepage: `http://www.dmg.org/faq.html`. Last accessed 30.07.2009.

[DMG, 2009b] DMG (2009b). Pmml 3.2 - general structure of a pmml document. The Data Mining Group Consortium Project Homepage: `http://www.dmg.org/v3-2/GeneralStructure.html`. Last accessed 30.07.2009.

[Gupta et al., 1999] Gupta, G. K., Strehl, A., and Ghosh, J. (1999). Distance based clustering of association rules. In *Proceedings of ANNIE 1999 (Intelligent Engineering Systems Through Artificial Neural Networks)*, pages 759–764, St. Louis, Missouri, USA.

[Han et al., 2004] Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1):53–87.

[Jorge, 2004] Jorge, A. (2004). Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In *Proceedings of the Fourth SIAM International Conference on Data Mining*, pages 178–187, Lake Buena Vista, Florida, USA.

[Kolmogorov and Fomin, 1999] Kolmogorov, A. N. and Fomin, S. V. (1999). *Elements of the Theory of Functions and Functional Analysis*. Dover Publications, Inc.

[Levandowsky and Winter, 1971] Levandowsky, M. and Winter, D. (1971). Distance between sets. *Nature*, 234:34–35.

[Li et al., 2004] Li, J., Shen, H., and Topor, R. W. (2004). Mining informative rule set for prediction. *Journal of Intelligent Information Systems*, 22(2):155–174.

[Mann, 1945] Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, 13(3):245–259.

[Montgomery and Runger, 2003] Montgomery, D. C. and Runger, G. C. (2003). *Applied Statistics and Probability for Engineers*. John Wiley & Sons, Inc., 3rd edition.

[Natarajan and Shekar, 2008] Natarajan, R. and Shekar, B. (2008). Tightness: A novel heuristic and a clustering mechanism to improve the interpretation of association rules. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*, pages 308–313, Las Vegas, Nevada, USA.

[Reichelt and Winkelmann, 2008] Reichelt, M. and Winkelmann, T. (2008). Ideal 2 ican - integration and extension of a temporal association-rule-mining-tool. Master's thesis, Fachhochschule Braunschweig/Wolfenbüttel.

[Sahar, 2002] Sahar, S. (2002). Exploring interestingness through clustering: A framework. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002)*, pages 677–680, Maebashi City, Japan.

[Toivonen et al., 1995] Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., and Mannila, H. (1995). Pruning and grouping discovered association rules. In *ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases*, pages 47–52, Greece.

[Zaki, 2000] Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390.

List of Figures