Discovering Interesting Temporal Changes in Association Rules

— Diplomarbeit —

vorgelegt von

cand. inf. Mirko Böttcher

mirko.boettcher@student.uni-magdeburg.de

24. April 2005

Erster Gutachter : Prof. Dr. Rudolf Kruse

Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Institut für Wissens- und Sprachverarbeitung
Universitätsplatz 2
D-39106 Magdeburg

Zweiter Gutachter : PD Dr. habil. Detlef Nauck

Intelligent Systems Research Centre
British Telecom Research & Venturing
Adastral Park, Martlesham
Ipswich IP5 3RE, UK

Abstract

Many businesses collect huge volumes of data. Commonly this data is continously gathered over long periods and therefore reflects changes in the domain from which it has been derived. To control their business operations and to gain a competitive edge it is crucial for companies to detect these changes. However, existing data mining approaches are not developed for this task. They typically assume that the domain under consideration is stable over time.

As a response to the lack of methods, some research in *rule change mining* has recently emerged. The idea of rule change mining is to analyse the temporal development of the data's underlying structure by discovering change patterns in the statistics of association rules derived from it.

This thesis reviews approaches to rule change mining and shows their deficiencies, particularly with respect to the requirements of business applications. The main achievement of this thesis is the development of a framework of methods which, for the first time, accounts for all the tasks linked to rule change mining: the discovery and long-term storage of rules, the detection of change patterns and their automatic post-processing in order to identify the most relevant and interesting ones. In particular, this thesis proposes novel methods for change pattern detection, pruning and interestingness assessment which are far superior to existing approaches in terms of their reliability and interpretability.

Experiments on real-life data taken from surveys show that the framework is very effective in discovering novel and useful change patterns which significantly help to understand the developments in the underlying domain.

Kurzfassung

Viele Unternehmen sammeln große Mengen an Daten. Diese werden zumeist kontinuierlich über lange Zeiträume erfaßt und reflektieren daher Änderungen in der ihnen zugrundeliegenden Domäne. Um ihre Geschäftsoperationen zu kontrollieren, aber auch um Wettbewerbsvorteile zu gewinnen, ist es für Unternehmen unabdingbar, derartige Änderungen zu entdecken. Vorhandene Verfahren des Data Mining sind für diese Aufgabe jedoch ungeeignet, da sie zumeist auf der Annahme basieren, daß die betrachtete Problemdomäne zeitlich invariant ist.

Als eine Antwort auf diesen Mangel an Methoden wurden in der letzten Zeit einige Verfahren zur Analyse von Regeländerungen entwickelt. Derartige Ansätze versuchen interessante Muster in der zeitlichen Entwicklung statistischer Eigenschaften von Assoziationsregeln zu entdecken. Diese lassen dann Rückschlüsse auf Änderungen in den ihnen zugrundeliegenden Daten zu.

Diese Arbeit stellt bestehende Ansätze zur Analyse von Regeländerungen vor und zeigt deren Defizite im Hinblick auf den Einsatz in der Praxis. Das wesentliche Ergebnis dieser Arbeit besteht in der Entwicklung eines Systems, welches zum ersten Mal alle mit der Analyse von Regeländerungen verbundenen Aufgaben berücksichtigt. Dies sind die Entdeckung und langfristige Speicherung von Assoziationsregeln, die Entdeckung von Änderungsmustern und deren automatische Bewertung, um die interessantesten unter ihnen zu identifizieren. Insbesondere stellt diese Arbeit neue Verfahren zur Entdeckung von Änderungsmustern, zum Erkennen redundanter Änderungen und zur Interessantheitsbewertung vor, die den bisherigen Verfahren hinsichtlich ihrer Verständlichkeit und Zuverlässigkeit deutlich überlegen sind.

Anhand der Analyse von Umfragedaten wird gezeigt, daß das vorgeschlagene System sehr wirkungsvoll in der Entdeckung zuvor unbekannter und überraschender Änderungen ist.

Acknowledgments

I'd like to thank all the people who were involved in this work. Special thanks to Prof. Dr. Rudolf Kruse for his support, and for the opportunity to work in the excellent and stimulating environment of the Intelligent Systems Research Centre of British Telecom. Special thanks also to my supervisor at British Telecom, Dr. Detlef Nauck, for giving me the freedom to pursue the project I proposed and to give guidance where necessary. Special thanks also to my supervisor at university, Dr. Christian Borgelt, for carefully proofreading drafts of this thesis and for many valuable comments. I am deeply grateful to Dr. Martin Spott for many inspiring discussions which substantially helped me to shape my ideas, for reviewing drafts of this document and for co-authoring my first submission to a conference.

I would like to express my gratitude to the following people for their support and assistance: Dr. Dymitr Ruta for reviewing early drafts and for many fruitful discussions; Simon B. Beech for correcting grammatical and spelling errors; Fabian Wickborn, Lothar Schlesier, Dani and Rainer Habrecht for valuable comments and suggestions.

Thanks to all the girls and guys with whom I spent my time in England. I had a great time there and realised that many other qualities of life in England compensate for the cold, wet weather and the 'exotic' food!

Last but not least, special thanks to my parents for supporting me throughout my studies at the university! Thanks also to all my friends with whom I spent my time here during my studies.

“To understand is to perceive patterns.”

Sir Isaiah Berlin

“The key in business is to know something that nobody else knows.”

Aristotele Onassis

Contents

Chapter 1

Introduction

Triggered by astonishing improvements in information and communication technology, the last decade heralded the change from an industrial to an information age. Mass customisation replaced the mass production of the industrial age and knowledge became the key driving economic input (Tjaden, 1996). Since the world with its markets and innovations is changing faster than ever before, the key to survival in the information age is the ability to detect, assess and respond to new trends and events – rapidly and intelligently. Discovering new information and using it before others do has become a strategical issue (Tjaden, 1997). As Don Keough, the former president of Coca-Cola, pointed out "Who[ever] has information fastest and uses it wins" (Watterson, 1995). Therefore it is crucial for companies to collect and stockpile data about processes, markets, customers and products in order to find new information.

As a response to those trends, significant research into developing processes for automatic information gathering is currently being conducted, from which the term *data mining* has been coined. The main goal of data mining is to find information in data that is interesting, novel and potentially useful (Fayyad et al., 1996). It utilises and combines methods from areas like statistics and artificial intelligence as well as from soft computing and machine learning. Cluster and regression analysis, Bayesian networks, decision trees and association rules are just a few examples.

Most methods assume implicitly that the domain under consideration is stable over time and thus provide a rather static view on the underlying structure of gathered data. This is undesirable in timestamped domains, since the data then captures and reflects external influences like management decisions, economic and market trends and changes in customer behaviour. Methods should account for this dynamic behaviour and hence consider the data as a sequence along the time axis. Such domains are very common in practice, since data is almost always collected over long periods. Or, as (Kimball, 1996) noted: "The time dimension is the one dimension virtually guaranteed to be present in every data warehouse, because virtually every data warehouse is a time series".

From this perspective the question *Which patterns exist?* is replaced by *How do patterns change?* Systematic pattern change is a pattern itself and obviously of high interest in the decision making process: its predictive nature allows for proactive business decisions, which are hardly possible with static approaches (Dong et al., 2003). In fact, the detection of interesting and previously unknown change patterns – which this thesis focuses on – not only allows the user to monitor the impact of past business decisions but also to prepare today's business for tomorrow's needs.

1.1 Temporal Change of Association Rules

Given a sequence of timestamped data, the traditional way to measure changes in a domain is to calculate indicators in regular intervals and to analyse the resulting sequences by means of time series analysis. Indicators are typically intended to monitor predefined business goals. The range of observable changes is therefore significantly limited by the user's expectations about its business – many other interesting changes may remain unrevealed. For example, if the business goal is to decrease the customer dissatisfaction the indicator could be chosen as the fraction of dissatisfied customers. However, changes within smaller populations, for example, male customers who use broadband, cannot be detected. In general, the obstacle connected with this type of analysis is the user's lack of knowledge about many interesting co-occurring attribute values.

This issue can be solved by *association rule mining* (Agrawal et al., 1993), which was originally developed for market basket data analysis, where each transaction consists of a set of purchased items. Its goal is to detect all those items which frequently occur together and to form rules which predict the co-occurrence of items. Nonetheless, association rule mining is not just bound to this specific purpose: it can be applied to every relational database. An example of an association rule might be that 5% of males who use broadband are dissatisfied customers. The attached probability is called the confidence of the rule, while the fraction of transactions that contain all three items (male, broadband user, dissatisfied) is called support. The advantage of association rule mining is the completeness of its results: it finds the exhaustive set of all patterns which exceed specified minimum support and confidence thresholds. For this reason it provides a rather detailed description of a dataset's structure. On the other hand, however, the set of discovered rules is typically vast and contains many obvious patterns.

The underlying idea of this thesis is to detect interesting changes in a dataset by analysing the support and confidence of association rules along the time axis. In contrast to the indicator method discussed above this allows for the detection of interesting changes in a dataset at nearly any level of granularity. The starting point of such a *rule change mining* approach is as follows: a timestamped dataset is divided according to intervals along the time axis. Association rule mining is then applied to each of those subsets. This yields sequences, or *histories*, of support and confidence for each rule which can be analysed further. Of particular interest are regularities in the histories which I will call *change patterns*. They allow statements about the future development of a rule and thus provide a basis for proactive decision making. For the rule mentioned above, for example, a change pattern may be that the fraction of dissatisfied customers among all male broadband users exhibits an upward trend.

1.2 Objectives of this Thesis

This thesis aims to develop a framework of methods for rule change mining which is applicable to a broad range of business problems. Rule change mining, in this context, is understood as the discovery of interesting patterns in the change of support and confidence of association rules over time.

Because the framework is intended to support proactive decision making in business processes, it is important that the discovered change patterns indicate the likely future development of a rule. On the other hand, the patterns should meet high requirements regarding their reliability, interpretability and interestingness. These requirements are necessary because the number of detected change patterns is usually vast and thus any subsequent manual verification, post-analysis or search step is deemed undesirable.

An analysis of the task of rule change mining shows that it consists of three key problems for which solutions have to be found: first, association rules have to be discovered from a timestamped dataset and their histories efficiently stored and managed; second, change patterns have to be reliably discovered; third, the detected change patterns have to be automatically post-processed in order to give a user support in identifying the most relevant and interesting ones.

Several methods for rule change mining have been proposed which solve certain aspects of the key problems above. However, a thorough analysis shows that they have significant drawbacks: some of them require extensive manual intervention or are only suitable for histories of two periods length, whilst others do not meet the abovementioned requirements regarding reliability and interpretability. This thesis therefore aims to develop new and improved methods, partially taking the innovative ideas and concepts of existing approaches as its basis. Furthermore, it has also to be analysed which other methods are necessary to meet the aforementioned requirements on change patterns.

Finally, the effectiveness of the framework and its methods has to be demonstrated on a real business problem.

1.3 Thesis Overview

Chapter 2 describes the concepts of association rules and their discovery, then proceeds to discuss typical problems related to them, which are relevant for rule change mining, and gives an overview of proposed solutions.

Chapter 3 provides a formal definition of "rule change mining" and pinpoints the requirements on the framework to develop. This is followed by an in-depth review of existing rule change mining approaches. In particular, the deficiencies of these approaches are analysed with regard to the requirements on the framework and their impact on the analysis of real business problems assessed.

Chapter 4 proposes the architecture of the rule change mining framework and motivates the business problem to which it will be applied for testing purposes. The basic design consists of three layers – the *mining layer*, *the detection layer* and the *evaluation layer*. A detailed discussion of each layer's methods is provided in the subsequent chapters, whereby each chapter covers a different layer. At the end of each chapter the proposed methods are experimentally evaluated.

Chapter 5 discusses the *mining layer*. The influence of different approaches from the field of association rule discovery on the quality of rule change mining results is analysed. Furthermore, a database scheme to efficiently manage rules and their histories is specified.

Chapter 6, which covers the *detection layer*, discusses the discovery of change patterns. Besides a discussion of *exponential smoothing* as a method for noise reduction, this chapter mainly focuses on statistical tests for trend and stability detection.

Chapter 7 describes the components of the *evaluation layer*. In the first part of this chapter a novel approach for pruning change patterns is proposed. The second part introduces several metrics for the interestingness assessment of stabilities and trends.

Chapter 8 summarises the results of this thesis and points out possible future work.

Chapter 2

Association Rules

2.1 Preliminaries

Roughly speaking, the problem of finding association rules can be stated as follows (Agrawal et al., 1993): given a database of sets of objects, discover the important associations among objects such that the presence of some objects will imply the presence of other objects within the same set. In market basket analysis – the use case association rule mining was originally intended for – the sets correspond to sales transactions and the objects correspond to purchased items. A typical rule would look like "A customer who buys products x_1 and x_2 will buy product x_3 with probability p". This knowledge can then be used to create special offers, to arrange products in a market or for customer-tailored advertisement, assuming that the sale of one item will influence the sale of others. However, association rule mining is not bound to this particular purpose. Given a database table with nominal attributes, every combination of an attribute and its value can be seen as an item and records as transactions. This opens it to a much wider range of business applications than just retail. Moreover, association rule mining is complete – in contrast to other methods, like decision trees, it has the ability to detect all patterns contained in data. Therefore it is widely agreed that association rule mining is a very flexible yet powerful tool.

Formally, association rule mining is applied to a set $\mathcal{D}$ of *transactions* $\mathcal{T} \in \mathcal{D}$. Every transaction $\mathcal{T}$ is a subset of a set of literals $\mathcal{L}$. These literals are commonly called *items* and a subset $\mathcal{X} \subseteq \mathcal{L}$ with $|\mathcal{X}| = k$ a *k-itemset*, or short *itemset*. It is said that a transaction $\mathcal{T}$ *supports* an itemset $\mathcal{X}$ if $\mathcal{X} \subseteq \mathcal{T}$.

An *association rule* r is an expression $\mathcal{X} \Rightarrow \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are itemsets, $|\mathcal{Y}| > 0$ and $X \cap Y = \varnothing$. Its meaning is quite intuitive: given a database $\mathcal{D}$ of transactions the rule above expresses that whenever $\mathcal{X} \subseteq \mathcal{T}$ holds, $\mathcal{Y} \subseteq \mathcal{T}$ is likely to hold too. If for two rules $r : \mathcal{X} \Rightarrow \mathcal{Y}$ and $r' : \mathcal{X}' \Rightarrow \mathcal{Y}$, $\mathcal{X} \subset \mathcal{X}'$ holds this is denoted by $r \succ r'$ and said that r is a *generalisation* of r' and accordingly r' is a *specialisation* of r. In this thesis I focus on association rules whose consequent is a 1-itemset, since rules of this kind are usually sufficient for most applications. A rule is then written as $\mathcal{X} \Rightarrow y$, with $\mathcal{X} \subset \mathcal{L}$ and $y \in \mathcal{L}$.

The predictive ability of a rule $r : \mathcal{X} \Rightarrow \mathcal{Y}$ is measured by its *confidence* $\operatorname{conf}(r)$ defined as the ratio of transactions that contain $\mathcal{Y}$ additionally to $\mathcal{X}$ with regard to the number of transactions that contain just $\mathcal{X}$:

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

Obviously, this is an estimate for $P(\mathcal{Y} \subset \mathcal{T} \mid \mathcal{X} \subset \mathcal{T})$, or short $P(\mathcal{Y} \mid \mathcal{X})$.

The significance of a rule $r : \mathcal{X} \Rightarrow \mathcal{Y}$ is measured by its *support* defined as the fraction of transactions that contain $\mathcal{X} \cup \mathcal{Y}$:

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

Obviously, the support estimates $P(\mathcal{X} \cup \mathcal{Y} \subseteq \mathcal{T})$, or short $P(\mathcal{XY})$. It should be noted that this definition of support depends merely on the itemset the rule has been generated from, but not directly on the rule itself. To stress this fact, I will sometimes denote the support of an itemset $\mathcal{XY}$, or a rule $\mathcal{X} \Rightarrow \mathcal{Y}$ by $\text{supp}(\mathcal{XY})$.

The above definition of support is somewhat technical, fitted to the needs of the mining algorithm. To express the significance of a rule r more intuitively the *antecedent support* $\text{asupp}(r)$ is sometimes used. It is defined as the fraction of transactions which contain $\mathcal{X}$:

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

When mining association rules theoretically a vast number of candidate rules must be considered. In fact their number grows exponentially with $|\mathcal{L}|$. Practically it is never required nor wanted to mine such a large number of rules. The search space can be significantly reduced if lower thresholds for support and confidence, supp_{min} and conf_{min}, are used. Based on this, the primary idea for association rule discovery proposed in (Agrawal et al., 1993) is to split the task into two steps:

1. The set of *frequent itemsets*

   $$\mathcal{I}(\mathcal{D}) := \{\mathcal{X} \ : \ \text{supp}(\mathcal{X}) \geq \text{supp}_{min}\} \tag{2.4}$$

   is generated from the dataset $\mathcal{D}$ by traversing the power set of $\mathcal{L}$. This step utilises support's downward-closure property: $\text{supp}(\mathcal{X}) \geq \text{supp}(\mathcal{X} \cup \{x'\})$ for any itemset $\mathcal{X}$ and $x' \in \mathcal{L}$. This property implies that every proper superset of an infrequent itemset is infrequent too. Therefore the traversal starts with 1-itemsets which are then expanded to 2-itemsets and so on, continuing until all frequent itemsets are discovered.

2. The set of *association rules*

   $$\mathcal{R}(\mathcal{D}) := \{\mathcal{X} \Rightarrow y \ : \ \mathcal{X} \cup \{y\} \in \mathcal{I}(\mathcal{D}) \wedge \text{conf}(\mathcal{X} \Rightarrow y) \geq \text{conf}_{min}\} \tag{2.5}$$

   is generated from $\mathcal{I}(\mathcal{D})$ in a straightforward manner.

Subsequent work has primarily concentrated on this approach. Common algorithms are, for example, *apriori* (Agrawal and Srikant, 1994), *eclat* (Zaki, 2000) and *FP-growth* (Han et al., 2004). But many more exist and also as many variations and extensions. The reason for this diversity is the very costly generation of frequent itemsets: a large number of candidates needs to be generated and the determination of their supports requires multiple dataset scans. For this reason the latter two algorithms solely discover frequent itemsets – the less expensive rule generation has to be done subsequently. It should be noted that there are many applications which do not require any predictive models. In these cases the sole generation of frequent itemsets is desired and sufficient.

Studies show that no single 'best' algorithm can be identified (Hipp et al., 2000; Goethals and Zaki, 2004). Basically the choice depends on the structure of the data, the constraints used and the efficiency of the actual implementation of the algorithm. It must be stressed, however, that most comparative studies[1] just assess the processor and memory utilisation, not the quality of discovered rules.

2.2 Constrained Mining and Pruning

The ability to discover all patterns is an association rule learner's strength and likewise its weakness. Usually – even for moderate confidence and support thresholds – the number of discovered associations can be immense, easily in the thousands or even tens of thousands. Clearly, the large numbers make rules difficult to examine by a human user. This rule quantity problem is particularly evident when dense datasets are analysed. In general, dense data sets have any or all of the following properties: (Bayardo et al., 2000)

- many frequently occurring items (e.g. Sex=Male)
- strong correlations between several items (e.g. between FamilyStatus=Single and Age=UnderTwenty)
- many items in each transaction

While market-basket data is mostly sparse, many other domains use dense data, e.g. census and telecommunication data analysis. Therefore significant research has been conducted into methods which reduce the number of rules generated. As a result a broad range of methods have been proposed, which can roughly be divided into constrained and pruning methods. However, it should be noted that this distinction is arbitrary and primarily influenced by the terms used in the corresponding publications. However, in order to have a clear notion for the discussion below, I will term all those techniques as 'constrained methods' which discard rules based on criteria that are independent of the underlying transaction set. On the other hand, I understand 'pruning methods' as techniques which utilise statistical properties of rules with respect to a transaction set. In the following the major works and particularly the concepts in both areas will be briefly presented. Later in this thesis I will explain how and why these concepts may cause problems for rule change mining approaches.

As the name implies, constrained mining approaches discover each rule that satisfies a set of user determined hard constraints – in addition to the thresholds on support and confidence. (Pei and Han, 2002) point out three commonly encountered classes of constraints:

- *Item constraints* specify which items or groups of items should or should not be present in a rule. Such constraints can, for example, be specified by Boolean expressions (Srikant et al., 1997) and are particularly useful if only specific target items should appear in the rule consequent.
- *Length constraints* specify an upper bound on the number of items contained in a rule. This is, for example, useful in domains where rules with many items are difficult to interpret.

[1] An exhaustive, annually repeated comparison can be found at http://fimi.cs.helsinki.fi.

- *Aggregate constraint* specify an aggregate of items in a rule. For example, in market basket analysis a user may be interested in every rule for which the overall price of its constituting items is above a certain threshold. A discussion about the utility of several common aggregation functions for association rule mining can be found in (Ng et al., 1998).

Generally, constrained approaches allow a user to obtain a rule set as small as he wishes, but likewise increase the risk of discarding potentially useful rules. Pruning approaches, on the other hand, simplify the association rule set by discarding non-interesting (Bayardo et al., 2000), non-significant (Liu et al., 1999) or redundant (Zaki, 2004; Li et al., 2004) rules. However, there is a lack of consensus on what terms like interestingness, significance and redundancy, actually mean – the used definitions vary greatly.

The pruning approach of (Liu et al., 1999) discards non-significant rules. A rule $\mathcal{X}\mathcal{Y} \Rightarrow \mathcal{Z}$ is non-significant if its specialising items $\mathcal{Y}$ and its consequent $\mathcal{Z}$ are not positively correlated with respect to a more general rule $\mathcal{X} \Rightarrow \mathcal{Z}$. This means the items $\mathcal{Y}$ are spurious and appear more by chance than by a true underlying association. The non-significance is tested by means of a χ^2 test, which utilises the support value of the less general rule.

One of the most significant pruning approaches has been published by (Zaki, 2004) and is based on *closed frequent itemsets*. A frequent itemset $\mathcal{X}$ is closed if there exists no other frequent itemset $\mathcal{X}\mathcal{Y} \supset \mathcal{X}$ with $\text{supp}(\mathcal{X}\mathcal{Y}) = \text{supp}(\mathcal{X})$ (Zaki and Hsiao, 2002). The set of closed frequent itemsets has the salient property of being lossless in the sense that each non-closed itemset can be derived from it. The method of (Zaki, 2004) takes as input the set of closed frequent itemsets and constructs a non-redundant rule set from it, whereby a rule is redundant if a more general rule with the same support and confidence exists.

A very recent and promising pruning approach is to discover the *informative rule set* (Li et al., 2004). Given a set of association rules with no more than one item in their consequents, the informative rule set is basically derived by discarding all rules r for which a more general rule $r' \succ r$ exists such that $\text{conf}(r') \geq \text{conf}(r)$. Note that this definition is very similar to the approach of (Bayardo et al., 2000) outlined below. However, compared to the informative rule set the latter approach requires all rules to have the same consequent. One can show that the approach of (Li et al., 2004) discards at least all rules which are redundant with respect to the aforementioned approach of (Zaki, 2004) and outputs hence a considerably smaller rule set.

A whole class of pruning approaches can be derived by introducing further rule quality measures – in addition to confidence and support. Most of these measures are also known as *objective interestingness measures*, generally discussed in the next section. Only those rules are discovered which exceed user-defined thresholds on all measures. For example, (Bayardo et al., 2000) introduce the *minimum improvement* constraint. The improvement of a rule is the minimum difference between its confidence and the confidence of any more general rule. Thus a negative improvement means that there is at least one more general rule which is more predictive. If the improvement is positive, then removing any item from the rule's antecedent leads to a drop in confidence of at least its improvement. A rule is therefore discarded if its improvement, support and confidence are below given positive thresholds.

2.3 Measuring Interestingness

Besides the rule-quantity problem, covered in the previous section, there is also a rule-quality problem: rules might be obvious, already known or not relevant. Rule pruning or constrained approaches do not satisfactorily solve this problem, since a rule can be non-redundant but still very obvious. Otherwise, the mere assessment of interestingness does not necessarily prevent redundant rules from being presented to a user (Shah et al., 1999). However, both problems are related in the sense that knowledge about the quality, or interestingness, of a rule may be used, for example, to discard rules.

It should be noted that the naive approach to increase the lower or to additionally introduce upper thresholds on confidence and support does not solve the rule-quality problem. On the one hand interesting rules commonly have a low support, for example if they express exceptions from the user's background knowledge (Hussain et al., 2000). On the other hand this needs not always be the case. For example, common-sense rules, which usually expose a high support, can be interesting if the problem domain is only partially understood by the user.

Many methods of measuring a rule's interestingness have been published to tackle the rule-quality problem. Interestingness is a highly subjective and hence psychological matter, but publications in the field of interestingness measures insufficiently account for this. For this reason this issue is very briefly addressed in the following, before some approaches are discussed.

2.3.1 Psychological Consideration

The notion of *interestingness* is rather unknown in psychological literature[2]. In fact, the notion that comes closest is *attention.* Loosely speaking, attention is selectivity in information processing. Or, as (James, 1890) noted, attention is "the withdrawal from some things in order to deal effectively with others". This selectivity is essential, since human information processing and storage is only able do deal with a limited number of stimuli at the same time, forcing us to focus only on some. The control of this selection is guided by the joint influence of top-down (task-driven) and bottom-up (stimulus-driven) considerations (Wilson and Keil, 1999). Top-down selection is guided from the inside depending on our current intentions and goals. A person will then just focus on the stimuli which he considers important to fulfill his goals and neglect others. In contrast, with bottom-up processing our attention is directed from the outside. It is – independently of our current goals or plans – drawn to stimuli with properties which stick out from the rest. From this perspective interestingness measures can be seen as some kind of pre-attentive tool, that supports a user to decide what to filter out and ignore. It should be emphasised, however, that almost all methods for measuring interestingness have been designed without psychological aspects in mind. I will try nonetheless to bridge this gap in the following discussion.

2.3.2 Objective Measures

Studies about interestingness measures can roughly be divided into two classes. The first class are objective measures. These are usually derived from statistics, information

[2] For example, it is not mentioned in comprehensive psychological encyclopedias like (Wilson and Keil, 1999)

theory or machine learning and assess numerical or structural properties of a rule and the data to produce a ranking. Objective measures do not take any background information into account and are therefore suitable if an unbiased ranking is required, e.g. in off-the-shelf data mining tools. Examples of such measures are lift, conviction, odds ratio and information gain. Overviews can be found in (Hilderman and Hamilton, 1999) and (Tan and Kumar, 2000).

In (Tan et al., 2004) it is empirically shown that some measures produce similar rankings while others almost reverse the order. This poses the problem of choosing the right measure for a given scenario. One solution is to discover all rules that are interesting to any measure out of a predefined set (Bayardo, Jr. and Agrawal, 1999). On the one hand this has the clear disadvantage of an increasing number of rules. On the other hand it shows more facets of the data. A different approach is presented by (Tan et al., 2004). They developed two preprocessing methods which – integrated into a mining algorithm – render many measures consistent with each other. The same publication also presents an algorithm which finds a measure that best fits the requirements of a domain expert. This is accomplished by an interactive interestingness rating of a small set of patterns.

From a psychological perspective objective interestingness measures might be interpreted as bottom-up filters, because those rules which show a salient property are highly-rated. But there are also arguments against this: such measures rarely fit a human's intuition and it can happen that many highly ranked rules have the same value, so that no salient best one can be identified.

2.3.3 Subjective Measures

In contrast to objective measures, subjective measures incorporate a user's background knowledge. In this class a rule is regarded interesting if it is either *actionable* or *unexpected*.

Actionability of a rule means that the user "can act upon it to his advantage" (Silberschatz and Tuzhilin, 1996). Since the focal point is on rules that are advantageous for the user's goals, this can obviously be seen as top-down processing. The actionability approach needs detailed knowledge about the current goals and also about the cost and risks of possible actions. Systems that utilise it are hence very domain specific, like the *KEFIR* system described in (Piatesky-Shapiro and Matheus, 1994).

A rule is unexpected if it contradicts the users knowledge about the domain. If we think of a rule as some kind of stimulus, then we are interested in salient rules with respect to our domain knowledge. On the one hand this can be seen as bottom-up processing. On the other hand it is top-down if the domain knowledge does not describe factual knowledge, but an expected or targeted state of the world.

Systems that build upon this approach require the user to express his domain knowledge – a sometimes difficult, long and tedious task. The methods are usually based on pairwise comparison of a discovered rule with rules representing the user knowledge. This comparison can be logic-based (Padmanabhan and Tuzhilin, 1999, 2000, 2002) or syntax-based (Liu et al., 1997). In logic-based systems a contradiction is determined by means of a logical calculus, whereas in syntax-based systems a rule contradicts if it has a similar antecedent but a dissimilar consequent.

In (Wang et al., 2003) the dynamic aspects of unexpectedness are analysed. They identified two issues that were ignored in previous systems: first, with every rule seen, the user's domain knowledge changes and his objectives may become clearer or change.

This could, for example, render some remaining rules no longer unexpected. Thus the set of unexpected rules should be revised after a rule has been presented. Second, whether a rule is unexpected or not depends on *how* a user likes to apply his prior knowledge to it – in addition to the prior knowledge itself.

Subjective measures strictly rely on prespecified user knowledge. To overcome the problem of its tedious collection, some research has been conducted into methods which discover unexpected patterns without forcing a user to specify his prior knowledge. For example, the approaches by (Hussain et al., 2000; Suzuki and Zytkow, 2000) are based on the observation that common sense rules usually exhibit a high support and confidence.

2.3.4 Conclusion

Overall, it should be noted that objective measures are the easiest, but also a considerably inflexible way to determine interestingness. Systems that build upon actionability and unexpectedness are more sophisticated and have psychological underpinnings, but a lot of effort is necessary to collect, organise and finally incorporate domain knowledge. Moreover, domain experts often forget certain key aspects or may not remember others which come into play under rarer circumstances. This problem can be termed 'expert dilemma' and has already been observed by designers of expert systems in the 1980s (Fogel, 1997).

Although the interestingness of discovered patterns was always one of data mining's key issues, no universal nor perfect method has been developed yet. The assessment of interestingness is one of the most difficult problems in data mining research and still experiencing only slow progress (Piatetsky-Shapiro, 2000). Moreover, its progress forecasts are pessimistic even on a long term scale and it thus will remain one of the most challenging topics in data mining research (Fayyad et al., 2003).

Chapter 3

Discovering Patterns in Rule Changes

The association rule mining approaches discussed in the last chapter typically assume that the underlying patterns hidden in the data are stable over time. However, in many real world domains data is collected over long periods and it is likely due to a variety of influences that its characteristics and hence the patterns hidden in it change significantly. Generally, this process is known as *concept drift* (Widmer and Kubat, 1996).

Due to the dynamics in the data's underlying structure, association rule mining is faced with new challenges but also with new opportunities: on the one hand, incremental mining methods are necessary to keep the discovered associations efficiently up-to-date in the presence of changes in the data characteristics without having to start a completely new data mining session. In the past years significant research has been conducted into this area and several algorithms have been proposed. However, the discussion of these approaches is outside the scope of this thesis; the interested reader is referred to the survey of (Spiliopoulou and Baron, 2002).

On the other hand, the changes in the patterns themselves can provide valuable, useful information. In fact, it is crucial for the success of most businesses to detect changes, correctly interpret their roots and finally to react or adapt to them. The changes in a dataset's underlying domain are reflected by changes in the patterns hidden within it. Therefore the problem of identifying changes in a certain domain can be reduced to the identification of changes in patterns. Association rule mining is a powerful basis for this task: its ability to discover all patterns contained in data allows a user to analyse changes in the data characteristics at virtually any level of granularity.

3.1 Problem Statement

Let $\mathcal{D}$ be a timestamped set of transactions and $[t_0,\ t_n]$ the minimum time span covering all its tuples. The interval $[t_0, t_n]$ is divided into $n > 1$ non-overlapping periods $T_i := [t_{i-1}, t_i]$, such that the corresponding subsets $\mathcal{D}(T_i) \subset \mathcal{D}$ each have a size $|\mathcal{D}(T_i)| \gg 1$. I will denote the set of all periods as $\hat{T} := \{T_1, \ldots, T_n\}$.

At the end of each period T_i a set of association rules $\mathcal{R}(\mathcal{D}(T_i))$ is generated from $\mathcal{D}(T_i)$. Because rule quality measures, like confidence and support, of every rule $r : \mathcal{X} \Rightarrow y$ are now related to a specific transaction set $\mathcal{D}(T_i)$ and thus to a certain time period T_i their notations need to be extended. Starting from (2.1), (2.2) and (2.3) this is straightforward and yields:

$$
\begin{aligned}
\operatorname{supp}(r,\ T_i) &= \operatorname{supp}(\mathcal{X}y, T_i) &\approx&\quad P(\mathcal{X}y|\ T_i) && (3.1)\\
\operatorname{asupp}(r,\ T_i) &= \operatorname{supp}(\mathcal{X},\ T_i) &\approx&\quad P(\mathcal{X}|\ T_i) && (3.2)\\
\operatorname{conf}(r,\ T_i) & &\approx&\quad P(y|\ \mathcal{X},\ T_i) && (3.3)
\end{aligned}
$$

It is likely that rules may not be present in some of the discovered rule sets $\mathcal{R}(\mathcal{D}(T_i))$, because their support or confidence is below the specified threshold of the rule miner. I will refer to the maximal subset of rules present in all rule sets as the *compound rule set* and denote it by

$$\hat{\mathcal{R}}(\mathcal{D}) := \bigcap_{i=1}^{n} \mathcal{R}(\mathcal{D}(T_i)) \tag{3.4}$$

Obviously, the temporal development of each rule $r \in \hat{\mathcal{R}}(\mathcal{D})$ is described by n values for conf, supp and asupp respectively. Imposed by T_i the values are inherently ordered and hence form sequences

$$
\begin{aligned}
H_{supp}(r) &:= (\operatorname{supp}(r, T_1), \ldots, \operatorname{supp}(r, T_n)) && (3.5)\\
H_{asupp}(r) &:= (\operatorname{asupp}(r, T_1), \ldots, \operatorname{asupp}(r, T_n)) && (3.6)\\
H_{conf}(r) &:= (\operatorname{conf}(r, T_1), \ldots, \operatorname{conf}(r, T_n)) && (3.7)
\end{aligned}
$$

Accordingly they are called *support history*, *antecedent support history* and *confidence history* of the rule r, respectively. If the semantic of a history's values is not relevant the subscript of H will be dropped.

The previous chapter showed that it is possible to generate for any transaction set virtually every association rule. Support and confidence then determine the rule's degree of validity, with zero support indicating that the rule does not hold at all. For this reason I will treat the mere symbolic representation of a rule itself as time-invariant, i.e. each rule is formally present at any point in time and no rule can change into another. Consequently, quality measures and other rule statistics are time-variant. For example, the rule MALE ⇒ VERYSATISFIED CUSTOMER cannot change into MALE ⇒ SATISFIED CUSTOMER as both rules are present at any point in time. On the other hand, the confidence of the first rule, for example, may decline over time, whereas the confidence of the latter inclines. This again, however, could be interpreted by an expert as the continuous change of the first rule into the other. Therefore, the term *rule change* refers to changes in the rule statistics, but not directly to the rule itself.

The aim of *rule change mining* is to detect patterns in rule histories. In this thesis I will understand any significant global regularity in the values of a history as a pattern. Trends, stabilities and cyclical variations are, for example, commonly encountered types. Because such patterns refer to the development of a rule, I will call them *change patterns*. It should be stressed that in my understanding of rule change mining the primary objective is to detect change patterns, but not to model the underlying process which generated a change pattern or more generally a history, for example, by a regression model. I will explain the reason for this distinction in the next section.

The rule-quantity problem, discussed in the previous chapter, has also implication for rule change mining: typically a large number of change patterns will be detected. The used change pattern types and their detection methods should account for that and, consequently, should therefore be chosen such that any manual verification, post-analysis and search of detected change patterns is avoided. Based on this, the following requirements can be identified:

- **Reliability:** Given a set of rules and their corresponding histories, the detection method should, on the one hand, detect all occurrences of a certain type of pattern. On the other hand, the probability of a false alarm, i.e. the probability that a pattern is wrongly detected, should be low and ideally controllable.

- **Interpretability:** Types of patterns and detection methods should be meaningful to a user and not contradict his intuition. Particularly, the types of discovered patterns should be as simple as possible and as complex as necessary.

- **Usefulness:** Based on the knowledge that a rule's history exhibits a certain pattern an expert should be able to roughly estimate its business value without the need to manually review the history or to apply further analysis. Particularly, the detected pattern types should be able to trigger proactive business decisions, in contrast to the reactive decision making supported by association rule mining.

- **Interestingness:** In compliance with the discussion in Section 2.3.1 those patterns should be presented to a user which would gain his attention if he had the possibility of reviewing all detected patterns manually.

This list is not complete, but reflects criteria considered important in business applications of rule change mining. Moreover, they are also basically supported by other authors, e.g. (Borgelt and Kruse, 2002).

In addition to the aforesaid about rule change mining, the following points should be stated:

1. The compound rule set $\hat{\mathcal{R}}$ can – similar to (3.4) – alternatively be defined as the union of all rule sets. This implies a performance problem: rule change mining needs complete histories for each rule. Because a rule's support and confidence change over the course of time, the rule may be present in the rule set of some periods, but missing in others. Consequently, values for support and confidence would be missing too. To calculate them the datasets have to be scanned once more. Depending on the size of the datasets and the number of items, periods and rules, this can be very time-consuming and may be, particularly with respect to typical large and dense business datasets, practically infeasible. Therefore this approach is not used in this thesis. However, all proposed methods are independent of the actual definition of $\hat{\mathcal{R}}$.

2. In addition to patterns it is also very interesting to search for certain events. An event could be, for example, the turning or starting point of a trend. However, although this sort of event might look simple, its timely and reliable detection turns out to require very sophisticated methods even in well-understood domains like the analysis of business cycles (Andersson et al., 2004). Generally, non-parametric solutions are still a current research issue and hence the detection of such events is outside the scope of this thesis.

3. Generally, rule change mining can also be applied to histories of other measures. In particular, all objective interestingness measures for rules are suitable. Nonetheless, compared to those measures, support and confidence have the advantage of being easy interpretable.

3.2 Some Notes on Time Series Analysis

A time series is a set of observations measured sequentially through time. Obviously, the history of a rule is a time series. For this reason the relationship between the detection of change patterns in rule histories and conventional time series analysis is discussed in the following.

The main objectives of classical time series analysis are (1) to describe the data, (2) to model the data-generating process, (3) to do forecasting and (4) to use the forecasts to control a given process (Chatfield, 1996). The latter two objectives depend mostly on the availability of a valid model. However, there are also forecasting methods which do not rely on a model, like *exponential smoothing* and *moving average*.

To fit a model a sufficiently large number of values both for parameter estimation and testing is necessary, but rarely available in rule change mining. Moreover, such a fitted model may not adequately describe the underlying process. For example, a regression line can be fitted to any pairs of values – independent of how the dependent value actually relates to the other. Therefore a model needs to be validated, but this typically involves human intervention. However, due to the rule-quantity problem rule change mining approaches must be able to cope efficiently with a vast number of time series. Thus it is practically infeasible to find a suitable model for each of them (Chatfield, 2001).

But even if it were possible to efficiently identify the models, it is doubtful that a user would really want to have such a vast number of sophisticated models. He commonly needs exact predictions for just a tiny fraction of rules and rarely has a precise idea about how rules should evolve, if and which target values should be reached. Hence the full potential of all the models would rarely be exploited and the user bothered with an unnecessary complexity.

Rule change mining, however, should be understood as in-line with the first of the above objectives to the extent that it provides a descriptive statement about a time series with the ability to deliver qualitative short term predictions and to suggest control actions. On the other hand, rule change mining and classical time series analysis are complementary in the sense that the latter should be used to provide a thorough analysis of histories that exhibit interesting change patterns.

3.3 Review of Related Approaches

Surprisingly, the number of publications related to rule change mining is rather limited and focused – as is this thesis – on the change in rule statistics. I will now review the available publications and provide a thorough analysis of the most relevant approaches. In particular, I will discuss their drawbacks with regard to the requirements stated in Section 3.1.

3.3.1 Shape Queries

In (Agrawal and Psaila, 1995) a query language for history shapes is introduced. The user has to specify several shapes, which he considers interesting, and the histories are then matched against them. In addition to this approach, which has been developed to query association rule histories, there are several similar methods in the field of shape retrieval from time series, like (Goldin and Kanellakis, 1995) and (Rafiei and Mendelzon, 1997).

However, all of these approaches have the disadvantage that the user has to know in advance which shapes occur and which of them might be interesting. Thus the discovered patterns are very biased towards the expectations of the user – other novel and unexpected patterns may remain unrecognised.

3.3.2 Emerging Itemsets

An itemset is emerging if its support increases sharply between two periods. Efficient algorithms which detect all itemsets whose support change is above a user-defined threshold have been proposed by (Dong and Li, 1999) and (Zhang et al., 2000). These approaches are not adaptable to any other measure apart from support and do not take more than two consecutive periods into account. From the perspective of rule change mining they are therefore unable to detect long-term change patterns. In fact, the purpose of emerging itemsets is to reveal the significant differences between two datasets.

3.3.3 A Fuzzy Approach

In (Au and Chan, 2002, 2005) a fuzzy approach to predict confidence and support changes on the basis of association rule histories is presented. The method basically expresses reoccurring change patterns in a set of fuzzy rules which are then used for prediction. Since confidence and support are treated equally, the following discussion focuses, without loss of generality, on support.

The proposed method measures a change as the absolute difference in support between two periods. By subtracting each value of a history from its successor, a support history is transformed into a history of the absolute changes between consecutive periods. Furthermore, the range of possible absolute changes is partitioned into fuzzy sets. Fuzzy rules for prediction are then generated as follows: a sliding window of fixed size is moved over the change history. It is convenient to think of every position within the window as an attribute. The value of each attribute – a absolute support difference – is then expressed in terms of membership degrees to the fuzzy sets defined above. Hence, each position of the window yields a fuzzy data tuple; all window positions together a data set. This data set is then used to learn fuzzy rules which predict the window's last value. The authors claim that these "fuzzy meta-rules" are able to predict the future development of association rules.

The algorithm creates several fuzzy meta-rules – precisely a fuzzy decision tree – for each association rule history. This will obviously lead to a further explosion in complexity – in addition to the already vast number of association rules. This effect worsens if the histories are noisy or exhibit non-linear trends. In order to prevent this, a user must decide in advance to which rules this method has to be applied, thus rendering it impossible to discover change patterns for rules which have not been considered.

Furthermore, change is measured as the absolute difference in support, which can be problematic. Because support, like confidence, is a relative measure, the interpretation of an absolute change varies with its value. For example, an increase by 0.05 would not be much for a support of 0.8. In contrast, an increase in support from 0.05 to 0.1 means a doubling of the transactions covered by the corresponding rule and therefore a significant change. The approach cannot distinguish between these cases, hence the precision and reliability of the induced “fuzzy meta-rules” declines with smaller support values.

The approach strictly relies on multiple reoccurrences of exactly the same change patterns in a history and the assumption that these patterns will occur in the future too. It is doubtful that this is always a realistic assumption. Moreover, as motivated at the beginning of this chapter, it is more reasonable to assume that change patterns are generally triggered by unique events, like management decisions or the launch of a competitor’s new product. Such change patterns are obviously non-repetitive and would therefore not be easily detectable by means of “fuzzy meta-rules”.

3.3.4 Temporal Description Length

The approach of (Chakrabarti et al., 1998) is different from the other methods discussed in this chapter, as it does not rely on a particular discretisation of time. Instead the method determines the best time segmentation for every itemset such that the correlation between its constituting items is maximally homogeneous within, but inhomogeneous between segments. This is done by application of the *minimum description length principle* (Rissanen, 1978). For this purpose a binary coding scheme for time segments is defined. It assigns a code length that declines with increasing homogeneity. The segmentation is then chosen such that the sum of all segment code lengths is minimised. Furthermore, the overall code length is used to assess the interestingness of itemsets: large code lengths hint at more dynamic correlations and hence at more interesting itemsets.

The advantage of this algorithm is clearly that is does not rely on particular time periods. On the contrary, each itemset gets a different time segmentation and thus renders it difficult to compare itemsets according to their change. The method detects and ranks changes on a very abstract and hardly interpretable level. In doing so it incorporates neither the type nor the systematics of change, although they can be regarded as crucial aspects for business decisions.

3.3.5 Detection of Semi-stable, Stable and Trend-Rules

The work described in (Liu et al., 2001b) is motivated by the observation that most users of association rules feel uncomfortable with the bare presentation of rule confidence and support. They are concerned about a rule’s validity. Thus additional information about how a rule changes, if it exhibits trend, stability or other systematics, need to be provided to assist a user in judging the rule’s value. The authors therefore argue that a rule’s interestingness should be assessed by the characteristics of its behaviour over time.

Given a measure like confidence or support, rules are assigned by means of statistical tests to one of three classes if possible. Subsequently, rules are ranked within each class according to an interestingness measure. The classes are:

- **semi-stable rules:** These are rules which are present in every period. Except that the chosen measure is always above the threshold used by the rule miner, their histories may not show any other patterns.

- **stable rules:** These are semi-stable rules whose history is (statistically) stable over time.

- **trend rules:** These are semi-stable rules whose history exposes a significant trend.

Particularly stability and trend are both change patterns, which are useful in many domains and meaningful to a user. For this reason the methods for their detection proposed in (Liu et al., 2001b) will be discussed in the following. Without loss of generality I will focus on support. Because the methods are basically the same for confidence and support the following arguments will hold for confidence as well.

Stable Rules. The support history of a rule r is classified as stable if it is semi-stable and a statistical test fails to reject the following null hypothesis:

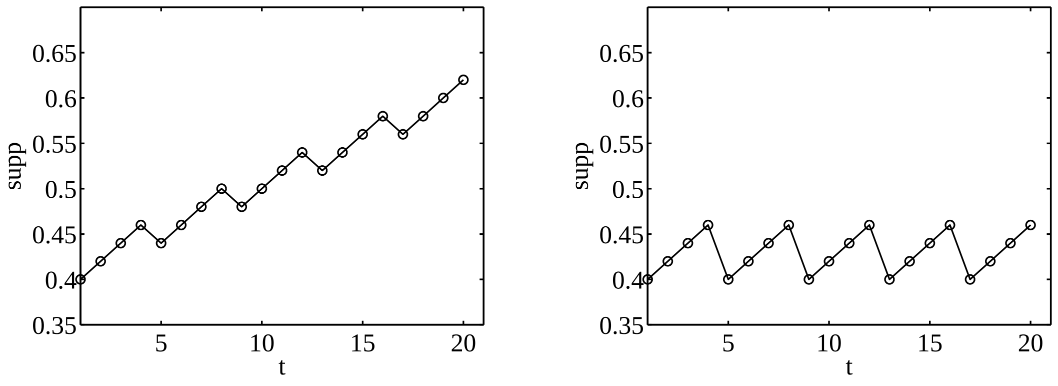
$$\begin{aligned} H_0 &: \text{supp}(r, T_1) = \text{supp}(r, T_2) = \ldots = \text{supp}(r, T_n) \\ H_1 &: \text{the supports are not all equal} \end{aligned}$$

Because support, like confidence, expresses a population proportion the χ^2 *test* for homogeneity of proportions is used. This test is described in detail in Section 6.3 and can also be found in (Liu et al., 2001b), or any textbook on statistics, like (Sheskin, 2003). I will therefore outline only the basic ideas in this section.

Given a rule and its history, the χ^2 test assumes that whether or not a transaction supports a rule is independent from the period during which a transaction has been generated. This is expressed by H_0 which basically states that the probability that a transaction supports a rule is equal in all periods. Based on this assumption the expected number of transactions which support and do not support the rule can be calculated. The weighted sum of squared differences between the expected and the observed number of transactions in each period is then calculated. If H_0 holds then the sum is χ^2 distributed. This distribution assumption is then used for testing.

The χ^2 test does not account for the inherent order of a history, since the test has been designed for nominal attributes. For example, the test result for the sequence (0.15, 0.14, 0.13, 0.12, 0.11) is the same as for (0.15, 0.12, 0.13, 0.14, 0.11), or any other of its permutations. This leads to an unreliable stability detection in the sense that a history may be classified as stable, although it actually exhibits a trend. If the χ^2 test, for example, is applied to the history above under the assumption that the number of transactions in each period is 1000, the test will not reject the null hypothesis at a significance level of 0.05. Therefore the history, which shows a rather pronounced trend, is classified as stable.

Another reliability related issue is caused by the way H_0 and H_1 are defined. To carry out a statistical test the proposition which has to be shown is generally formulated as the alternative H_1, whereas the null hypothesis H_0 is merely formulated to be rejected. This allows to control the probability that the test rejects H_0, although it is in fact correct, by the significance level. Because the method above connects the decision for stability to the non-rejection of the null hypothesis, the significance level does *not* control the probability that a rule is wrongly identified as stable. Moreover, the probability might even increase by lowering the significance level. However, the theoretical foundations of statistical tests require an equality relation in the null hypothesis. Since any test on stability will be obviously based on (statistical) equality between values, there seems to be no way to circumvent the problem described above.



(a) A history with a trend.

(b) A stable history.

Figure 3.1: Example to point out that the *runs test* should not be used for trend detection. The runs test yields identical test results for both histories.

Trend Rules. To test for a trend rule (Liu et al., 2001b) use the *runs test* because it is an "often used test whether a sequence ... exhibits a trend". The runs test takes a sequence of values and generates another sequence by setting a '+', if a value increases and a '−', if it decreases. For example, (7, 8, 9, 10, 5, 1) would yield $(+, +, +, -, -)$. Without going into further detail only the latter sequence is then used for the test.

However, the runs test is actually used to test whether a sequence is non-random (Sheskin, 2003). Being non-random does not necessarily mean that a trend exists. For example, Figure 3.1(b) and Figure 3.1(a) both show support histories. The plot to the left exhibits a trend, while the plot to the right does not. In fact, the latter could be judged as stable. Although both sequences obviously perform differently they yield identical sequences of '+' and '−'. The test result would therefore be the same for both. This is not surprising because both sequences are non-random. In Chapter 6 I will discuss two statistical tests which are – in contrast to the runs test – specifically designed to test for trends.

It is reasonable to assume that a user is particularly interested in the clarity of detected trends. Therefore the authors use the test statistic of the runs test, i.e. the number of '+' or '−', respectively, as an interestingness measure. However, because the runs test is applied to a purpose it is not designed for, the histories in the above example would get the same ranking.

3.3.6 Fundamental Rule Changes

One of the basic problems of rule change detection is that commonly a huge number of changes is reported, but most of them are merely the effect of other changes. A method proposed by (Liu et al., 2001a) aims to tackle this problem by detecting the fundamental changes, namely those that cannot be explained by others. However, the method is limited to histories of length two, i.e. it just detects the fundamental changes between two consecutive periods. Because the method is basically the same for confidence and support, the discussion will be restricted once more to support.

Given a rule $r : \mathcal{X} \Rightarrow y$, it can be seen as a combination of two more general rules

(a) change of rule supports

	T_1	T_2	$rel.change(\%)$
r	0.08	0.06	−25%
r'	0.1	0.1	0%
r”	0.7	0.77	+10%

(b) contingency table

	r	r'	r”
1	60	80	88
2	940	920	912

Table 3.1: Counterexample for fundamental rule change detection

$r' : \mathcal{X}' \Rightarrow y$ and $r'' : \mathcal{X}'' \Rightarrow y$, whereby $\mathcal{X}' \cup \mathcal{X}'' = \mathcal{X}$ and $\mathcal{X}' \cap \mathcal{X}'' = \varnothing$. Under the assumption that the proportional relationships among the constituting itemsets of r, r' and r'' should stay the same (“constant proportion assumption”) the *expected supports* $E_{r'}$ and $E_{r''}$ of r in period T_2 are computed as follows[1]:

$$
\begin{aligned}
E_{r'}(\operatorname{supp}(r, T_2)) &= \min\left(1, \frac{\operatorname{supp}(r, T_1)}{\operatorname{supp}(r', T_1)} \operatorname{supp}(r', T_2)\right) \\
E_{r''}(\operatorname{supp}(r, T_2)) &= \min\left(1, \frac{\operatorname{supp}_1(r, T_1)}{\operatorname{supp}(r'', T_1)} \operatorname{supp}(r'', T_2)\right)
\end{aligned}
$$

The rule r is defined as non-fundamental if rules r' and r'' exist, such that a χ^2 test fails to reject the following null hypothesis:

$$
\begin{aligned}
H_0 &: E_{r'}(\operatorname{supp}(r, T_2)) = E_{r"}(\operatorname{supp}(r, T_2)) = \operatorname{supp}(r, T_2) \\
H_1 &: \text{the three values are not all equal}
\end{aligned}
$$

Obviously the test is very similar to those for stability detection in Section 3.3.5. Therefore the above arguments, which are related to the choice of the hypothesis, hold for the approach discussed in this section too.

The criterion for detecting non-fundamental rules is rather strict from an intuitive point of view. All rules within a rule set are related by their constituting itemsets, so that if one rule changes many others are effected too – directly and indirectly. Furthermore, not all itemset are equally likely to be key-drivers for change as the above definition of fundamental changes implies. In fact, some items might influence change more strongly than others. For this reason it is surprising that the experimental results of (Liu et al., 2001a) show a significant reduction in the number of rules.

The reason for these results is that the χ^2 test counterintuitively classifies many changes as explainable. Before I address the underlying reasons, the claim will be illustrated by an example.

Table 3.0(a) shows the supports of the rules r, r' and r'' in periods T_1 and T_2. As can be seen in the third column, r decreases by one fourth, whereas r' stays constant and r'' increases by one tenth. Therefore a user would assume that r is not explained by r' and r'', because neither r' nor r'' decreases in its support. The expected supports for r in T_2 are $E_{r'}(\operatorname{supp}(r, T_2)) = 0.08$ and $E_{r"}(\operatorname{supp}(r, T_2)) = 0.088$ respectively. If the χ^2 test is applied, the null hypothesis that the support and the two expected supports are

[1] The definitions of r' and r'' by (Liu et al., 2001a) are questionable, because support is not a property of a rule, but of the itemset a rule has been built from. Considering this yields more possibilities to calculate the expected support.

all equal would not be rejected.[2] Consequently the rule r is discarded – against intuition of the user.

There are several reasons why the approach of (Liu et al., 2001a) fails. First, although it is crucial for our intuition, information about the direction of change is not used. Second, the χ^2 test internally operates with absolute numbers and neglects that support is a relative measure. Therefore, if the support is small enough, even large relative differences between expected and true support, would yield just small differences in the absolute values of supported transactions. These small absolute differences are then regarded as noise by the test.

3.3.7 The Pattern Monitor PAM

The framework *PAM* (**pa**ttern **m**onitor), which is proposed in the dissertation of (Baron, 2004), aims to monitor rule histories in order to detect interesting short and long term changes. In contrast to the other approaches discussed in this chapter and the initial problem statement, the monitor treats a history not as a static, but as a dynamic object which is extended permanently after every new mining session. For this reason the author makes the early detection of long term changes a key issue.

The monitor's architecture basically consists of two layers. The first layer discovers, stores and maintains association rules and their histories. Each association rule is stored using the *Generic Rule Model*, which is a record containing the discovered rule, its statistics for a particular period, a timestamp to indicate the period and a unique identifier for each rule (Baron and Spiliopoulou, 2001).

The second layer builds upon the Generic Rule Model and provides several methods to detect interesting short- and long term changes of rules. A short term change refers to the change between two consecutive periods, whereas long term change means that the values of a rule's measure do not immediately return to their former level (Baron, 2004). However, the detection methods for long term changes are basically an extension of those for short term changes. The methods are independent of the used measure, so that the following discussion will focus, without loss of generality, on support.

Significant Changes. To detect a significant change a two-sided binomial test is applied to decide if a change in support of a rule r between two consecutive periods is non-random, that is, if the following null hypothesis is rejected:

$$\begin{aligned} H_0 &: \text{supp}(r, T_i) = \text{supp}(r, T_{i+1}) \\ H_1 &: \text{supp}(r, T_i) \neq \text{supp}(r, T_{i+1}) \end{aligned}$$

If the test rejects H_0, it is additionally tested whether $\text{supp}(r, T_{i+2})$ differs significantly from $\text{supp}(r, T_i)$. If the test rejects the null hypothesis again, the change from $\text{supp}(r, T_i)$ to $\text{supp}(r, T_{i+1})$ is assumed to hold for a longer period and hence the user is notified.

This detection method never uses values from more than three consecutive periods and thus detects only isolated changes, i.e. the detected changes are independent of the global context given by the other values of the history. In fact, information from just

[2] The brief outline of the corresponding test is as follows: the underlying datasets $\mathcal{D}(T_1)$ and $\mathcal{D}(T_2)$ each having a size of 1000 transactions. The corresponding contingency table is shown in Table 3.0(b). The resulting test statistic is $\hat{\chi}^2 = 5.92$. The test statistic is smaller than the threshold value of 5.99 at significance level 0.05 for two degrees of freedom. Therefore the test would not reject H_0.

three periods is not enough to reliably detect change patterns and so this must be done subsequently.

Occurrence Based Grouping. Since the support and confidence of a rule change over the course of time it is possible that the rule is not present in some periods, i.e. it is not contained in the discovered rule sets. The method of *occurrence based grouping* aims to detect interesting changes in the occurrence frequency of a rule.

Let n be the overall number of periods and $m < n$ the number of periods in which a rule is present. The *occurrence frequency* f is then defined as $f := m/n \in [0, 1]$. Rules are grouped according to their occurrence frequency, whereby the groups are determined by a segmentation of the range of f. The user is notified about an interesting long term change if a rule changes its group and does not return to it for at least p additional periods. For example, in (Baron and Spiliopoulou, 2003) the intervals $[0, 0.5)$, $[0.5, 0.75)$, $[0.75, 0.9)$, $[0.9, 1)$ and $[1, 1]$ are used for grouping. According to the interpretation of the authors, the groups contain rules with a low, medium, high, very high and permanent occurrence, respectively. The user would, for example, be notified if a rule's occurrence frequency changes from the interval $[0.75, 0.9)$ to $[0.9, 1)$.

The occurrence frequency is a continuous measure and every group defined over it is linked to a linguistic term, like *high* or *low*. A user will rarely have a clear notion about how to strictly separate, for example, a medium from a high occurrence frequency. Notably *fuzzy theory* is motivated by similar observations. Therefore it would be intuitive to assign rules to groups in terms of real-valued degrees of membership and to consequently regard group change as a continuous process. However, the approach of (Baron, 2004) assigns every rule to exactly one group and so treats group change as a discrete event. For example, a rule which has a high frequency in one, can suddenly have a medium frequency in the next. In fact, the continuous nature of the change is transparent to a user. For this reason, and due to the rather arbitrary choice of interval borders, it is questionable if a detected group change is use- or meaningful to a user without manual post-analysis of the histories.

Corridor-based Heuristic. For this heuristic an expected interval for a history's most recent value is calculated. If the value is outside this interval the values in the following p periods are examined in the same way. In case all of them are outside of their expected intervals too, the user will be notified about an interesting long term change.

The method to calculate the interval is heuristic and the definitions for it vary among the publications. In the following let $H = (v_1, \dots, v_n)$ be a history, whereby v_n is the value of interest. Furthermore let $\bar{v} := 1/n \sum_{i=1}^{n} v_i$ denote the sample mean and $s^2(v) := 1/(n-1) \sum_{i=1}^{n} (v_i - \bar{v})^2$ the sample variance. The expected interval for v_n is in (Baron, 2004) defined as:

$$[\bar{v} - 0.5s(v),\ \bar{v} + 0.5s(v)] \tag{3.8}$$

On the contrary, in previous publications, like (Baron and Spiliopoulou, 2003) and (Baron, 2003), the following definition can be found:

$$[\bar{v} - s(v),\ \bar{v} + s(v)] \tag{3.9}$$

This means that the heuristic either uses an interval size of one sample standard deviation, or of two sample standard deviations. However, no guidelines are given on how to choose an appropriate interval for a given task and the author provides neither a clear motivation nor a theoretical framework for it. The latter is particularly desirable, because it would allow for a quantitative statement about the expected errors. In the following I will provide a theoretical model for this heuristic finally leading to an estimate of the probability of an interesting long term change being presented to a user, although the history is stable apart from noise.

Let $(v_1, \ldots, v_n)$ be a sequence of values generated by the stochastic process $V_i = c+\varepsilon_i$, i.e. the sequence is stable apart from random noise.The random variables ε_i, which model the noise, are assumed to be independent and identically distributed. Furthermore, it is assumed that the noise is normally distributed with zero mean and non-zero variance σ^2, i.e. $\varepsilon_i \sim N(0, \sigma^2)$. Hence the random variables $V_i \sim N(c, \sigma^2)$ are also independent and identically distributed. The sample mean $\bar{v}$ and the sample variance $s^2(v)$ are unbiased estimators for the parameters c and σ^2 of a normal distribution. Hence, the probability that a value is outside the interval is

$$P(V < \bar{v} - 0.5s(v) \ \vee \ V > \bar{v} + 0.5s(v)) \approx 0.62$$

for the interval (3.8) and accordingly

$$P(V < \bar{v} - s(v) \ \vee \ V > \bar{v} + s(v)) \approx 0.32.$$

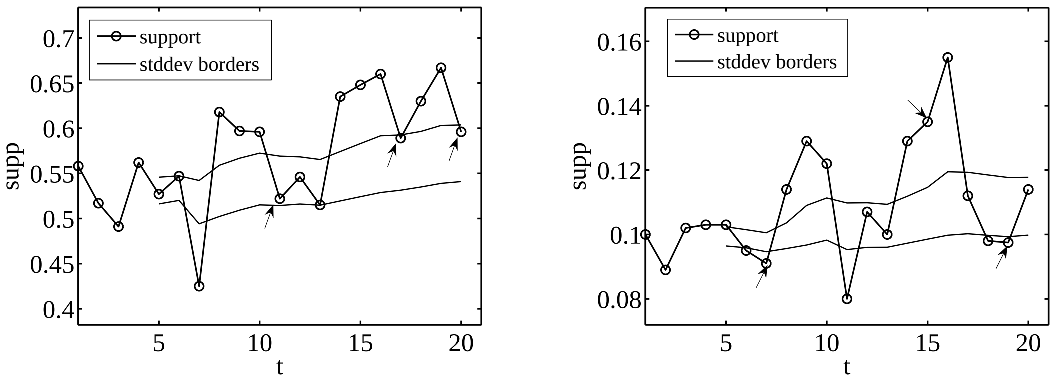
for the interval (3.9). The probability that p successive values are outside the interval is therefore 0.62^p (0.32^p). For example, to early detect long term changes the parameter p could have been chosen as $p = 5$. Depending on the actual definition of the interval to be used the probability that random noise will be falsely detected as an interesting long term change is 0.091 (0.003). In other words, if the interval (3.8) is used, it can be expected that approximately ten percent of all stable histories are presented to a user.

Without any model assumptions and thus for arbitrary sequences the reliability of the corridor based heuristic is rather difficult to assess, because such an analysis relies on the user's subjective notion of what constitutes an interesting long term change. However, I will provide two examples in the following which illustrate two other reliability related issues.

Figure 3.2(a) shows a plot of a support history drawn from a real life dataset. Additionally, the intervals (3.8) of its values are presented. The apparent upward trend of the history will not be detected by the heuristic if an arbitrary $p > 4$ is used. This is due to some values, marked with arrows, that occasionally break the count of successive values which are outside their intervals.

Figure 3.2(b) shows another plot of a support history. This time I will assume that $p = 2$. In contrast to the previous example which reported no change at all, three interesting changes are now reported. The values about which a user is notified are marked with arrows. However, although several significant values lie above the interval and the history maybe has a slight upward trend, two notifications are due to values below the interval. The plot hints at artefacts as the possible cause of these notifications, thus they might have little practical value for a user.

Overall, the reliability of the heuristic is very tightly connected to the threshold p for successive outside-the-interval occurrences. On the one hand, small values of p are necessary to early detect changes but they lead to a significant decrease in reliability. On the other hand, larger values for p would not justify the rather general definition of *long term change* anymore, because in this case more sophisticated detection methods can be utilised to detect more specific change patterns.



(a) Although the history exhibits an upward trend, no change will be detected.

(b) The user will be notified three times about interesting changes.

Figure 3.2: Examples for non-detected and detected changes by the *corridor based heuristic.*

Figure 3.2(b) points also to a problem regarding the interpretability and usefulness of the obtained results. As already noted, it shows that the first and third notification are connected to values below and the second notification to a value above the interval. A user (who does not know about the future development of the rules) might therefore expect the start of a declining or inclining trend, respectively. Because both expectations would trigger contrary business decisions, the risk of a wrong decision increases. To reduce this risk, the user has to manually analyse each history for which an interesting long term change has been detected. This has already been pointed out by (Baron, 2004, p. 89) who recommends to conduct a statistical analysis to the history whenever a change is reported in order to clarify the existence of specific change patterns.

Interval-Based Heuristic. For this heuristic the range of support or confidence, respectively, is partitioned into multiple equally spaced intervals. A user is notified about an interesting long term change if all of the following three conditions are met. First, the most recent value of a history is outside the interval of its predecessor. Second, the *absolute* difference between the value and its predecessor exceeds a user-defined limit. Third, the values of the following p periods do not return to the previous interval.

This method uses several concepts which have been previously discussed in the context of other methods. Therefore I will briefly address the concepts and discuss their implications for the interval-based heuristic in the following, but refer to the corresponding sections for an in-depth discussion.

The segmentation into intervals can be seen as a type of rule grouping and consequently the notification of a user as the result of a group change. The concept of grouping has already been discussed in the context of *occurrence-based grouping* on page 23. Like in occurrence-based grouping the segmentation is done over continuous measures and the change of the measure reduced to the discrete event of an interval change. The same arguments as in the case of occurrence-based grouping can therefore be applied to challenge the interpretability and usefulness of detected changes. Moreover, the requirement of equally spaced intervals and their missing semantics may intensify the problem.

Another problem, which also has implications for the interpretability, is the use of equally spaced intervals and absolute differences for relative measures, like support and confidence. This issue has already been discussed in the context of the fuzzy approach on rule change mining in Section 3.3.3.

The approach of (Baron, 2004) aims to discover interesting short and long term changes in rule histories. As shown it fulfills this task to a certain extent, but with several drawbacks:

1. All of the proposed detection methods yield results that are difficult to interpret. Moreover, to assess their business value an additional, possibly manual analysis of the history may be necessary. This is basically due to the fact that all methods merely detect *that* something has changed using rather simple heuristics. The more crucial question for business applications, *how* something changes, remains unanswered.

2. For the corridor-based heuristic I have shown that it may unreliably detect long-term changes. Statements about the reliability of the other heuristics are hardly possible for two reasons. First, no theoretical foundation is given for them and due to their ad hoc nature it is questionable if one exists. Second, it is impossible to state what these heuristic should detect, because no clear definition of “interesting long term change” is provided by the author.

3. Interestingness is basically a detection of a change. This means *PAM* assumes that every discovered change is inherently interesting. Thus it does not provide any alternative methods for interestingness assessment. Moreover, if a user’s notion of interestingness changes new detection methods need to be developed.

Chapter 4

Framework for Rule Change Mining

The last chapter provided a thorough review of the available publications on rule change mining. Although a variety of problems is addressed in the publications, the solutions offered mostly have drawbacks regarding the interpretability, usefulness and reliability of their results. Moreover, all approaches apart from the *PAM* framework are rather isolated in the sense that the advantages of synergies and interoperability between different methods are neglected.

Based on the problem statement in Section 3.1 this chapter proposes the architecture of a framework for rule change mining. On the one hand, it solves the basic tasks identified in the previously discussed publications, but aims to avoid their drawbacks by proposing completely new approaches. On the other hand, the architecture also accounts for problems caused by or inherited from the underlying discovery of association rules. Analogous to association rule mining one focal point will therefore be the interaction of mining, pruning and interestingness assessment to obtain valuable results.

4.1 Basic Tasks and Workflow

Basically, the objective of rule change mining is to reliably discover useful, interesting and interpretable change patterns. This objective can be broken down into several tasks, which are interconnected and build upon each other. They will be motivated and stated in the following.

As already said in the problem statement in Section 3.1 are histories of association rule measures, like support and confidence, the basis for rule change mining. To derive a history, datasets collected during many consecutive periods have to be mined. After each mining session, the discovered rules have to be compared to those discovered in previous periods and their histories extended. On the other hand, history values may be discarded if their age exceeds an application dependent threshold. Consecutive update operations may not be temporally close to another because each mining session may take place directly after the collection period ended. For this reason rules and histories have to be stored on a long term basis. Taking all of the aforesaid into account the first task can be formulated as:

1. Association rules have to be *discovered* and their histories efficiently stored, managed and maintained.

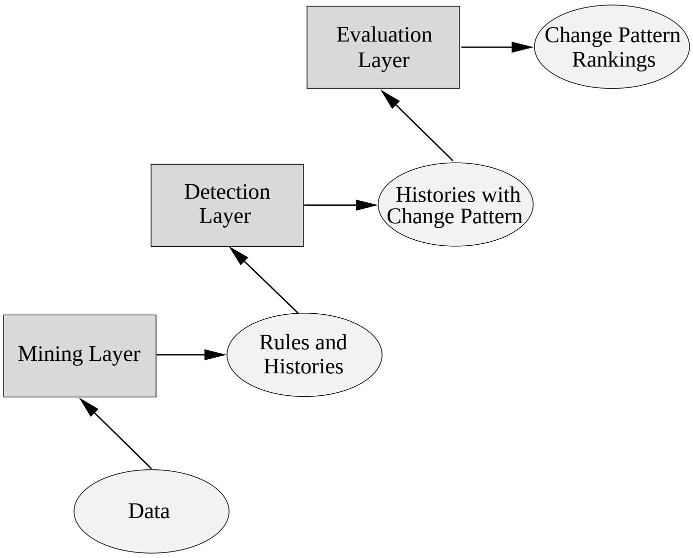


Figure 4.1: Layers and workflow of the rule change mining framework

If histories with a sufficient length are available, the next task is straightforward and constitutes the core component of rule change mining:

2. Histories which exhibit specific change patterns have to be reliably *detected*.

The discussion of association rule discovery in Chapter 2 showed that significant research is dedicated to the rule-quantity and the rule-quality problem. In fact, it is reasonable to say that primarily the development of methods for constrained mining, pruning and interestingness assessment made association rule mining suitable for a broad business usage. Since a history is derived for each rule, the rule quantity problem also affects rule change mining: it has to deal with a vast number of histories and thus many change patterns are likely to be detected. Moreover, as I will discuss in Section 5.1 pruning methods for association rules cannot be used in rule change mining.

Furthermore, there is also a quality problem: not all of the detected change patterns are equally interesting to a user and the most interesting are hidden among many irrelevant ones. This problem is partly inherited from association rule mining, because the change patterns of obvious rules, for example, might be obvious too. On the other hand, however, a change pattern may increase the interestingness of a previously uninteresting rule significantly. Overall, the third task can be formulated as:

3. Histories with a change pattern have to be pruned and *evaluated* according to their interestingness.

Because the aforementioned tasks build upon each other, they can be seen as the layers of the framework. According to their task the layers are termed *mining layer*, *detection layer* and *evaluation layer*, respectively. Figure 4.1 illustrates them and summarises the workflow.

Additionally, it should be noted that it is desirable to choose the methods within each layer such that maximum flexibility in terms of their interchangeability is attained. Particularly the evaluation layer benefits from this. Interestingness assessment is, as

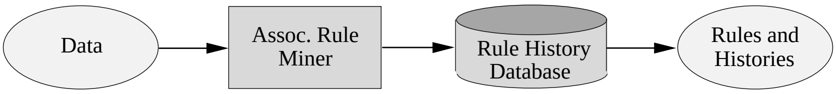


Figure 4.2: Design of the mining layer

discussed in Section 2.3, extremely user and application dependent. As opposed to the *PAM* framework, reviewed in Section 3.3.7, such a design allows the introduction of new notions of interestingness without the need to develop or modify methods in the detection layer. In contrast, a clever combination of change pattern detection and interestingness assessment, possibly by utilising the relations between rules, may lead to a decreasing number of histories to be examined and thus to a significant performance improvement. However, such approaches still need to be developed and will not be investigated further in this thesis.

4.2 The Layers

4.2.1 Mining Layer

The mining layer takes as its input a dataset collected during a specified period, applies association rule mining to it and stores the results. Its design is shown in Figure 4.2.

To obtain the dataset the period length has to chosen first. This choice is crucial and difficult because two aspects have to be considered: on the one hand a long period leads to many tuples in the dataset and thus enhances the reliability of support and confidence. On the other hand, short periods allow the measurement of a rule's statistics more frequently. This again may lead to a more reliable detection of change patterns. Nevertheless, due to practical convenience periods in business applications are commonly measured in days, weeks, months et cetera. Additionally, the maximum number of periods is often restricted, for example, to fit a quarter or a year.

After the dataset is available, association rule mining is applied to it. Basically all approaches discussed in Chapter 2 can be used. However, it is necessary to analyse if and how the requirements on the rule change mining framework constrain the choice of methods for association rule discovery. This will be done in Section 5.1.

The discovered rules and their statistics need to be stored and efficiently managed. Because the number of rules and thus the number of histories to analyse is vast, a method has to be employed which provides fast access as well as effective long-term storage. Furthermore, it is desirable that rule browsing and simple descriptive statistical analysis on histories are supported. All of these requirements can be fulfilled by a database management system in combination with *SQL* as its powerful query language. An efficient database scheme for managing rules and histories will be introduced in Section 5.2.

4.2.2 Detection Layer

The detection layer, which is illustrated in Figure 4.3, takes as its input histories from the database and applies detectors for certain change pattern types to them.

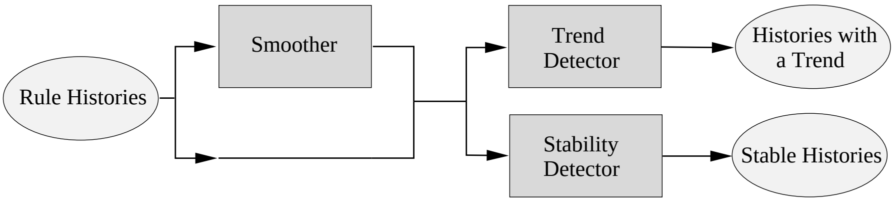


Figure 4.3: Design of the detection layer

Although change patterns can be detected directly from the histories, it is advisable to consider the application of noise reduction techniques to enhance the reliability of the subsequent change pattern detection. Particularly the reviews in the last chapter identified some existing rule change mining methods as susceptible to noise, e.g. the fuzzy approach (see Section 3.3.3) and the heuristics of the *PAM* framework (see Section 3.3.7). Many techniques for noise reduction have been developed, mainly in the fields of signal processing (Oppenheim et al., 1999) and time series analysis (Chatfield, 1996). However, in rule change mining the noise reduction method has to be applied to a vast number of histories. For this reason it is crucial that it requires no strong assumptions about the process that generated a history. Furthermore, it needs to be fast yet effective. A broadly used method that meets these requirements is *double exponential smoothing*. It will be discussed in Section 6.1.

The core components of the detection layer are the detectors for different types of change patterns. The reviews in the previous chapter showed that the vast majority of change pattern types or, in a wider context, "types of change" detected by existing approaches is difficult to interpret and less useful for business applications. In this thesis the main focus will be on trend and stability detection due to the following reasons. First, these change pattern types are meaningful with respect to the domain from which the survey data used for testing in this thesis is derived. Second, as (Liu et al., 2001b) already pointed out, they qualitatively indicate the likely future performance of a history and thus support proactive business decisions. On the other hand, they are intuitive and simple. Third, the overall time span covered by the available test data is too short to detect, for example, any kind of cyclical variations. Fourth, the decision whether a history is stable or exhibits a trend is an integral part of descriptive time series analysis (Chatfield, 1996). Sections 6.2.1 and 6.2.2 discuss methods for trend detection, whereas Section 6.3 discusses the detection of stable histories.

4.2.3 Evaluation Layer

The evaluation layer, which is illustrated in Figure 4.4, takes histories exhibiting change patterns as its input, discards redundant ones and assesses the interestingness of the remaining. Consequently, the layer consists of two building blocks.

The first block helps to cope with the vast number of histories by discarding redundant ones. A history of a rule is considered redundant if it can be derived from the histories of more general rules. For this reason I term such redundant histories also as *derivative histories*. Technically, derivative history detection could also be carried out directly before the detection of change patterns. However, the computational effort in-

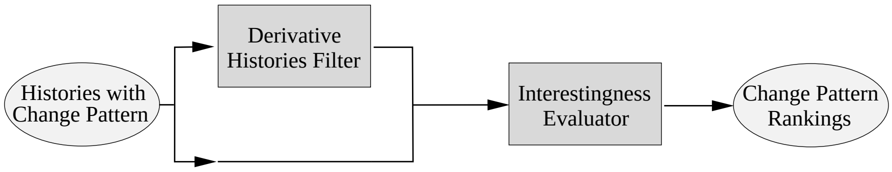

Derivative Histories Filter
Histories with Change Pattern
Interestingness Evaluator
Change Pattern Rankings

Figure 4.4: Design of the evaluation layer

volved in the detection of redundant histories by far exceeds that necessary for change pattern detection and increases significantly with the number of histories to examine. Therefore, in order to reduce the computational effort, the approach is only applied to histories containing a change pattern. Moreover, it fits conceptually into the evaluation layer, because redundant histories will rarely be of interest to a user. The approach for derivative rule detection is presented in Section 7.1.1.

The second block provides a collection of measures to assess the interestingness of detected change patterns. Rankings are generated as a result of this assessment. The interestingness measures are motivated and explained in Section 7.2.

4.3 Experimental Evaluation: Survey Analysis

Companies continuously launch new products, improve their services or place advertising campaigns but they are simultaneously aware that their customers are exposed to many other influences, like changes in lifestyle or the products and campaigns of competitors. It is crucial for companies to detect trends and stabilities in consumer behaviour and in the attitude of customers towards certain products and services. For this reason customer surveys are regularly conducted.

A typical survey consists of several question blocks, each block related to a certain topic. In order to analyse mid and long term changes the same questions are repeatedly asked over a longer period. A business analyst usually connects two objectives with the analysis of surveys. First, he aims to detect answers which frequently occur together. Particularly, he is interested in answers which significantly influence the answer to a key question. For example, he might discover that people, who reported difficulties in filling out an online fault report, are likely to be generally dissatisfied with the fault handling of the company. Second, he aims to detect changes in the frequency of answers in order to draw a conclusion on general changes in the attitude of customers.

To solve the first problem an analyst typically has several possible dependencies among answers in mind which he then verifies on the survey data. Nevertheless, many unexpected and useful associations may not be considered and hence remain undiscovered. Moreover, the dynamics of the domain are not taken into account. To solve the second problem it is common practice to calculate a set of indicators in regular intervals, whereby each indicator aggregates answers by means of descriptive statistics. The resulting time series are then examined manually. An indicator is typically intended to monitor the effects of business decisions on business goals, so that it incorporates merely a few related questions or answers, respectively. The discovered changes are therefore biased to what a user expects from its business. Many other changes, particularly those

Domain Size	#Attributes
2	7
3	6
4	4
5	6
6	1
7	5
8	1
11	1
36	1
39	1

Table 4.1: Distribution of attribute domain sizes.

Period	#Tuples	Period	#Tuples	Period	#Tuples
1	2470	9	2046	17	1511
2	2542	10	2040	18	1575
3	2946	11	2221	19	1493
4	2558	12	2194	20	1480
5	2936	13	2412		
6	2129	14	1906		
7	2229	15	1955		
8	2904	16	1749		

Table 4.2: Number of tuples in each period.

triggered by external influences, remain unrevealed.

Rule change mining supports both of the aforementioned objectives, but does not have the disadvantages of common approaches for survey analysis. In fact, survey analysis is an important application area of rule change mining. For this reason the effectiveness of the proposed rule change mining framework and its methods are tested on customer survey data. The properties of the data set will be described in the following.[1]

The dataset contains a combination of survey and internal process data collected over a period of 40 weeks. The survey is conducted on customers for whom a repair job has been finished and mainly queries their degree of satisfaction with certain aspects of fault and repair handling. The process data refers to details of the repair job itself. Each tuple is described by 33 attributes of which 19 are related to questions in the survey and 14 to internal processes. Most of the attributes are nominal. Numeric attributes are discretised by equal-frequency binning. Table 4.1 shows the domain size of the attributes

[1] For reasons of data protection it is not possible to disclose any details about the attributes. This includes their name, meaning and values. Furthermore, it is not possible to disclose interpretations or explanations of any discovered change pattern. All examples given in this thesis are completely fictitious and therefore not related to British Telecom or its customers. Exceptions are plots of histories which, unless stated otherwise, are derived from the above dataset and thus are non-fictitious.

after discretisation.

The dataset is split into 20 subsets, each related to a period of two weeks. Table 4.2 shows the size of each of the resulting subsets. For simplicity each data subset will be called dataset in the following if the context is clear. The datasets are transformed into transaction sets by recoding each (attribute, attribute value) combination as an item. The overall number of items is therefore 213. If a tuple contains missing values no item will be generated for the related attributes. Consequently, transactions theoretically contain between 1 and 33 items. In practice, however, the lower border is significantly larger.

Furthermore, the dataset has some remarkable characteristics which are challenging for both association rule and rule change mining:

- The value distribution of many attributes is rather skew. For a few attributes the tuples contain predominantly the same value, yet are the infrequent values sometimes more interesting, e.g. statements of dissatisfaction.
- Some attribute values are strongly correlated. If, in addition, the large number of attributes is taken into account, the dataset can be regarded as dense according to the description of density in Section 2.2.
- Tuples with missing values for survey related attributes are relatively frequent. In addition, the relative number of missing values for a few attributes shows an inclining trend over time. This in turn may lead to difficult interpretable trends in the frequency of attribute values.

Chapter 5

Mining Layer

Given a timestamped dataset collected during a certain period, the task of the mining layer is to discover and store the association rules hidden in it. One of the two components of this layer is therefore an association rule mining system. Typical methods of such a system have already been discussed in Chapter 2. However, it still needs to be clarified if the requirements on the rule change mining framework constrain the choice of these methods. This will be done in Section 5.1. The second component of the mining layer is a database which stores and manages rules and their histories. The design of the database will be explained in Section 5.2.

5.1 The Choice of the Association Rule Mining System

The core component of the mining layer is an association rule mining system. As laid out in Chapter 2, a typical system for association rule mining may not only consist of the rule miner itself but also of methods for pruning, constrained mining and interestingness assessment. This section discusses which of these methods are useful for a rule mining system with respect to the objectives and requirements of rule change mining.

The aforementioned methods have been developed to cope with the rule quantity and rule quality problem, respectively, in each period. The rule quality obviously does not affect rule change mining: whether or not a rule is interesting in a certain period does not directly influence the interestingness of its history. In fact, as it will be explained in Chapter 7, it is assumed that the interestingness of a change pattern primarily influences the interestingness of the underlying rule. For this reason, interestingness measures for association rules should not be part of the rule mining system.

On the other hand, the rule quantity problem affects rule change mining: a huge number of histories has to be processed and consequently far too many change patterns will be reported. Apparently, pruning and constrained mining approaches seem to be very useful in solving this problem at an early stage within the rule change mining framework. One of the requirements on rule change mining, however, is to reliably discover all histories which exhibit an interesting change pattern. In the following it will be discussed if this requirement can be met if approaches for pruning or constrained mining, respectively, are part of the rule mining system.

For the following discussion it has to be kept in mind that in order to derive rule histories, association rules have to be discovered for many transactions sets each relating to a different period. Transaction sets will be processed by a rule mining system inde-

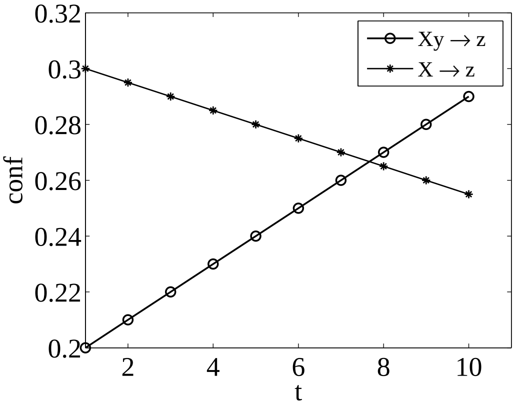


Figure 5.1: The criterion for the *informative rule set* prunes the more specific rule until period 7.

pendently from another. For pruning and constrained mining approaches this means in particular: whether a rule is discarded in a certain period does not dependent on the rule's properties with respect to other transaction sets.

Due to that, it is possible that a pruning approach discards a rule in each period, although the history of the rule is exceptionally interesting. For example, the pruning approach based on the so-called *informative rule set* discards a rule $\mathcal{X}y \rightarrow z$ if a more general rule $\mathcal{X} \rightarrow z$ with larger confidence exists (Li et al., 2004). Figure 5.1 shows two hypothetical confidence histories. Obviously, the more specific rule is pruned up to period seven and consequently no values are obtained for confidence and support. A user will therefore be unaware that the confidence history of the rule has an upward trend. Moreover, it will be not discovered that the trend contradicts the trend of the more general rule, which in turn is a rather interesting constellation. Starting from period eight the rule will not be discarded anymore, but it is impossible to state whether it was previously pruned or whether its statistics were below the support or confidence threshold. This distinction is crucial for the interpretation of the newly emerged rule. Clarification, however, is only possible by accessing historic transaction sets. Since this needs to be done for each emerged rule, the effort linked to it is rather high and may be undesirable for many practical applications.

Generally, an association rule is pruned if a certain property of it matches a given criterion. This property – in general a numerical values – depends on the transaction set. It may therefore vary for a rule over time, but still match the pruning criterion in each period. Although these changes may qualify this rule as interesting, it is discarded by association rule pruning from each rule set. Overall, pruning approaches for association rules should therefore not be used in conjunction with rule change mining.

In contrast to pruning methods, approaches for constrained mining access only time-invariant properties of a rule. Item, length and aggregate contraints are typical examples (see Section 2.2). They basically allow a user to select rules that he a priori considers interesting to be analysed by rule change mining. For example, a user may only be interested in change patterns of rules with less than five items because they are easier to interpret than those of longer rules. Overall, constraint mining will detect change patterns in all rules considered as relevant by a user.

In summary, the requirements on rule change mining constrain the choice of methods for association rule discovery. While approaches for constrained rule mining are considered to be uncritial, interestingness assessment and in particular pruning approaches may discard rules with a rather interesting history. For this reason they should not be used in conjunction with rule change mining.

5.2 Efficient Management of Association Rules

To allow fast access and long-term storage, rules and their histories are managed within a database. The most frequent operations on such a database are:

- Determining if a given rule is contained in the database.
- Retrieving the histories for a given rule

The tables and indexes of a database scheme which efficiently supports the above operations will be explained in the following. The core component is the table

RINV(RULE_ID, RBODY_ID, RHEAD_ID)

which stores all discovered rules whereby a unique identifier is assigned to each. A rule is defined by its antecedent ('body') and consequent ('head'), each of which is described by another identifier. Besides the reduction of redundancy this also supports the retrieval of rules with either the same antecedent or consequent. A fast yet simple way to derive such identifiers is to lexicographically sort and then concatenate the descriptors of the constituting items. The mapping of each antecedent identifier to single items is stored in the table.

RBODY(RBODY_ID, ITEM_ID)

To check if a rule is already contained in the database, its antecedent and consequent identifiers need to be calculated and the corresponding columns of the RINV table queried. To increase the speed of this operation it is recommended to generate an index for the combination of both.

The start date and the end date of, and a unique identifier for each period are stored in the table

PERIOD(PERIOD_ID, PSTART, PEND)

A rule's statistics for each period in which it has been discovered is stored in the table

RHISTORY(RULE_ID, PERIOD_ID, SUPP, ASUPP, CONF, ...)

In order to retrieve the histories for a rule its identifier must first be obtained from the RINV table. The identifier is then used to query the corresponding column of the RHISTORY table. Since this table is typically very large it is recommended to create an index on RULE_ID.

Period	#Rules	Period	#Rules	Period	#Rules
1	204516	9	236919	17	184208
2	223021	10	233194	18	179777
3	224640	11	210875	19	179338
4	221278	12	194439	20	175013
5	214162	13	192414		
6	218946	14	178058		
7	211106	15	186025		
8	235147	16	188339		

Table 5.1: Number of rules discovered in each period.

5.3 Experimental Evaluation

To obtain rule histories as the basis for all further experiments in this thesis, Ch. Borgelt's implementation of the well-known *apriori* algorithm[1] has been period-wise applied to the transaction sets described in Section 4.3. The algorithm's parameters – lower thresholds for support and confidence – are chosen as $\text{supp}_{\text{min}} = 0.05$ and $\text{conf}_{\text{min}} = 0.2$, respectively. Such low thresholds are necessary because some items of interest, like statements of dissatisfaction, are rather infrequently contained in the survey data. To restrict the number of generated rules several item constraints have been defined based on background knowledge about the underlying business scenario. They state whether certain items should either appear in a rule's antecedent or in its consequent. Table 5.1 shows the cardinality of the rule sets discovered in each period. The compound rule set is generated from these rule sets by intersecting them. Its cardinality is 77401, i.e. it contains only 44% of the rules of the smallest discovered rule set.

[1] The program can be obtained from http://fuzzy.cs.uni-magdeburg.de/∼borgelt/apriori.html.

Chapter 6

Detection Layer

The task of the detection layer is to discover change patterns in rule histories. In this thesis, however, only the detection of histories which are stable or exhibit a trend is discussed. The detection layer fulfills its task by a two step approach. In the first step a filter is applied to the histories to reduce the contained noise. Section 6.1 discusses the broadly used method of *double exponential smoothing.* In a second step statistical tests for trend and stability are conducted. The tests are described in Sections 6.2 and 6.3, respectively. The chapter closes with an experimental evaluation of the methods.

6.1 Smoothing

Inherent with any kind of data collected over time, like rule histories, are random variations also referred to as *noise.* These random variations may influence subsequent analysis steps such that wrong and misleading results are produced. Particularly some existing rule change mining approaches are vulnerable to noise (cp. Section 3.3). An often used technique for reducing this effect is *smoothing.* When properly applied it reveals more clearly any underlying trend or stability.

One popular and widely used class of methods to smooth a sequence is *exponential smoothing.* Let $(v_0, \ldots, v_n)$ be a sequence, for any $i = 0, \ldots, n$ the smoothed value $\hat{v}_i$ is found by computing:

$$\hat{v}_i = \begin{cases} v_0 & : \quad i = 0 \\ \alpha v_i + (1-\alpha)\hat{v}_{i-1} & : \quad i = 1, \ldots, n \end{cases} \tag{6.1}$$

This scheme is called *single exponential smoothing.* The constant $\alpha \in [0,1]$ is also called the *smoothing constant*, or *smoothing parameter*, and controls the influence of past observations. This becomes clearer if (6.1) is rewritten as:

$$\hat{v}_i = \alpha v_i - \alpha \sum_{j=1}^{i-1} (1-\alpha)^{i-j} v_j + (1-\alpha)^i v_0 \tag{6.2}$$

Obviously, the smoothed value $\hat{v}_i$ is calculated by combining the recent values v_i with the weighted sum of all past values. Since the weights $(1-\alpha)^k$ decrease exponentially, a more recent value has a relatively higher weight than an older one. A small smoothing parameter α leads to a slow decrease of the weights and thus to a stronger dampening

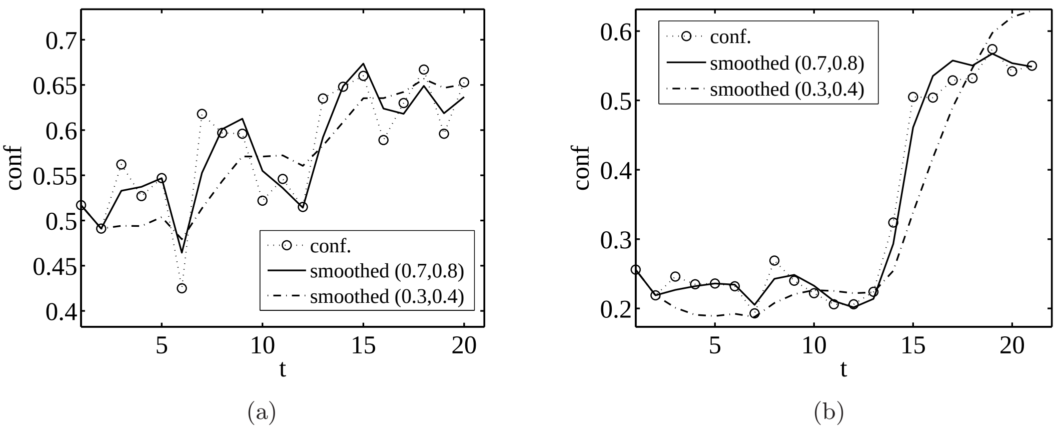


Figure 6.1: Example for exponential smoothing for different shapes and parameter settings, denoted as (α, γ)

effect while a large α will emphasise recent value. The marginal case $\alpha = 1$ yields $\hat{v}_i = v_i,\ i = 0, \ldots, n$, i.e. smoothed and original sequence are identical.

Single exponential smoothing is not very effective in presence of a trend. It tends to overestimate a downward and to underestimate an upward trend (NIST, 2004). This problem is resolved by the following generalisation, known as *double exponential smoothing*:

$$\hat{v}_i = \begin{cases} v_0 & : \quad i = 0 \\ \alpha v_i + (1-\alpha)(\hat{v}_{i-1} + b_{i-1}) & : \quad i = 1, \ldots, n \end{cases} \tag{6.3}$$

$$b_i = \begin{cases} v_1 - v_0 & : \quad i = 0 \\ \gamma(\hat{v}_i - \hat{v}_{i-1}) + (1-\gamma) b_{i-1} & : \quad i = 1, \ldots, n-1 \end{cases} \tag{6.4}$$

Compared to Equation (6.1) for single exponential smoothing, (6.3) adjusts $\hat{v}_i$ directly for the (smoothed) change b_{i-1} of the previous period by adding it to the last smoothed value $\hat{v}_{i-1}$. Equation (6.4) shows that b_{i-1} is calculated by single exponential smoothing, whereas now changes $(\hat{v}_i - \hat{v}_{i-1})$ are smoothed. Consistently with the notes on parameter choice for single exponential smoothing a large γ emphasises recent changes, while a low γ leads to more equally distributed weights. In the latter case the smoothed sequence, for example, follows a trend more aggressively and may overshoot, if the sequence's shape has a rather pronounced curvature. Moreover, it takes longer until the influence of the initial value b_0 vanishes and hence the first smoothed values might be unreliable. Both effects can be seen in Figure 6.1(b) for the parameters $\alpha = 0.3$ and $\gamma = 0.4$. In contrast, the same parameters perform considerably well for the sequence shown in Figure 6.1(a).

As this already indicates, the choice of a best pair of smoothing parameters can be a rather tedious task. In fact, both parameters are interrelated and must be chosen accordingly. Optimal values are often not found without applying non-linear optimisation techniques (Chatfield, 2001). Particularly when applied to the vast number of sequences typically encountered in rule change mining, the necessary effort would be enormous. Therefore, the parameters have to be chosen manually based on experience. Since the smoothed sequence converges for $\alpha \to 1$ and $\gamma \to 1$ towards the original sequence with the marginal case $\hat{v}_i = v_i,\ i = 0, \ldots, n$ for $\alpha = \gamma = 1$, a conservative parameter choice leads to more predictable smoothing effects. Figure 6.1 shows two examples for the effect of the parameters $\alpha = 0.7$ and $\gamma = 0.8$.

Whenever exponential smoothing is used to denoise rule histories, the following implications have to be considered:

1. After smoothing, statistical measures may be inconsistent among each other. For example, given support, antecedent support and confidence histories, the equation $\operatorname{conf}(r, T_i) = \operatorname{supp}(r, T_i)/\operatorname{asupp}(r, T_i)$ will generally not hold after smoothing at least one of them.

2. Although such scenarios are not considered in this thesis, the focal point in some applications may be on significant outliers.[1] In this case smoothing should be applied very carefully, since it might disturb or discard them.

Overall, double exponential smoothing is a simple, fast yet effective method, which can easily be automated. It has the problem of choosing appropriate smoothing parameters. For this reason the choice should be conservatively unless further knowledge about the sequences is available or non-linear optimisation techniques can be applied with reasonable effort.

6.2 Trend Detection

A trend is present when a sequence exhibits steady upward growth or a downward decline over its whole length. This definition is rather loose, but in fact there exists no fully satisfactory definition for trend (Chatfield, 2001). For example, the perception of trend depends partly on the length of a sequence, i.e. what seems to be a trend over a short time span, might be on a long-term scale just part of a cyclical variation.

From the data mining perspective a trend describes the pattern that each value is likely to be larger or smaller than all its predecessors within a sequence, depending if the trend is upward or downward. Thus it is a qualitative statement about the current and likely future development of a sequence. From the aspects of interpretability and usefulness such a statement is sufficient in the case of rule change mining. Facing the vast number of rules and their corresponding histories, an user often has a basic expectation if and what sort of trend they should exhibit. Moreover, by comparing his expectations with the reality he will mostly be able to roughly judge the implications for its business. On the other hand a user will rarely know in advance how trends should look like quantitatively, e.g. their shape or target values. For this reason he may be unable to exploit the advantages of more sophisticated trend descriptions, like regression models.

To choose a method for trend detection, it has be considered that the number of sequences to examine is vast. Whenever a trend is reported the user is basically forced to rely on the correctness of this statement, because it is infeasible for him to verify each trend manually. In addition to the requirement of reliable detection, the method should incorporate no assumptions about any underlying model, because it is very unlikely that it will hold for all or at least the most sequences. For these reasons non-parametric statistical tests are the appropriate choice for trend detection, because they operate

[1] To say it loosely: ‘One researcher’s noise is another’s signal.’

without any model assumptions. Furthermore, the probability that a trend is detected although the sequence actually contains none is bounded by the confidence level.

In Section 3.3.5 an approach has been reviewed which utilises a nonparametric statistical method known as the *runs test* to detect trends. It has been shown that this test is inappropriate for trend detection. There exist several other, more appropriate nonparametrical statistical tests, most of them widely used in ecology and environmental science. While rarely discussed in classical statistical textbooks, in-depth descriptions of them can be found in textbooks on environmental statistics, like (Gilbert, 1987) and (Helsel and Hirsch, 1992). The commonly used methods of *Mann-Kendall* and *Cox-Stuart* will be discussed in the following.

6.2.1 The Method of Mann-Kendall

Let $H = (v_1, \dots, v_n)$ be a history for a rule. If its values v_i are independently drawn, the following method proposed by (Mann, 1945) can be used to test the following hypotheses:

$$\begin{array}{lll} A) & H_{01} & : \text{There is } \textbf{no} \text{ upward trend present in } H \\ & H_{11} & : \text{There is } \textbf{an} \text{ upward trend present in } H \\ \\ B) & H_{02} & : \text{There is } \textbf{no} \text{ downward trend present in } H \\ & H_{12} & : \text{There is } \textbf{a} \text{ downward trend present in } H \\ \\ C) & H_{03} & : \text{There is } \textbf{no} \text{ trend present in } H \\ & H_{13} & : \text{There is } \textbf{a} \text{ trend present in } H \end{array}$$

To test if the null hypothesis has to be rejected, the following test statistic is calculated:

$$C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(v_j - v_i) \tag{6.5}$$

whereby sgn denotes the following function:

$$\operatorname{sgn}(v_j - v_i) = \begin{cases} 1 & : \ v_j - v_i > 0 \\ 0 & : \ v_j - v_i = 0 \\ -1 & : \ v_j - v_i < 0 \end{cases} \tag{6.6}$$

The intuition behind the test statistic will be explained on an upward trend. If a clear upward trend is present, then for each value v_i most of its succeeding values $v_{i+1}, \dots, v_n$ are larger and hence most of the differences $(v_j - v_i)$, $j = i+1, \dots, n$ positive. With decreasing clarity of the trend, the number of smaller successors and thus the number of negative difference increases. If, finally, no trend is present at all, the number of positive and negative differences is approximately equal. All this is encoded in the test statistic, which counts cases when a value is lower than, but discounts cases when a value is greater than a successor. From this it follows that the likeliness of an upward trend grows with the test statistic's value.

Given the test statistic C and the confidence level α, the null hypothesis is rejected if

$$\begin{array}{llllll} A) & C & > & +K_{n;1-\alpha} & & \\ B) & C & < & -K_{n;1-\alpha} & & \\ C) & C & < & -K_{n;1-\alpha/2} & \vee \quad C \quad > & K_{n;1-\alpha/2} \end{array} \tag{6.7}$$

n	$K_{n;0.995}$	$K_{n;0.99}$	$K_{n;0.975}$	$K_{n;0.95}$	$K_{n;0.90}$
4	6	6	6	6	5
5	10	9	8	8	7
6	14	13	11	9	7
7	17	15	13	11	9
8	21	19	16	14	11
9	25	23	19	16	13
10	29	26	23	19	15
15	52	47	40	34	27
20	80	72	59	49	38

Table 6.1: Values of Kendall's K-statistic for the significance levels $\alpha = 0.005$, $\alpha = 0.01$, $\alpha = 0.025$, $\alpha = 0.05$ and $\alpha = 0.1$

.

The values $K_{n;1-\alpha}$ denote the $(1-\alpha)$-quantiles of Kendall's K-statistic which also depends on the length n of the sequence. For their values for up to $n = 20$ are shown in Table 6.1. If n is sufficiently large ($n > 20$) the transformed test statistic

$$C^* = \frac{C}{\sqrt{n(n-1)(2n+5)/18}} \tag{6.8}$$

is approximately $N(0,1)$ distributed. In this case C^* and the $(1-\alpha)$-quantiles of the standard normal distribution can be used to replace their counterparts in (6.7).

Unlike parametrical tests for trend, non-parametrical tests cannot use metric information about the change between values, since this would require model assumptions. Instead they have to exploit the greater-than and lower-than relations between values. From this perspective the method of Mann-Kendall has the desirable property that it incorporates the relation between *all* pairs of values and thus – if triples et cetera are not considered – uses the maximum on features of the sequence. For this reason it can be expected that the method is relatively robust to noise. The price for this property is the computational effort to carry out the test:

$$\sum_{i=1}^{n-1} i = (n-1)n/2 = O(n^2), \tag{6.9}$$

This means that the effort increases quadratically with the length of the sequence.

The power of the Mann-Kendall test, i.e. the probability of rejecting the null hypothesis when it is false, has been empirically analysed by (Yue et al., 2002). They came to the result that the power of the test is an increasing function of the sample size and of the absolute slope of the trend, whereas the power approaches zero if the slope goes to zero. Furthermore, the power is a decreasing function of the sequence's variation, because it simply masks the existence of a trend.

6.2.2 The Method of Cox-Stuart

Let the sequence $H = (v_1, \ldots, v_n)$ be a rule history. Like the method of Mann-Kendall, the method proposed by (Cox and Stuart, 1955) assumes that the values v_i are independently drawn from each other. It can be used to test the following hypotheses:

$$\begin{array}{lll} A) & H_{01} : & \text{There is } \textbf{no} \text{ upward trend present in } H \\ & H_{11} : & \text{There is } \textbf{an} \text{ upward trend present in } H \\ \\ B) & H_{02} : & \text{There is } \textbf{no} \text{ downward trend present in } H \\ & H_{12} : & \text{There is } \textbf{a} \text{ downward trend present in } H \\ \\ C) & H_{03} : & \text{There is } \textbf{no} \text{ trend present in } H \\ & H_{13} : & \text{There is } \textbf{a} \text{ trend present in } H \end{array}$$

The first step of the approach is to form pairs:

$$(v_{m+1},\ v_1),\ (v_{m+2},\ v_2), \ldots, (v_n,\ v_{n-m}) \tag{6.10}$$

whereby $m = n/2$, if n is an even number, and $m = (n+1)/2$, if n is an odd number. Thereafter a plus sign replaces each pair $(v_{i+m},\ v_i)$ for which v_{i+m} is greater than v_i, and a minus sign replaces each pair for which v_{i+m} is lower than v_i. Ties are omitted.

The basic idea is to regard the plus and minus signs as the outcome of a Bernoulli experiment which is repeated under the same conditions. If the null hypothesis of no trend holds, it is reasonable to assume that the results of the trials are neither biased to a plus nor to a minus sign, i.e. the first value in the pairs above does not generally tend to be larger or smaller than the second. Hence each of the two signs should occur with the same probability of 0.5. If the observed distribution of signs is very unlikely under this assumption, the null hypothesis will be rejected.

Let n_p and n_m be the observed counts of plus and minus signs. They are instantiations of random variables, N_p and N_m respectively. Because N_p and N_m model the counts of outcomes of a Bernoulli experiment, they are binomial distributed. Provided that the null hypothesis of no trend holds the distributions are identical, i.e. $N_p \sim B(n_p+n_m, 0.5)$ and $N_m \sim B(n_p + n_m, 0.5)$.

The actually used test statistic depends on the tested null hypothesis. The test statistic for hypothesis A is the number n_p of plus signs, and the test statistic for B the number of minus signs n_m. The test statistic for hypothesis C is the number of plus, or the number of minus signs, whichever is larger.

Let $n' = n_p + n_m$. Depending on the chosen null hypothesis the critical values for a given significance level α are obtained as follows:

A) Choose the smallest value $B_{n',1-\alpha}$ such that $P(N_p > B_{n',\alpha}) < \alpha$.

B) Choose the smallest value $B_{n',1-\alpha}$ such that $P(N_m > B_{n',\alpha}) < \alpha$.

C) Choose the smallest $B_{n',\alpha/2}$ such that $P(N_p > B_{n',\alpha/2}) < \alpha/2$, if $n_p > n_m$, or $P(N_m > B_{n',\alpha/2}) < \alpha/2$ otherwise.

The null hypothesis is rejected if:

$$\begin{array}{lll} A) & n_p & > B_{n',\alpha} \\ B) & n_m & > B_{n',\alpha} \\ C) & \max(n_p, n_m) & > B_{n',\alpha/2} \end{array} \tag{6.11}$$

For large n' the random variables for the test statistics, N_p and N_m, can be transformed such that they approximately follow a standard normal distribution. Without loss of generality the transformation is only given for N_p:

$$N_p^* = \frac{(N_p + 0.5) - 0.5n'}{0.5\sqrt{n'}} \tag{6.12}$$

The transformed test statistic and the quantiles of the standard normal distribution can then be used to replace their counterparts in (6.11).

In contrast to the Mann-Kendall test, the Cox-Stuart test utilises less features of the sequence. While the former uses $(n-1)n/2$ relations between pairs of values, the latter uses maximally $n/2$. This leads to a computational effort for the Cox-Stuart test of $O(n)$, i.e. it increases linearly with the sequence length. On the one hand, it may render the approach susceptible to noise, because the influence of each utilized relation on the test result is more significant than for the Mann-Kendall test. On the other hand, the Cox-Stuart test is in particular for long sequences faster. Depending on the actual application scenario is has therefore to be determined if which of the two issues is more important.

A further difference to the Mann-Kendall test is the treatment of pairs of equal values. While the test statistic of Mann-Kendall accounts for them, they are explicitely excluded by the Cox-Stuart test. Particularly in sequences with many zero differences, the number of actually used pairs of values might be low. This in turn increases the susceptibility to noise and thus leads to unreliable test results.

6.3 Stability Detection

Roughly, a history is considered stable if its mean level and variance are constant over time and additionally the variance reasonable small. Similar to trends, a clear definition of stability is difficult. For example, a sequence may exhibit a cyclical variation, but nonetheless be stable on a long term scale. Depending on the problem domain, either the one or the other would be emphasised.

From the data mining perspective stability describes the pattern that each value is likely to be close to a constant value, estimated by the mean of its predecessors. Thus it is, like a trend, a qualitative statement about the future development of a sequence. However, in contrast to a trend it can be easily modeled in an interpretable and useful way, e.g. by the sequence's sample mean and sample variance. As (Liu et al., 2001b) pointed out, stable rules are more reliable and can be trusted – an eminently useful and desirable property for long term business planing.

An apparent, simple approach to detect a stable sequence is to define a fix interval around its mean. A history is tested stable if all values are within the interval. This approach has the weakness that histories with a slight trend will be classified as stable. Furthermore, the choice of a reasonable interval to express the maximum tolerable noise level may be difficult. Particularly this is complicated by the fact, that a user would intuitively judge a history stable even if a few, isolated values are outside the interval.

Obviously, more flexibility for stability detection is needed. This can be provided by the χ^2 test based method suggested by (Liu et al., 2001b). Nonetheless, in Section 3.3.5 it has been shown that this method sometimes detects trends as stabilities. To circumvent this problem I propose the following two step approach.

	1	...	n
#supp.	$\lvert\mathcal{D}(T_1)\rvert\ \mathrm{supp}(r, T_1)$	...	$\lvert\mathcal{D}(T_n)\rvert\ \mathrm{supp}(r, T_n)$
#not supp.	$\lvert\mathcal{D}(T_1)\rvert\ (1 - \mathrm{supp}(r, T_1))$	...	$\lvert\mathcal{D}(T_n)\rvert\ (1 - \mathrm{supp}(r, T_n))$

Table 6.2: $2 \times n$ contingency table for support

	1	...	n
#supp.	$\lvert\mathcal{D}_{\mathcal{X}}(T_1)\rvert\ \mathrm{conf}(r, T_1)$	...	$\lvert\mathcal{D}_{\mathcal{X}}(T_n)\rvert\ \mathrm{conf}(r, T_n)$
#not supp.	$\lvert\mathcal{D}_{\mathcal{X}}(T_1)\rvert\ (1 - \mathrm{conf}(r, T_1))$	...	$\lvert\mathcal{D}_{\mathcal{X}}(T_n)\rvert\ (1 - \mathrm{conf}(r, T_n))$

Table 6.3: $2 \times n$ contingency table for confidence

1. Test the sequence for trend, for example with the methods described in Section 6.2. Generally, if other change patterns are mined as well, test for all patterns that contradict stability.
2. If no change patterns are detected in Step 1, test for stability with the χ^2 test.

The basic idea and the pros and cons of utilising the χ^2 test for stability detection have already been discussed in Section 6.2. The technical details will be outlined in the following, first for support histories and then for confidence histories.

Support Histories. The support history $(\mathrm{supp}(r, T_1), \ldots, \mathrm{supp}(r, T_n))$ of the rule $r : \mathcal{X} \Rightarrow y$ is considered stable, if the χ^2 test fails to reject the following null hypothesis:

$$\begin{aligned} H_0 &: \mathrm{supp}(r, T_1) = \mathrm{supp}(r, T_2) = \ldots = \mathrm{supp}(r, T_n) \\ H_1 &: \text{the supports are not all equal} \end{aligned}$$

The first step of the test is to create the $2 \times n$ contingency table shown in Table 6.2. The first row contains for each $\mathcal{D}(T_i)$ the absolute number of transactions which support $\mathcal{X} \cup \{y\}$. The second row contains for each $\mathcal{D}(T_i)$ the absolute number of transactions which do *not* support $\mathcal{X} \cup \{y\}$. These numbers are called *observed frequencies* $O_{ij},\ i = 1, 2\ ;\ j = 1, \ldots, n$. If H_0 is true, the expected number of supported and accordingly not supported transactions in each cell is

$$E_{ij} = \begin{cases} (m_j \sum_{j=1}^{n} O_{ij})/|\mathcal{D}| & : \quad i = 1 \\ (m_j \sum_{j=1}^{n} O_{ij})/|\mathcal{D}| & : \quad i = 2 \end{cases} \tag{6.13}$$

The χ^2 test statistic is defined as

$$\hat{\chi}^2 = \sum_{i=1}^{2} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{6.14}$$

If the null hypothesis holds it follows a χ^2 distribution with $(2 - 1)(n - 1)$ degrees of freedom. The null hypothesis H_0 is rejected at significance level α, if the test statistic exceeds the threshold value $\chi^2_{1-\alpha}$ of the χ^2 distribution with $(n - 1)$ degrees of freedom.

Confidence Histories. The test for stability of a confidence history is similar to the one for support. Merely the hypotheses and the contingency table need to be changed, whereas the testing procedure itself stays the same. Let $(\text{conf}(r, T_1), \ldots, \text{conf}(r, T_n))$ be the confidence history of the rule $r : \mathcal{X} \Rightarrow y$. The history is considered stable, if the χ^2 test fails to reject the following null hypothesis:

$$\begin{aligned} H_0 &: \text{conf}(r, T_1) = \text{conf}(r, T_2) = \ldots = \text{conf}(r, T_n) \\ H_1 &: \text{the confidences are not all equal} \end{aligned}$$

The $2 \times n$ contingency table is shown in Table 6.3, whereby the subset of transactions in each $\mathcal{D}(T_i)$ that is supported by the antecedent $\mathcal{X}$ is denoted by $\mathcal{D}_{\mathcal{X}}(T_i)$. The first row of the table contains for each $\mathcal{D}_{\mathcal{X}}(T_i)$ the absolute number of transactions which support the consequent y. The second row contains for each $\mathcal{D}_{\mathcal{X}}(T_i)$ the absolute number of transactions which do *not* support the consequent y. The subsequent steps of the testing procedure are equal to those for support.

6.4 Experimental Evaluation

To test the effectiveness of the proposed methods, the Mann-Kendall test, the Cox-Stuart test and the stability test have been applied to the histories described in Section 5.3. All experiments have been carried out for both smoothed and unsmoothed sequences. Several objectives are linked to the evaluation. First, the number of trends and stabilities contained in histories has to be determined. Second, the results of the Mann-Kendall and Cox-Stuart test have to be compared to each other. Third, the influence of smoothing on the change pattern detectors has to be analysed. It should be clear that a qualitative comparison between methods is hardly possible due to missing formal definitions of trend and stability. For this reason the analysis will be quantitatively, based on the percentage of histories for which a certain pattern has been detected. However, some histories will be used to exemplify certain issues.

The results for the Mann-Kendall and Cox-Stuart test are shown in Table 6.4 and Table 6.5, respectively. Furthermore, the results for stability detection are included in both tables, since they depend on the outcome of a prior trend detection.

By comparing the corresponding columns for unsmoothed histories in Table 6.4 and Table 6.5 it can be seen that the number of change patterns detected by the Mann-Kendall test is very similar to those detected by the Cox-Stuart test. Roughly 50% of support and antecedent support histories exhibit a trend, whereas the number of confidence histories with a trend is considerably smaller. On the other hand, around 30% of confidence histories are stable, compared to under 3% for both types of support. The significant difference can be explained with the density of the data. Since some items are highly correlated, it is very likely that many rules have a stable history of high confidence values. The support history of such rules, nonetheless, may exhibit a trend.

Apparently, more trends are detected by the Mann-Kendall than by the Cox-Stuart test. In this context it is interesting to analyse how histories, for which a trend has been detected, are distributed among both tests, i.e. how many trends are detected by both tests and how many by exclusively one. Table 6.6 displays the results of this analysis. In most cases, obviously both methods detected a trend. However, the trends exclusively detected by the Mann-Kendall test significantly outnumber those exclusively discovered by the Cox-Stuart test. One reason for this difference may be the already mentioned noise susceptibility of the Cox-Stuart test. For example, Figure 6.2 shows a history for which a trend is detected by the method of Mann-Kendall. The test statistic is $C = -73$,

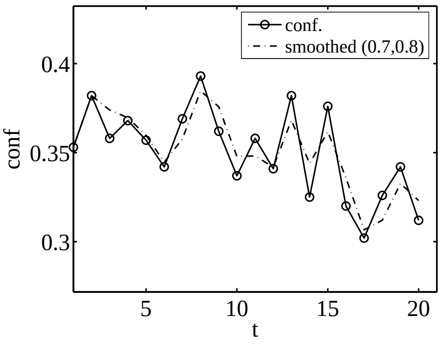


Figure 6.2: As opposed to the Cox-Stuart test, the Mann-Kendall test detects a trend in this history. After smoothing *a* trend is detected by both methods.

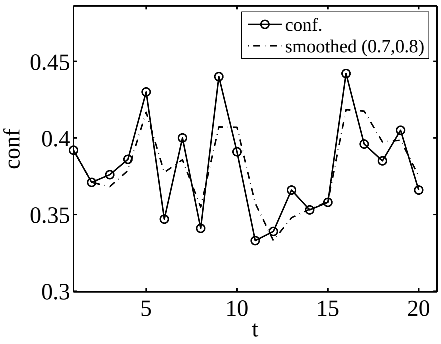


Figure 6.3: As opposed to the Mann-Kendall test, the Cox-Stuart test detects a trend in this history. After smoothing *no* trend is detected by both methods.

which is smaller than the threshold value of -43 at a significance level of 0.05. This trend, however, is not detected by the method of Cox-Stuart, for which the test statistic $n_m = 7$ is not greater than the threshold value of 7. The trend would be detected by the Cox-Stuart test if, for example, the value in period three was greater than the one in period eleven. In contrast, Figure 6.3 shows a trend detected by the Cox-Stuart test. The test statistic is $n_p = 8$ which is greater than the threshold value of 7 at a significance level of 0.05. The method of Mann-Kendall detects no trend, its test statistic is $C = 3$. Both examples have in common that the outcome of the Mann-Kendall test is consistent with the user's intuition.

To analyse the effects of smoothing on change pattern detection the above analysis was additionally carried out on smoothed histories. The method of double exponential smoothing was used with the smoothing parameters $\alpha = 0.7$ and $\gamma = 0.8$. As it can easily be seen in Table 6.4 and Table 6.5 smoothing leads to an increase in the number of detected trends and stabilities. In particular the number of stable support and antecedent support histories is almost doubled. In Section 6.1 smoothing was motivated by the need of noise reduction methods to improve the reliability of change pattern discovery. Therefore it has to be determined how many trends are newly detected and how many trends are no longer detected after smoothing. Table 6.7 shows the results. As already expected from the figures in Table 6.4 and Table 6.5 are trends which are no longer detected after smoothing outnumbered by newly detected trends. Particularly for the Cox-Stuart is the relative number of no longer discovered trends significantly larger than for the Mann-Kendall test. This, in turn, could be a further evidence for the test's noise sensitivity.

To illustrate the influence of smoothing on the test outcome, the two examples from above are continued. Both, the Mann-Kendall and the Cox-Stuart test, detect a trend in the history displayed in Figure 6.2 after smoothing. On the other hand, no trend is detected in the history shown in Figure 6.3 by both methods anymore.

	no smoothing				smoothing			
	trend(%)			stable(%)	trend(%)			stable(%)
	down	up	all		down	up	all	
conf	21.2	18.0	39.3	28.1	25.12	20.5	45.6	32.6
asupp	36.8	20.2	57.0	2.1	38.87	21.1	60.0	4.3
supp	37.5	17.1	54.7	2.6	40.07	18.1	58.2	5.1

Table 6.4: Mann-Kendall test: Fraction of unsmoothed and smoothed histories with trend or stability.

	no smoothing				smoothing			
	trend(%)			stable(%)	trend(%)			stable(%)
	down	up	all		down	up	all	
conf	16.3	18.2	34.5	29.3	18.9	20.4	39.2	35.0
asupp	31.8	19.2	51.0	2.2	32.1	19.8	51.8	4.9
supp	31.2	16.2	47.4	2.8	32.5	17.0	49.5	5.8

Table 6.5: Cox-Stuart test: Fraction of unsmoothed and smoothed histories with trend or stability.

	MK and CS		MK – CS		CS – MK	
	no sm.(%)	sm.(%)	no sm.(%)	sm.(%)	no sm.(%)	sm.(%)
conf	28.4	33.3	10.9	12.3	6.1	6.0
asupp	46.4	48.0	10.6	12.0	4.6	3.8
supp	43.1	45.8	11.6	12.4	4.3	3.7

Table 6.6: Comparison of the Mann-Kendall (MK) and Cox-Stuart (CS) test: Fraction of unsmoothed (no sm.) and unsmoothed (sm.) histories for which a trend has been detected by both the MK and CS test, only by the MK test and only by the CS test, respectively.

	Mann-Kendall		Cox-Stuart	
	no sm.(%)	sm.(%)	no sm.(%)	sm.(%)
conf	2.4	8.8	4.4	9.1
asupp	2.6	5.6	4.7	5.5
supp	2.4	5.9	4.4	6.5

Table 6.7: Influence of Smoothing: Fraction of histories for which (1) a change pattern is only detected if smoothing is not applied (no sm.) and (2) a change pattern is only detected if smoothing is applied (sm.).

Chapter 7

Evaluation Layer

Because usually a vast number of change patterns will be detected is it necessary to provide methods which reduce their number and identify potentially interesting ones. This is the task of the evaluation layer. To reduce the number of change patterns the evaluation layer contains a novel pruning approach, based on so-called derivative histories. The approach is proposed and evaluated in Section 7.1. Although this approach discards many change patterns, their number may still be to large for manual examination. For this reason the evaluation layer provides a set of interestingness measures which are described in Section 7.2.

7.1 A Pruning Approach

Association rule mining discovers the exhaustive set of all hidden associations within the data which exceed thresholds on support and confidence. This set is usually vast and hardly presentable, the truly interesting rules are hidden within many redundant or obvious ones (see Section 2.2). As the experimental evaluation in the preceeding chapter showed, this rule quantity problem transfers directly to rule change mining: a vast number of histories have to be analysed and finally far too many change patterns are reported.

Generally, most changes captured in a history and consequently also change patterns are simply the snowball effect of the changes of other rules. For example, given the database of surveys the following rule could have been discovered:

$$r_1 : \textsc{Broadband} \rightarrow \textsc{High Internet Usage}$$

A change pattern could be that the support of the rule, i.e. the fraction of users with broadband internet and high internet usage shows an upward trend. However, if the fraction of males among all broadband users with high usage is stable over time, the history of

$$r_2 : \textsc{Male, Broadband} \rightarrow \textsc{High Internet Usage}$$

shows qualitatively the same trend. In fact, the history of r_2 can be *derived* from the one of r_1 by multiplying it with a gender related constant factor.

It is reasonable to assume that a user will generally be interested in rules with non-derivative histories, because they are likely key drivers for changes. This is also supported by similar statements by many researchers (Shah et al., 1999; Liu et al., 2001a). As

(Chakrabarti et al., 1998) pointed out: "If the support of the itemset changes over time, it is not considered interesting if the changes are totally explained by the changes in the support of smaller subsets of items".

Moreover, derivative rules may lead to wrong business decisions. In the above example a decision based on rule r_2 would account for the gender as one significant factor for the observed trend. In fact, the gender is completely irrelevant.

As the discussion in Section 5.1 showed, conventional approaches for association rule are incapable of solving the problems stated above. For this reason there is a significant need for approaches to detect and discard rules which are derivative with respect to their histories and thus provide more clarity about possible key drivers for change. In a way this can be seen as a form of *pruning*.

The approach of (Liu et al., 2001a) aims at the same objective. Its review in Section 3.3.6, nonetheless, showed that it can only be applied to support and confidence histories of exactly two periods. Furthermore, it yields counterintuitive results, mainly due to the chosen criterion and the utilised statistical test.

In the following I present a formal notion of derivativeness and propose a collection of meaningful criteria and an adequate testing procedure. The effectiveness of the proposed framework is tested on the survey data. The results are discussed in Section 7.1.4.

7.1.1 Detecting Derivative Rules

As laid out in the previous section the aim is to find rules that on the one hand display an interesting history of support or confidence and that on the other hand are fundamental in the sense that their history is not a derivative of related rules' histories. In a way, the approach is looking for the rules which are a root cause of a change pattern. All other rules can be discarded, which in turn can be seen as a form of pruning.

In order to find derivative rules the following two properties have to be defined:

1. What is meant by *related* rules?

2. What makes a history a *derivative* of other histories?

In order to explain the history of one rule with the history of another, the second rule must obviously be more general (less specific) than the first one. Otherwise, the second rule cannot make any assertions about the validity of the first one. By looking at related rules I therefore mean looking at less specific rules as defined in Section 2.1, i.e. r' is related to r iff $r \prec r'$.

For the definition of derivative history I consider history of measures like support or confidence of a rule. The following definition includes derivative measure histories of itemsets as a generalisation from rules. Thereby, the superset relation is used to define *related itemsets*: an itemset $\mathcal{Y}$ is related to an itemset $\mathcal{X}$ iff $\mathcal{X} \prec \mathcal{Y} := \mathcal{X} \supset \mathcal{Y}$. As before, $\mathcal{X}\mathcal{Y}$ is written for $\mathcal{X} \cup \mathcal{Y}$.

Definition 1: Let s, $s_1, s_2 \ldots s_p$ be rules or itemsets with $s \prec s_i$ for all i. In case of rules, let the antecedent itemsets of the s_i be pairwise disjoint, in case of itemsets let the s_i be pairwise disjoint. The history $H_m(s)$ regarding a measure m is called *derivative* iff a function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ exists with

$$m_s(T) = f(m(s_1, T), m(s_2, T) \ldots m(s_p, T)) \tag{7.1}$$

for all $T \in \hat{T}$. ■

The main idea behind the definition is that the history of a rule (itemset) is derivative, if it can be constructed as a mapping of the histories of less specific rules (itemsets). To compute the value $m(s, T)$ the values $m(s_i, T)$ are thereby considered. The function should be 'simple' in the sense that it is composed of basic operations, like quotients and products. More complex functions would have the disadvantage of being difficult to interpret. It should also be noted that a trivial function can always be defined such that each history is derivative. However, such functions will not be considered. For simplicity, I call a rule or itemset *derivative with respect to a measure m* iff its history of m is derivative.

In the following three criteria for detecting derivative histories are introduced. The first two criteria deal with itemsets and can therefore be directly applied to the support and antecedent support of rules, as well. The last criterion is related to histories of rule confidences.

The first criterion checks if the support of an itemset can be explained with the support of exactly one less specific itemset.

Criterion 1: The term $\operatorname{supp}(\mathcal{XY}, T)/\operatorname{supp}(\mathcal{Y}, T)$ is constant over $T \in \hat{T}$ given disjoint itemsets $\mathcal{X}$ and $\mathcal{Y}$. ■

Rewriting the criterion as

$$c = \frac{\operatorname{supp}(\mathcal{XY}, T)}{\operatorname{supp}(\mathcal{Y}, T)} = \frac{P(\mathcal{XY} \mid T)}{P(\mathcal{Y} \mid T)} = P(\mathcal{X} \mid \mathcal{Y}T)$$

with a constant c reveals its meaning. The probability of $\mathcal{X}$ is constant over time given $\mathcal{Y}$, so the fraction of transactions containing $\mathcal{X}$ in addition to $\mathcal{Y}$ constantly grows in the same proportion as $\mathcal{Y}$. This definition is also closely related to confidence, and states that the confidence of the rule $\mathcal{Y} \rightarrow \mathcal{X}$ should not change. For this reason the influence of $\mathcal{X}$ in the itemset $\mathcal{XY}$ on the support history is not important. Due to

$$\operatorname{supp}(\mathcal{XY}, T) = f(\operatorname{supp}(\mathcal{Y}, T)) = c \cdot \operatorname{supp}(\mathcal{Y}, T) \tag{7.2}$$
$$\text{with constant } c = \frac{\operatorname{supp}(\mathcal{XY}, T')}{\operatorname{supp}(\mathcal{Y}, T')} \text{ for any } T' \in \hat{T}$$

$\mathcal{XY}$ is obviously a derivative of $\mathcal{Y}$ with respect to support history as defined in Definition 1.

Figures 7.1 and 7.2 show an example of a derivative support history of a rule taken from the survey data. For reasons of data protection, the underlying rule cannot be revealed. The reader is referred to the example given in the introduction to this section, instead, to illustrate the meaning of Criterion 1. Figure 7.1 shows the support histories of the less specific rule at the top and the more specific rule below over 20 time periods. The shape of the two curves is obviously very similar and it turns out that the history of the more specific rule can be approximately reconstructed using the less specific one based on (7.2). As shown in Figure 7.2, the reconstruction is not exact due to noise. As a result, a statistical test is employed in Section 7.1.3 to test the validity of the criteria.

In contrast to the criterion above, the following is based on the idea of explaining the support of an itemset with the support values of two subsets.

Criterion 2: The term

$$\frac{\operatorname{supp}(\mathcal{XY}, T)}{\operatorname{supp}(\mathcal{X}, T)\operatorname{supp}(\mathcal{Y}, T)}$$

is constant over $T \in \hat{T}$ given disjoint itemsets $\mathcal{X}$ and $\mathcal{Y}$. ■

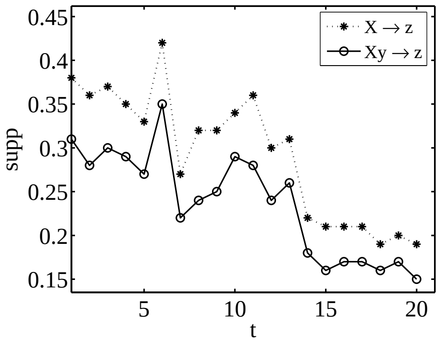


Figure 7.1: Histories of the rule $\mathcal{X} \to z$ and its derivative rule $\mathcal{X}y \to z$

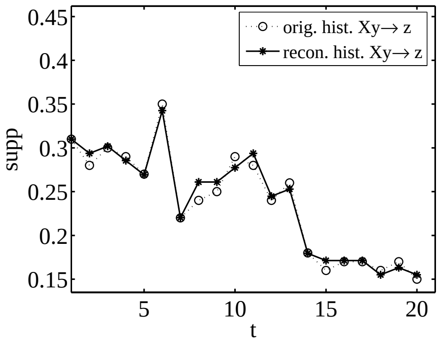


Figure 7.2: Reconstructed history of $\mathcal{X}y \to z$ using the history of $\mathcal{X} \to z$

$\operatorname{supp}(\mathcal{XY}, T)$ measures the probability of the itemset $\mathcal{XY}$ in period T which is $P(\mathcal{XY} \mid T)$. The term

$$\frac{\operatorname{supp}(\mathcal{XY}, T)}{\operatorname{supp}(\mathcal{X}, T), \operatorname{supp}(\mathcal{Y}, T)} = \frac{P(\mathcal{XY} \mid t)}{P(\mathcal{X} \mid T) P(\mathcal{Y} \mid T)}$$

is quite extensively used in data mining to measure the degree of dependence of $\mathcal{X}$ and $\mathcal{Y}$ at time T. Particularly in association rule mining this measure is also known as *interest factor* (Silverstein et al., 1998) and *lift* (Webb, 2000). The criterion therefore expresses that the degree of dependence between both itemsets is constant over time.

The support history of $\mathcal{XY}$ can then be constructed using

$$\begin{aligned} \operatorname{supp}(\mathcal{XY}, T) &= f(\operatorname{supp}(\mathcal{X}, T) \operatorname{supp}(\mathcal{Y}, T)) \qquad (7.3) \\ &= c \cdot \operatorname{supp}(\mathcal{X}, T) \operatorname{supp}(\mathcal{Y}, T) \\ \text{with constant } c &= \frac{\operatorname{supp}(\mathcal{XY}, T')}{\operatorname{supp}(\mathcal{X}, T') \operatorname{supp}(\mathcal{Y}, T')} \text{ for any } T' \in \hat{T} \end{aligned}$$

that is the individual support values of the less specific itemsets are used corrected with the constant degree of dependency on another. According to Definition 1 the support history of $\mathcal{XY}$ is therefore derivative.

Overall, an itemset is considered derivative with respect to support if more general itemsets can be found such that at least one of the above criteria holds.

Finally, the last criterion deals with derivative confidence histories of rules.

Criterion 3: The term

$$\frac{\operatorname{conf}(r, T)}{\operatorname{conf}(r', T)}$$

is constant over $T \in \hat{T}$ given two rules r and r' with $r \prec r'$. ■

Assuming the rules $r = \mathcal{XY} \Rightarrow z$ and $r' = \mathcal{Y} \Rightarrow z$ with disjoint itemsets $\mathcal{X}$ and $\mathcal{Y}$, the criterion translates to

$$\frac{P(z \mid \mathcal{XY}T)}{P(z \mid \mathcal{Y}T)}$$

being constant over time. This basically means that the contribution of $\mathcal{X}$ in addition to $\mathcal{Y}$ to predict z relative to the predictive power of $\mathcal{Y}$ remains stable over time and can therefore be neglected. The confidence history of r is derivative because of

$$\operatorname{conf}(r,T) = f(\operatorname{conf}(r',T)) = c \cdot \operatorname{conf}(r',T) \tag{7.4}$$
$$\text{with constant } c = \frac{\operatorname{conf}(r,T')}{\operatorname{conf}(r',T')} \text{ for any } T' \in \hat{T}$$

Remark 1: One recent pruning approach for association rules is based on so-called *informative rule sets* (Li et al., 2004). A rule $r \ : \ \mathcal{XY} \Rightarrow z$ is not in the informative rule set iff there exists a rule $r' \ : \ \mathcal{Y} \Rightarrow z$, such that $\operatorname{conf}(r) - \operatorname{conf}(r') \leq 0$. This is equivalent to $\operatorname{conf}(r)/\operatorname{conf}(r') = c$ for any $c \leq 1$ which matches (7.4) if c is constant over all periods. Criterion 3 for detecting derivative rules is therefore consistent with the criterion for non-informative rules. □

Remark 2: Another approach in the field of association rule mining to restrict the number of discovered rules is to generate the *non-redundant rule set* based on closed frequent itemsets (Zaki, 2004). A rule with only one item in its consequent is thereby considered redundant iff a more general rule with the same support and confidence exists. The two conditions of this criterion obviously match (7.2) and (7.4), respectively, for $c = 1$. Therefore if a rule would be considered derivative only if both Criterion 1 and Criterion 3 are satisfied with respect to the same more general rule it can be stated that the intersection of all periods' non-redundant rule set is a subset of the obtained non-derivative rule set. □

7.1.2 Implementation Issues

The last section proposed a set of criteria to detect derivative rules which again rely on a search over the set of related itemsets or rules, respectively. Generally, this search is exhaustive and thus a potentially exponential number of comparisons is required (e.g. for every frequent itemset all subsets have to be enumerated in the worst case). The approach's apparent complexity may evoke questions about its feasibility. Actually, the approach works within reasonable time, considering that rule change mining for business data is typically carried out infrequently. Nonetheless, in some domains complexity may be an issue. In this case the number of comparisons can be reduced by the following simplification adopted from (Liu et al., 2001a): instead of all related itemsets (rules) only *closely related* are considered. Criterion-wise this means that for an itemset $\mathcal{X}$ or rule $\mathcal{X} \rightarrow z$ and any $y \in \mathcal{X}$:

- Criterion 1: the itemset $\mathcal{X} \setminus \{y\}$ is considered
- Criterion 2: the itemsets $\mathcal{X} \setminus \{y\}$ and $\{y\}$ are considered
- Criterion 3: the rule $\mathcal{X} \setminus \{y\} \rightarrow z$ is considered

Independent of which of the two approaches – the general or the above simplified – is actually used, due to performance reasons it is always assumed that derivativeness is transitive. For example, rule r_1 may explain the support history of rule r_2, and rule r_2 may explain the support history of rule r_3. In this case, r_3 is treated as a derivative of r_1. In fact, the criteria for derivative rules will due to noise rarely be exactly satisfied. The transitivity assumption may therefore, due to the Poincaré paradox, not hold. On the other hand, however, a generally conducted further test between any r_1 and r_3 would drastically reduce the approach's overall performance.

7.1.3 Testing the Criteria

To check if the history of a rule r or an itemset $\mathcal{XY}$ is derivative with respect to support or confidence, it needs to be tested if the criteria introduced in Section 7.1.1 hold. The choice of a suitable testing procedure was influenced by two requirements:

- The procedure should not significantly depend on difficult to understand and to adjust parameters.

- The procedure should be reliable.

While the latter requirement is straightforward, the first needs some explanation: due to the vast number of histories a user is virtually unable to validate the resulting non-derivative rule set. Particularly, it may be difficult for him to qualitatively judge the strictness of the test for certain parameters, because the boundary between derivativeness and non-derivativeness is due to noise rather fuzzy. For this reason an ad hoc parameter choice – the likely implication of incomprehensible parameters – can lead to counterintuitive results.

The proposed testing procedure will not use the definitions of Criteria 1–3 directly, but an equivalent representation, which finally leads to a reliable and comprehensible test method based on regression analysis. To underline the choice it will be contrasted to some alternative approaches at the end of this section.

Let

$$\Delta_i \operatorname{supp}(\mathcal{X}) := \frac{\operatorname{supp}(\mathcal{X}, T_i)}{\operatorname{supp}(\mathcal{X}, T_{i-1})}$$

be the relative change in support for itemset $\mathcal{X}$ and let

$$\Delta_i \operatorname{conf}(r) := \frac{\operatorname{conf}(r, T_i)}{\operatorname{conf}(r, T_{i-1})}$$

be the relative change in confidence for the rule r, both between two periods T_{i-1} and T_i. As it can be easily verified the respective Criterion 1,2 or 3 holds, iff for any $T_i \in \hat{T} \backslash \{T_1\}$:

$$\begin{aligned} &\textbf{Criterion 1:} \quad \Delta_i \operatorname{supp}(\mathcal{XY}) = \Delta_i \operatorname{supp}(\mathcal{Y}) && (7.5)\\ &\textbf{Criterion 2:} \quad \Delta_i \operatorname{supp}(\mathcal{XY}) = \Delta_i \operatorname{supp}(\mathcal{X}) \Delta_i \operatorname{supp}(\mathcal{Y}) && (7.6)\\ &\textbf{Criterion 3:} \quad \Delta_i \operatorname{conf}(r) = \Delta_i \operatorname{conf}(r') && (7.7) \end{aligned}$$

This means that if Criterion 1 or 3 holds for an itemset $\mathcal{XY}$, or a rule r, then the relative changes in its history are equal to the temporally related relative changes in the history of a more general itemset $\mathcal{X}$, or rule r'. If Criterion 2 holds, then the relative changes in the history of $\mathcal{XY}$ are equal to the product of the corresponding relative changes in the histories of $\mathcal{X}$ and $\mathcal{Y}$.

Obviously, (7.5)-(7.7) are following the same general scheme $y_i = x_i;\ i = 2, \ldots, n$, whereas the quantities y_i and x_i stand for the left and accordingly right hand side of the equations.

It is convenient for the following discussion to imagine x_i and y_i in a plot, whereby y_i is – as implied by Definition 1 – the dependent quantity. If $y_i = x_i$ holds, then all points in the plot should be on a straight line with slope 1 and intercept 0. As the already suggested earlier in the section, in practice this equality will rarely hold due to noise. In fact, the underlying relationship will be $y_i = x_i + \epsilon$ where ϵ is a random error with zero mean and unknown, but low variance.

Under the assumption that the dependence of y_i on x_i can be generally described by $y_i = ax_i + b + \epsilon$, a regression line $y = \hat{a}x + \hat{b}$ is fitted. The parameters $\hat{a}$ and $\hat{b}$ are estimates for a and b and obtained, for example, by minimising the sample variance of the error $e_i := y_i - (\hat{a}x_i + \hat{b})$ between the actual vertical point y_i and the fitted value $\hat{a}x_i + \hat{b}$, i.e.

$$s^2(e) = \frac{1}{n-2}\sum_{i=2}^{n}(y_i - \hat{a}x_i - \hat{b})^2$$

It is then tested whether x_i is statistically equal to y_i by carrying out the following two steps:

1. Based on the estimates $\hat{a}$ and $\hat{b}$ the hypothesis is tested that the true parameters of the model are $a = 1$ and $b = 0$.

2. Additionally, it is tested whether the variance of ϵ is small, i.e. whether the points (x_i, y_i) are sufficiently close to the regression line.

Step 1: To test hypotheses about the true, but unknown parameters a and b, two statistical tests are employed which are, for example, described in (Montgomery and Runger, 2002). Both tests transform the estimates $\hat{a}$ and $\hat{b}$ such that they are approximately t-distributed. The transformed values are then used as the test statistic. In the following I denote the standard errors of a and b as

$$s(a) := s(e)\sqrt{\frac{1}{\sum_{i=2}^{n}(x_i - \bar{x})^2}}$$

and

$$s(b) := s(e)\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=2}^{n}(x_i - \bar{x})^2}}$$

respectively.

The null hypothesis $H_0 : \ a = 1$ is tested against the alternative $H_1 : \ a \neq 1$. If H_0 holds the test statistic $|(\hat{a} - 1)/s(a)|$ is t-distributed with $(n-3)$ degrees of freedom. Thus H_0 is rejected at significance level α if

$$\left|\frac{\hat{a}-1}{s(a)}\right| > t_{n-3;1-\alpha/2}$$

where $t_{n-3;1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the t-distribution.

The null hypothesis $H_0 : \ b = 0$ is tested against the alternative $H_1 : \ b \neq 0$. If H_0 holds the test statistic $|\hat{b}/s(b)|$ is t-distributed with $(n-3)$ degrees of freedom. Thus H_0 is rejected at significance level α if

$$\left|\frac{\hat{b}}{s(b)}\right| > t_{n-3;1-\alpha/2}$$

where $t_{n-3;1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the t-distribution.

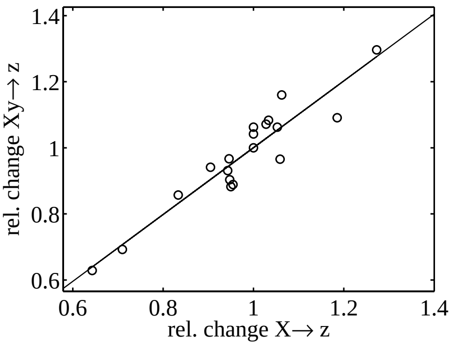


Figure 7.3: Scatter plot of the relative changes of the histories shown in Figure 7.1. The regression line is $y = 1.0107x - 0.0103$ and the correlation coefficient $r = 0.97$

Step 2: To test how close the points (x_i, y_i) are scattered around the regression line Pearson's correlation coefficient $r(x, y)$ is used, which, in contrast to the regression error's sample variance $s^2(e)$, has the advantage of being dimensionless and thus more comprehensible. However, $r(x, y)$ is related to $s^2(e)$ by

$$r(x, y) = \frac{s(e)}{s(y)}$$

$$\text{with } s^2(y) = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ as the sample variance of } y_i.$$

Because $r(x, y) \to 1$ if $s(e) \to 0$ and $r(x, y) \in [0, 1]$, a lower boundary $\tilde{r} < 1$ is used for testing.

Figure 7.3 illustrates the testing procedure. It shows the scatter plot of the relative changes of the support histories shown in Figure 7.1. The regression line is $y = 1.0107x - 0.0103$ and the correlation coefficient $r \approx 0.97$. The above test procedure using $\alpha = 0.05$ and $\tilde{r} = 0.95$ shows that the more specific rule is indeed derivative with respect to the history of the less specific one.

A similar approach to test the criteria can be directly derived from (7.2), (7.3) and (7.4). It utilises that each derivative history is a linear function with zero intercept of the history of one or two more general rules or itemsets, respectively. Therefore it has only to be tested if the intercept is zero and the correlation coefficient considerably small, whereas the slope can have an arbitrary, but positive value. Experiments, however, indicate that such an approach sometimes counterintuitively marks histories as derivative. For example, it has been observed that some histories were classified derivative, although significant changes and other remarkable episodes are not adequately modeled by the history of the more general rule or itemset. The reason is that the approach does not account for the inherent order of histories. In contrast, the technique proposed in the first part of this section does account for the order by incorporating the relative changes between values and so it is much stricter.

Finally, it should be noted that the decision for derivativeness of a history is bound to the null hypothesis. The drawbacks connected with it have already been laid out in Section 3.3.5 and it has also been stressed that such a configuration is unavoidable if statistical equality needs to be shown. The implication, however, is basically that the detection of derivative histories is optimistically biased. An alternative to generally circumvent hypothesis testing would be the introduction of user-defined thresholds for slope and intercept. Nonetheless, their reasonable definition may be a tedious and complicated task for a user.

7.1.4 Experimental Evaluation

In Section 7.1 the concept of derivative rule histories has been motivated in particular by the immense number of histories to be examined by rule change mining approaches and by the consequently vast number of detected change patterns. The proposed detection method can either be applied before the actual change pattern detection or between change pattern detection and interestingness assessment (cp. Section 4.2.2). Therefore the first question to be answered experimentally is how many rule histories are derivative.

Algorithms for association rule mining typically use minimum support and confidence thresholds as parameters to control the number of discovered rules. Obviously, they also influence the number of more general rules discovered for each rule. Since the proposed approach utilises the generalisation relationship between rules, the second question to be answered is how it scales for different parameter settings of the mining algorithm.

Finally, the third question is based on Remark 2 and regards the extent to which the combined application of Criterion 1 and 3 is superior to the straightforward temporal extension of the non-redundant rule set approach by (Zaki, 2004) in terms of the number of discarded rules.

For all experiments a significance level of $\alpha = 0.05$ and a correlation threshold of 0.95 were used. The experiments where applied to all histories (cp. Section 5.3) as well as only to histories which contain a change pattern (cp. Section 6.4). Because Criterion 1 and Criterion 2 are defined for itemsets, the notion of a 'relatedness' needs to be adapted to rules. For support histories a rule is regarded as 'related' if its constituting itemset is a subset of those of the rule of interest. For antecedent support histories a rule is regarded as 'related' if its antecedent itemset is a subset of those of the rule of interest and both rules have the same consequent.

To answer the first question only histories which exhibit change patterns were tested if they are a derivative of another history. The first row of Table 7.1 shows the obtained results for trends separately for support, antecedent support and confidence histories. As it can be seen are between 40.7% (for confidence) and 66.3% (for support) of the histories derivative. The second row shows that these numbers are considerably smaller for stable histories; ranging from 19.2% (for antecedent support) to 39.6% (for support).

To answer the second question the mining phase described in Section 5.3 was repeated with two further parameter settings for the *apriori* algorithm. Overall, the $(\text{supp}_{min}, \text{conf}_{min})$ combinations $(0.1, 0.6)$, $(0.05, 0.4)$ and $(0.05, 0.2)$ were used, where the latter is the one used throughout this thesis. In contrast to the above, derivativeness was tested for all histories obtained in each setting. The results for support and antecedent support histories are shown in Table 7.2 and Table 7.3, respectively. Each row in the tables refers to a different parameter setting. The third column shows the number of histories to analyse. The fourth and fifth column show the number of histories classified as derivative by Criterion 1 and Criterion 2, respectively. The sixth and seventh column show the overall number of derivative histories, absolute and relative. As it can be seen, the figures for derivative histories are relatively insensitive to changes in the thresholds for support and confidence. The same holds for confidence histories for which the results are presented in Table 7.4.

To answer the third question the approach suggested in Remark 2 was applied to all histories obtained with the rule mining parameters $(0.05, 0.2)$. This means, a rule is only discarded if both Criterion 1 and Criterion 3 are satisfied with respect to the same more general rule. Of 77401 rules this method discards 24426, which is approximately 31.5%. To contrast this number with the size of the non-redundant rule set proposed by (Zaki, 2004) all rules were discarded for which a more general rule with identical support and confidence history exists. Of 77401 rules this method discards 3567, which is approximately 4.6%. Obviously, the pruning technique proposed in this thesis leads to a drastic increase in the number of pruned rules compared to the straightforward temporal extension of the approach by (Zaki, 2004).

	Support		Antecedent Support		Confidence	
Pattern	#histories	deriv.(%)	#histories	deriv.(%)	#histories	deriv.(%)
trend	42307	66.3	44091	52.5	30387	40.7
stable	2019	39.6	1642	19.2	21753	26.7

Table 7.1: Fraction of derivative histories among all histories which have a trend or are stable.

$\mathrm{supp}_{\mathrm{min}}$	$\mathrm{conf}_{\mathrm{min}}$	#histories	|C1|	|C2|	$\|C1 \cup C2\|$	ratio(%)
0.1	0.6	13586	9286	5249	10035	73.9
0.05	0.4	49282	32975	22559	36253	73.6
0.05	0.2	77401	52740	37638	58144	75.1

Table 7.2: Number of derivative support histories detected by each criterion (C1, C2) with different parameter settings of the rule miner.

$\mathrm{supp}_{\mathrm{min}}$	$\mathrm{conf}_{\mathrm{min}}$	#histories	|C1|	|C2|	$\|C1 \cup C2\|$	ratio(%)
0.1	0.6	13586	7136	5723	9283	68.3
0.05	0.4	49282	23737	22139	31578	64.1
0.05	0.2	77401	37645	39833	51735	66.8

Table 7.3: Number of derivative antecedent support histories detected by each criterion (C1, C2) with different parameter settings of the rule miner.

$\mathrm{supp}_{\mathrm{min}}$	$\mathrm{conf}_{\mathrm{min}}$	#histories	|C3|	ratio(%)
0.1	0.6	13586	6108	44.9
0.05	0.4	49282	20536	41.6
0.05	0.2	77401	34142	44.1

Table 7.4: Number of derivative confidence histories detected by criterion 3 (C3) with different parameter settings of the rule miner.

7.2 Interestingness Assessment

As documented by the experimental results in Section 7.1.4, the proposed pruning approach significantly reduces the number of histories with change patterns. The experiments revealed furthermore that the number of non-derivative change patterns is still rather large and so the search for the most crucial change patterns a rather tedious task for a user. Earlier it has also been noted that derivative histories in general are commonly less interesting for a user. Nonetheless, this does not mean that all non-derivative histories which exhibit a change pattern are equally interesting. In fact the most interesting ones are still hidden among many irrelevant.

To assess the interestingness of detected trends and stabilities it has to be considered that each history is linked to a rule which itself is prior to rule change mining differently interesting to a user, yet the detection of a certain change pattern may significantly influence this prior interest. In this sense a rule can have different degrees of interestingness, each related to another history. However, in Section 2.3 it has been discussed that no broadly accepted and reliable way of measuring a rule's interestingness has been developed up to now, yet there are two commonly accepted concepts: interestingness measures for association rules can be divided into two classes, subjective and objective, depending whether or not the user has to specify its background knowledge about the problem domain (cp. Section 2.3.2). It was also discussed that the choice of the actually used type depends not only on the required and practically feasible degree of interaction between the user and the data mining system, but also on the sort of data mining task.

In the following two sections it will be discussed whether and how the concepts of subjective and objective interestingness measures can be transferred to both stable and trend histories. As a result several objective interestingness measures will be proposed.

7.2.1 Assessment of Stabilities

As already indicated above, every statement about the interestingness of a history is also a statement about the interestingness of its related rule. With this in mind two things should be considered: first, association rule mining typically assumes that the domain under consideration is stable over time. Second, measures like support and confidence, for which histories are analysed are itself objective interestingness measures for rules.

Taking all this into account, a stable history is in some way consistent with the above stated assumption of association rule mining. It is summarised by the mean of its values, which in turn can then be treated as an objective interestingness measure. Here the variance of the history can be neglected, since it is constrained by the stability detection method proposed in Section 6.3.

A stable history itself has very few significant properties. Besides the mean and variance, the most basic one is a Boolean proposition about the existence of stability. The latter in turn can be easily expressed by a user. Therefore it is thinkable to extend the syntax- and logic based subjective interestingness approaches mentioned in Section 2.3.3 by a Boolean variable expressing a stability expectation.

7.2.2 Assessment of Trends

If the concept of subjective interestingness measures is transferred to histories which exhibit a trend, a user would not only be required to specify association rules, but also his expectations about the development of their histories. As discussed in Section 2.3.4, this task is rather complicated even for rules itself, such that it is questionable if a user has the expertise and the time to specify trend descriptions in addition: most trends have complex shapes which are, in contrast to rules, a non-symbolic and thus difficult expressible type of information. A simplistic alternative is to solely specify logical assertions about the expected existence of an upward or downward trend. However, like specifying shapes is overly complex, this method is overly simple because it does not account for almost any property of a trend's shape.

Objective interestingness measures for trends, in contrast, would not have this drawback of complexity. They rank trends by numerical properties or, for example, by their relation to the histories of more general rules. However, due to their numeric nature properties of trends are, compared to those of association rules, larger in number and more sophisticated. Therefore it cannot be expected to find a universal objective measure that fits the needs of all users and mining tasks.

To develop objective interestingness measures for trends, it is essential to account for their psychological background. As laid out in Section 2.3.1 should objective measures be interpreted as tools which support a user's bottom-up control of attention. This means that only trends which exhibit a salient feature will gain the user's focus. Unarguably it is virtually impossible to state whether a feature is salient without providing any reference point for comparison. As such I chose the assumptions a user naively consults in lack of knowledge about the changes in rules' histories. From a psychological perspective they can be seen as the anchors against which histories with a trend are assessed: a trend becomes more interesting with increasing inconsistency between its features and the user's naive assumptions. It is therefore crucial to identify such assumptions, define the corresponding properties and finally assess their inconsistency in order to derive objective interestingness measures for trends. This may sound confusing: *subjective* assumptions are incorporated into *objective* measures. However, the difference to subjective interestingness measures is basically that the assumptions are far fewer in number and much more general in the sense that they are not bound to the mining task and just loosely connected to the problem domain. In the following three assumptions will be presented:

- **Stability:** Unless other information is provided, the user assumes that histories are stable over time. This assumption does not mean that he expects no trends at all, but expresses his naive expectations in lack of precise knowledge about a trend. It should be noted that this is consistent with conventional association rule mining which implicitly assumes that the associations hidden in the data are stable over time.

- **Non-rapid Change:** Since the user shapes its business, he will be aware that the domain under consideration changes over time. However, he will assume that any change is continuous in its direction and moderate in its value. For example, if a business starts a new campaign, it will probably assume that the desired effect evolves moderately, for example, because not all people will see a commercial immediately. On the other hand, a rapid change in this context attracts more attention, because it may hint at an overwhelming success or an undesired side effect.

- **Homogeneous Change:** If the support of a rule (itemset) changes over time, it is assumed that the rate and direction of changes in the support of all its specialisations are the same. This basically means that the observed change in the rule (itemset) does not depend on further items. For example, a user may know that the fraction of satisfied customers increases. The homogeneous change assumptions states that the observed change in satisfaction affects all customers and not only selected subpopulations, e.g. females over forty.

If, on the other hand, the confidence of a rule changes over time, it is assumed that the confidence of all more specialised rules changes at the same rate. For example, the fraction of satisfied males among all male customers may increase. According to the homogeneous change assumption a user would conclude that among all married male customers the fraction of satisfied ones increases at the same rate.

This list is not complete. In fact, it enumerates the assumptions considered meaningful to analyse the survey data used in this thesis. The assumptions are inspired by the way how surveys are commonly analysed and thus kind of mimic the routine of an analyst. However, due to their commonality they are applicable to many other domains. It should also be noted that the assumptions are neither supposed to be consistent nor mutually exclusive, because whether and how they are employed depends on the given context.

In the following several measures are introduced which assess the consistency of histories with the above assumptions. They can be applied to trends in common but also to upward and downward trends separately. The measures all have in common that a trend's interestingness increases with the measure's value. Furthermore, it will be assumed that the interest in the rule itself is primarily determined by the interestingness of its change over time – a claim which is supported by many researchers like (Chakrabarti et al., 1998), (Liu et al., 2001b) and (Dong et al., 2003). In this sense a rule can have different degrees of interestingness, each related to one of its histories.

The measures have been extensively used to analyse the survey data by applying them to the histories for which a trend has been detected (cp. Section 6.4). They turned out to be very effective and greatly helped to identify several previously unknown and surprising trends. However, due to reasons of data protection it is not possible to reveal any discovered knowledge. Instead only examples of top and bottom rank rules are provided for each measure without naming the related rule.

Clarity

This metric assesses the clarity of a detected trend and thus the certainty that it indeed exists. Maximum clarity is reached for an upward trend if each value is greater than its predecessor and for an downward if each value is smaller than its predecessor. Obviously, such a measure targets the stability assumption which becomes more uncertain with increasing certainty of a trend.

The test statistics for the Mann-Kendall and the Cox-Stuart test, introduced in Section 6.2.1 and Section 6.2.2, respectively, both provide a suitable basis for assessing a trend's clarity. However, the range of the Cox-Stuart test statistic's possible values is significantly smaller than those of the Mann-Kendall test. Particularly for short histories this leads to a very coarse ranking, with the same value assigned to many histories. For example, the histories obtained from the survey data have a length of 20 such that the Cox-Stuart test statistic can take on values from the set $\{0, \ldots, 10\}$. If additionally the test's threshold value is taken into account the size of this set reduces further,

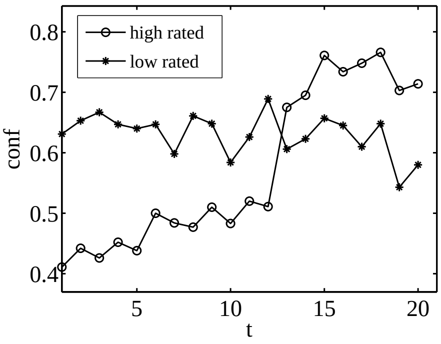


Figure 7.4: Clarity: trends ranked high and low by measure (7.8)

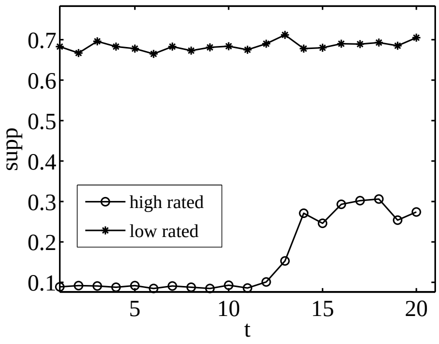


Figure 7.5: Pronouncedness: trends ranked high and low by measure (7.14)

for example, to $\{8, 9, 10\}$ for a significance level of $\alpha = 0.05$. The Mann-Kendall test statistic, on the other hand, does not have this disadvantage and so it will be incorporated into a clarity measure.

Let C denote the Mann-Kendall test statistic for the history H, the clarity ψ_{clarity} of a trend is defined as

$$\psi_{\text{clarity}}(H) := |C| \tag{7.8}$$

Figure 7.4 shows two histories from the survey data of which one is rated highly interesting ($\psi_{\text{clarity}} = 150$), whereas the other has a very low degree of interestingness ($\psi_{\text{clarity}} = 50$).

Pronouncedness

A measure which targets both the stability and the non-rapid change assumption is derivable by assessing the pronouncedness of a trend. Pronouncedness, in this context, means the deviation of a trend from stability which in turn can be described by the mean line. In the following I will provide two measures of which one is heuristic but applicable to every history, whereas the other is based on information theory but restricted to support histories.

Let $H = (v_1, \ldots, v_n)$ be a support, antecedent support or confidence history. Because these are relative measures, all histories should be scaled such that on the one hand they have the same mean level but on the other hand all relative changes are preserved. The scaled sequence will be denoted by $(v'_1, \ldots, v'_n)$, where v'_i can be derived by

$$v'_i = \frac{v_i}{\sum_{i=1}^{n} v_i} \tag{7.9}$$

It is easy to verify that this scaling yields sequences with the above properties. The mean line of the transformed sequence is described pointwise by the sequence $(1/n, \ldots, 1/n)$. One method to assess the pronouncedness of a trend in the history H is to calculate the distance between the transformed sequence above and the mean line sequence, for example:

$$\psi_{\text{pron1}}(H) := \sum_{i=1}^{n} \left| v'_i - \frac{1}{n} \right| \tag{7.10}$$

Any other norm to measure the distance could be used as well, but experiments with different p-norms showed no significant change in the final ranking of trends.

If the sequence H is a support (or antecedent support) history the probabilistic meaning of the support can be utilised to derive a pronouncedness measure with underpinnings in information theory. Let $\mathcal{X}$ be an itemset and $\operatorname{supp}(\mathcal{X}, T_i)$ its support in each period $T_i, i = 1, \ldots, n$. As stated in Section 3.1 is the support an estimate for the probability that a transaction supports $\mathcal{X}$ in period T_i, i.e.

$$\operatorname{supp}(\mathcal{X},\ T_i) \approx P(\mathcal{X}|\ T_i) \tag{7.11}$$

Using Bayes' rule this is rewritten as

$$P(T_i|\ \mathcal{X}) = \frac{P(\mathcal{X}|\ T_i)P(T_i)}{P(\mathcal{X})} \tag{7.12}$$

The left side can be interpreted as the probability that a randomly drawn transaction, of which it is known that it supports $\mathcal{X}$, has been generated in period T_i. In case that the history H is perfectly stable $P(\mathcal{X}|\ T_i) = P(\mathcal{X})$ obviously holds and consequently is

$$P(T_i|\ \mathcal{X}) = P(T_i) \tag{7.13}$$

This basically means that the information that a transaction supports $\mathcal{X}$ provides no additional knowledge about the period to which the transaction belongs. Given an arbitrary support history with a trend for an itemset $\mathcal{X}$, (7.13) can be used to derive a pronouncedness measure by comparing the a posteriori distribution $P(T|\ \mathcal{X})$ with the a priori distribution $P(T)$.

A well-known and broadly used measure to compare two distribution is *relative entropy*, also called the *Kullback-Leibler distance* (Kullback and Leibler, 1951). The relative entropy can be considered as a sort of a distance between two probability distributions, though it is not a true metric because it is not symmetric. Using the relative entropy, the following pronouncedness measure is defined:

$$\psi_{\text{pron2}}(H) := \sum_{i=1}^{n} P(T_i|\ \mathcal{X}) \log_2 \frac{P(T_i|\ \mathcal{X})}{P(T_i)} \tag{7.14}$$

According to the information theoretic interpretation of the relative entropy it measures the average number of additional bits necessary to encode the period T if the coding is based on the stability assumption, compared to a coding based on the true distribution $P(T|\ \mathcal{X})$ induced by the trend (Cover and Thomas, 1991). Only if (7.13) holds the measure takes on its minimal value, i.e. $\psi_{\text{pron2}} = 0$.

Figure 7.5 shows two support histories derived from the survey data which both exhibit a trend. The ψ_{pron2} metric has been used to assess them together with all other support histories with a trend. As it can be seen, the top ranked rule differs significantly from the mean line in contrast to the lower rated rule.

Another issue that has been analysed is to which extent rankings produced by (7.10) and (7.14) differ. A ranking of all trend histories from the survey data where obtained for each measure. *Spearman's rank correlation coefficient* between both rankings is approximately 0.96 and thus they are rather similar. In terms of the obtained ranking this indicates that (7.10) is a very good substitute for (7.14).

Dynamic

The dynamic of a trend embodies the rate of its incline or decline, respectively. Therefore it targets the assumption of non-rapid change. As opposed to pronouncedness, however, it measures the change rate only for the most recent values of a history, because a user is likely to be more interested in how its business currently evolves, than how it performed in the past.

Let $H = (v_1, \ldots, v_n)$ be a history. Since support, antecedent support and confidence are relative measures the history is scaled using the same procedure as for the pronouncedness heuristic. As before the scaled sequence is $(v'_1, \ldots, v'_n)$.

To assess the dynamic of a trend only the last n' values are considered. A linear regression line is fitted to them, where the time is the independent variable. Let m denote the slope of the obtained regression line; the interestingness measure is then defined as:

$$\psi_{\text{dyn}}(H) := |m| \tag{7.15}$$

Alternatively, the slope of the secant passing through the points v'_n and v'_{n-h} can be used, though experiments indicate that this approach is less robust to noise.

Additional insight into the detected trends and a further discrimination of the most interesting ones can be gained by comparing the rankings by dynamic and pronouncedness. For example, a history which has a high pronouncedness but a low dynamic is likely to have rapidly in- or declined in the past but is now stabilising. In contrast, a rule with a low pronouncedness but a high dynamic may hint at a current rapid increase in the trend's change rate.

To test the effectiveness of the dynamic as an interestingness measure it has been applied among others to all confidence histories with a trend. $n' = 5$ was used, i.e. only the last 5 periods where taken into account. Figure 7.6 shows two trends to one of which a high interestingness degree has been assigned, while the other received a low interestingness degree. On the other hand, if only the first 15 periods are considered the histories the ranking would be the other way round. The trend with the low dynamic, however, is one of the top ranked trends with respect to its pronouncedness. As discussed before, this would indicate a user that a trend which showed a rather rapid change in the past now stabilises. The top rated rule, on the other hand, has in addition the interesting feature that its current dynamic shows, opposed to the global upward trend, a downward direction.

Homogeneity

A user will typically be interested in those subpopulations which change differently than the population to which they belong, rather than in the information that a population has subsets which change differently. Transferred to support and confidence histories this means that they should be compared to the histories of each more general rule in order to detect inconsistencies with the homogeneity assumption. Furthermore, it has to be noted that the homogeneity assumption can generally be utilised to assess any history, regardless of whether it contains a change pattern.

Let m be a measure, like confidence, support or antecedent support and r the rule which has to be assessed. It is assumed that $H_m(r)$ exhibits a change pattern, this means, it is either stable or has a trend. Furthermore, let $R'(r)$ be the set of all rules

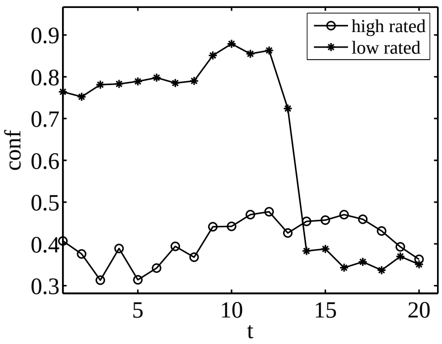


Figure 7.6: Dynamic: trends ranked high and low by measure (7.15) using $n' = 5$

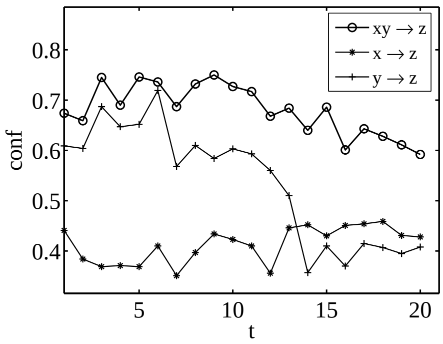


Figure 7.7: Homogeneity: a high ranked trend and the histories of the more general rules

$r' \succ r$ with exactly one item less in their antecedent whose histories $H_m(r')$ contain a change pattern. In case that $R'(r)$ is the empty set, the rule r will be ignored.

To assess the interestingness of a rule relative to the homogeneity assumption, a two step approach is proposed in the following. In the first step the deviation of $H_m(r)$ from $H_m(r')$ is calculated for each $r' \in R'(r)$ as defined below. In the second step the obtained deviations are aggregated into a single value. This aggregation allows a user to specify how the combination of obtained inhomogeneity values should be interpreted in terms of their interestingness. For example, the minimum operator can be interpreted as *all* histories within R' should be inconsistent with the assumption to obtain a high degree of interestingness. Likewise the maximum operator can be interpreted as *at least one* and the average as *some*. Such interpretations are in a way similar to the meanings of *ordered weighted average* (OWA) operators given in (Yager, 2000). There is also a more technical advantage connected to it: the length of a ranking solely based on the outcome of the first step would easily outnumber the assessed rules due to the potentially high cardinality of $R'(r)$. This, in turn, is undesirable with regard to the vast number of change patterns typically encountered.

Step 1. Calculating the deviation between two histories is an apparent extension of the methods for pronouncedness assessment, where the history of a more general rule replaces the mean line. As before two deviation measures are introduced of which one is a heuristic but applicable to every history, whereas the other is based on information theory but restricted to support histories. Both derivations are rather similar to those in the section on pronouncedness and therefore only briefly outlined in the following.

Let $H_m(r) = (v_1, \ldots, v_n)$ and $H_m(r') = (w_1, \ldots, w_n)$ be the histories of the rules r and r' with $r' \in R'(r)$. Both histories are scaled using (7.9) yielding sequences $(v'_1, \ldots, v'_n)$ and $(w'_1, \ldots, w'_n)$. The deviation between $H_m(r)$ and $H_m(r')$ is then defined as:

$$\phi_{\text{heuristic}}(H_m(r), H_m(r')) := \sum_{i=1}^{n} |v'_i - w'_i| \tag{7.16}$$

It should be noted that any other norm can be used as well, though experiments with different p-norms showed no significant change in the final ranking.

Under the restriction that both $H_m(r)$ and $H_m(r')$ are (antecedent) support histories a deviation measure with an information theoretic underpinning can be derived. Let $\mathcal{X}$ and $\mathcal{X}y$ be the corresponding itemsets of r' and r, respectively, with supports $\operatorname{supp}(\mathcal{X},\ T_i) \approx P(\mathcal{X}|\ T_i)$ and $\operatorname{supp}(\mathcal{X}y,\ T_i) \approx P(\mathcal{X}y|\ T_i)$ in each period T_i. Applying Bayes' rule yields the probabilities

$$P(T_i|\ \mathcal{X}) \quad = \quad \frac{P(\mathcal{X}|\ T_i)P(T_i)}{P(\mathcal{X})} \tag{7.17}$$

$$P(T_i|\ \mathcal{X}y) \quad = \quad \frac{P(\mathcal{X}y|\ T_i)P(T_i)}{P(\mathcal{X}y)} \tag{7.18}$$

If the homogeneity assumption holds it is

$$\frac{P(T_i|\ \mathcal{X}y)}{P(T_i|\ \mathcal{X})} = \frac{P(T_j|\ \mathcal{X}y)}{P(T_j|\ \mathcal{X})} \text{ for any } T_i, T_j \in \hat{T} \tag{7.19}$$

and it follows that

$$P(T_i|\ \mathcal{X}) = P(T_i|\ \mathcal{X}y) \text{ for any } T_i \in \hat{T} \tag{7.20}$$

This basically means that if the homogeneity assumption holds, the information that a transaction supports $\{y\}$ in addition to $\mathcal{X}$ gives no additional knowledge about the period to which the transaction belongs.

Proof:

$$\begin{aligned}
P(T_i|\mathcal{X}y) &= \frac{P(\mathcal{X}y|T_i)P(T_i)}{P(\mathcal{X}y)} \\
&= \frac{P(\mathcal{X}y|T_i)P(T_i)}{\sum_{j=1}^{n} P(\mathcal{X}y|T_j)P(T_j)} \text{ [law of total probability]} \\
&= \frac{P(\mathcal{X}|T_i)P(\mathcal{X}y|T_i)P(T_i)}{P(\mathcal{X}y|T_i)\sum_{j=1}^{n} P(\mathcal{X}|T_j)P(T_j)} \text{ [using (7.19)]} \\
&= \frac{P(\mathcal{X}|T_i)P(T_i)}{P(\mathcal{X})} \\
&= P(T_i|\mathcal{X})
\end{aligned}$$

■

Given support histories for the itemsets $\mathcal{X}y$ and $\mathcal{X}$, (7.20) can be used to derive a deviation measure by comparing the distribution $P(T|\mathcal{X}y)$ with $P(T|\mathcal{X})$. As in the section about the pronouncedness measure the relative entropy is utilised yielding the following deviation measure:

$$\phi_{\text{entropy}}(H_m(r), H_m(r')) := \sum_{i=1}^{n} P(T_i|\ \mathcal{X}y) \log_2 \frac{P(T_i|\ \mathcal{X}y)}{P(T_i|\ \mathcal{X})} \tag{7.21}$$

From an information theoretic perspective it measures the average number of additional bits necessary to encode the period T if the coding is based – according to the homogeneity assumption – on the distribution $P(T|\ \mathcal{X})$, compared to a coding based on the true distribution $P(T|\ \mathcal{X}y)$. Only if the homogeneity assumption holds the measure takes on its minimal value, i.e. $\phi_{\text{entropy}} = 0$.

Step 2. After for each $r' \in R'(r)$ the deviation of its history to the one of r has been calculated, the resulting set $\Phi(r) := \{\phi(H(r), H(r')) : r' \in R'\}$ is aggregated to the final homogeneity interestingness measure

$$\psi_{\text{homogeneity}}(H(r)) = \text{agg}(\Phi(r)) \tag{7.22}$$

The actually used aggregation operator agg is application dependent, yet should be well-behaved, i.e. $\min(\Phi) \leq agg(\Phi) \leq \max(\Phi)$. Some possible choices and their interpretation have already been discussed above. However, experiments with different aggregation operators, including min, max and avg, were indicating only a moderate influence of the actual choice on the final ranking result.

The above approach is related to the pruning method based on derivable histories proposed in Section 7.1: iff the history of a rule r is derivable, according to Criteria 1 or 3, respectively, with respect to the history of a more general rule r', the proposed deviation measures ϕ_{entropy} and $\phi_{\text{heuristic}}$ are both zero. Therefore, if the minimum is used as the aggregation operator, derivable histories will get the lowest interestingness rating. The apparent difference between both approaches is, however, that the pruning approach either rejects a history or keeps it. Therefore many other interestingness measures or generally post-processing techniques can be subsequently applied to non-derivable rules, e.g. the homogeneity measure with the average as the aggregation operator. In fact, the homogeneity interestingness measure's strength is its ability to aggregate information related to multiple more general rules into one value, while the pruning approach does not provide this flexibility.

Figure 7.7 shows the confidence histories of a top ranked rule and its two more general rules. To obtain the ranking the deviation measure (7.16) has been utilised in combination with the average as the aggregation operator in order to rank all histories which are stable or exhibit a trend. The Figure 7.7 illustrates a typical scenario for inhomogeneity: the slightly inclining value of $x \rightarrow z$ and the drastic downward trend in the history of $y \rightarrow z$ significantly contrasts the moderate decline in the history of their common specialisation $xy \rightarrow z$.

Chapter 8

Conclusions

8.1 Summary

Many businesses collect huge volumes of time-stamped data. This data reflects changes in the domain from which it has been derived. It is crucial for the success of most businesses to detect these changes, correctly interpret their roots and finally to adapt or react to them. For this reason there is a significant need for data mining approaches which are capable to find the most relevant and interesting changes within a dataset.

In this thesis, I designed and tested a framework for discovering interesting trends and stabilities in the support and confidence histories of association rules.

I discussed that trend and stability are both change patterns which allow a qualitative statement about the future development of a rule and thus support proactive business decisions. Furthermore, a review of publications on rule change mining yielded that most of the change patterns or, in a wider context, "types of change" suggested in them are inferior to the change patterns I used in terms of their interpretability and usefulness in business applications.

The framework's basic architecture is made up of three layers conceptually building upon each other. A different task is assigned to each layer which itself consists of several methods. In the following paragraphs I will summarise each layer, its methods and the main results.

The task of the *mining layer* is to discover, store and manage association rules and their histories. Its core component is a system for association rule discovery. In this context, I discussed that pruning approaches for association rules which are based on time-variant properties of the rules should not be used in conjunction with rule change mining, because rather interesting histories may be discarded. Surprisingly, this issue has never been mentioned nor discussed in any other publication. Furthermore, I proposed a database scheme which efficiently stores and manages the discovered rules and their histories.

The task of the *detection layer* is to discover change patterns in histories. For trend detection I employed two non-parametric statistical tests for trend, known as the Mann-Kendall and Cox-Stuart test, respectively. They are far superior to the "runs test" used in other publications which, as I demonstrated, is not suitable for trend detection. A comparison between both trend tests showed that the Mann-Kendall test, compared to the Cox-Stuart test, is linked with a higher computational effort, but seems to be less susceptible to noise. For this reason, choosing the right method depends on which of the two aspects, reliability or speed, is more important for a certain business application. For stability detection I developed an approach based on the χ^2 test. The approach is an enhancement of a method suggested by (Liu et al., 2001b) for which I demonstrated that it may also classify clear trends as stable behaviour. To generally improve the reliability of the described detection methods, I applied double exponential smoothing – a common denoising technique – to the histories.

The task of the *evaluation layer* is to post-process the detected change patterns in order to support a user in identifying the most relevant and interesting ones. Due to the typically vast set of detected change patterns this task can be considered vital for practical applications of rule change mining. Surprisingly, it is rather unaccounted for in research so far. Furthermore, I demonstrated that the only approach published in this area, the pruning method of (Liu et al., 2001a), sometimes counterintuitively discards histories.

I proposed a pruning approach based on the notion of a derivative rule history, that is a history that can be explained with histories of more general rules and can therefore be discarded. I introduced three different criteria for the derivativeness of histories and showed how to implement statistical tests to check if a history meets them. The proposed method thereby also overcomes another limitation of the aforementioned approach by (Liu et al., 2001a) as the criteria take entire rule histories into account. Furthermore, I showed that although existing pruning techniques for association rules cannot be directly applied in rule change mining, the method is consistent with some of them.

To identify the most interesting trends, I proposed a collection of metrics for interestingness assessment. Such metrics are novel for rule change mining, though similar concepts are broadly used in the area of association rules. To underpin the metrics psychologically, I introduced a model of the user's naive assumptions about rule change and defined interestingness as the degree of deviation of a trend's properties from the model.

Using real-life data taken from surveys I showed the effectiveness of the framework and its usefulness in solving real business problems. The goal of the analysis conducted was the discovery of interesting changes in the attitude and satisfaction of customers. In the following I will briefly summarise the experimental results presented in each chapter. The mining layer derived 77401 rules from the data and their support and confidence histories. Based on confidence histories the detection layer discovered a trend for 30394(39.3%) and stability for 21753(28.1%) of them. The proposed pruning approach significantly reduced these figures to 12369(23.2%) for trends and 15943(20.5%) for stabilities, where the percentages relate to the overall number of histories (77401). Based on support histories the detection layer discovered a trend for 43207(47.4%) and stability for 2019(2.8%) of them. Again, the proposed pruning approach significantly reduced these figures to 14242(18.2%) for trends and 1220(1.5%) for stabilities. The pruned histories with trend have been assessed using the proposed interestingness metrics and some of the high ranked trends have been assessed by domain experts. The experts, in turn, confirmed their interestingness by judging them as previously unknown and surprising. Overall, the proposed framework helped considerably to analyse the dataset and gave valuable insight into the problem domain which would not have been so easily possible with other data mining methods.

In summary, this thesis contributes to research on rule change mining – a promising and likewise challenging new area of data mining – in several ways: first, it is the first publication which proposes a framework that integrates all tasks connected to rule change mining – from association rule discovery to the interestingness assessment of change patterns. Second, it critically surveys the state of the art in rule change mining from a business perspective and questions some existing approaches in this context. Third, it proposes solutions for several previously unaccounted for or insufficiently solved problems, e.g. pruning and interestingness assessment.

8.2 Future Work

Research on rule change mining is still in its early stages and consequently there are many issues which still need to be addressed and solved. I have identified four areas that I believe merit future work which would significantly enhance the capabilities of the proposed framework. The first two areas suggest enhancement to the proposed methods, and the remaining two address possible extensions of the framework's architecture.

1. The detection layer should be extended to detect a greater variety of change pattern types, for example, cyclical variations or sudden changes in the level of the sequence. Currently, the latter is detected as a trend.

2. The proposed pruning approach is currently based on three criteria. However, it is likely that further criteria can be found which classify a history as derivative yet are meaningful and interpretable. Furthermore, in order to extend the framework to histories of other rule measures, like *lift* or *conviction*, effective derivativeness criteria have to be developed for them.

3. To interpret an interesting change pattern found for a rule it is useful to examine the histories of related rules. Therefore, a possible future extension to the framework is a further layer which provides a user interface for rule and history browsing.

4. For many business applications it may be useful not only to detect change patterns, but also the turning points between them. This can be, for example, the time point at which an upward trend turns into a downward trend or a stability turns into a trend. To support this kind of analysis it may be necessary to extend the framework's architecture by a database which manages detected change patterns.

Nomenclature

Association Rule Mining

$\mathrm{asupp}(r)$	antecedent support of the rule r
$\mathrm{conf}(r)$	confidence of the rule r
$\mathrm{conf}_{\mathrm{min}}$	lower threshold for confidence
$\mathcal{D}$	transaction set
$\mathcal{I}(\mathcal{D})$	set of frequent itemsets generated from $\mathcal{D}$
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$	itemsets
$\mathcal{X}\mathcal{Y}$	union of the itemsets $\mathcal{X}$ and $\mathcal{Y}$
$\mathcal{R}(\mathcal{D})$	set of association rules generated from $\mathcal{D}$
$\mathrm{supp}(r)$	support of the rule r
$\mathrm{supp}_{\mathrm{min}}$	lower threshold for support
$\mathcal{T}$	transaction
$r \succ r'$	r is a generalisation of r'
r	association rule
x, y, z	items

Rule Change Mining

$\mathrm{asupp}(r, T)$	antecedent support of the rule r in period T
$\mathrm{conf}(r, T)$	confidence of the rule r in period T
$\mathcal{D}(T)$	transaction set collected in period T
$\hat{\mathcal{R}}$	compound rule set, i.e. the maximal subset of rules discovered in all periods
$\hat{T}$	sequence of consecutive, non-overlapping periods
$\mathrm{supp}(r, T)$	support of the rule r in period T
$H_{\mathrm{asupp}}(r)$	antecedent support history of the rule r

$H_{\text{conf}}(r)$	confidence history of the rule r
$H_{\text{supp}}(r)$	support history of the rule r
T	time period
t	time point
ψ	interestingness measure for change patterns

Statistics

$\bar{v}$	sample mean of v
$B(n, p)$	binomial distribution of n Bernoulli trials with a success probability of p
C	test statistic of the Mann-Kendall test
$N(\mu, \sigma^2)$	Gaussian distribution with mean μ and variance σ^2
$P(\mathcal{X})$	probability that the itemset $\mathcal{X}$ is contained in a transaction
s	sample standard deviation, standard error
s^2	sample variance

Bibliography

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

Agrawal, R. and Psaila, G. (1995). Active data mining. In Fayyad, Usama, M. and Uthurusamy, R., editors, *Proceedings of the 1st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3–8, Montreal, Quebec, Canada. AAAI Press, Menlo Park, CA, USA.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In Bocca, J. B., Jarke, M., and Zaniolo, C., editors, *Proceedings 20th International Conference on Very Large Databases*, pages 487–499, Santiago, Chile.

Andersson, E., Bock, D., and Frisen, M. (2004). Detection of turning points in business cycles. *Journal of Business Cycle Management and Analysis*, 1(1):93–107.

Au, W.-H. and Chan, K. (2002). Fuzzy data mining for discovering changes in association rules over time. In *Proceedings of the 11th IEEE International Conference on Fuzzy Systems*, pages 890–895.

Au, W.-H. and Chan, K. (2005). Mining changes in association rules: a fuzzy approach. *Fuzzy Sets and Systems*, 149(1):87–104.

Baron, S. (2003). Identifizierung kurz- und langfristiger Musteränderungen in Transaktionsdaten. In *GI-Workshopwoche LLWA*.

Baron, S. (2004). *Temporale Aspekte entdeckten Wissens – Ein Bezugssystem für die Evolution von Mustern*. PhD thesis, Humboldt University Berlin.

Baron, S. and Spiliopoulou, M. (2001). Monitoring change in mining results. In *Proceedings of the 3rd International Conference on Data Warehousing and Knowledge Discovery*, Munich, Germany.

Baron, S. and Spiliopoulou, M. (2003). Monitoring the evolution of web usage patterns. In *1st European Web Mining Forum, Workshop at ECML/PKDD 2003*.

Bayardo, R., Agrawal, R., and Gunopulos, D. (2000). Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4(3):217–240.

Bayardo, Jr., R. J. and Agrawal, R. (1999). Mining the most interesting rules. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 145–154.

Borgelt, C. and Kruse, R. (2002). *Graphical Models.* John Wiley & Sons.

Chakrabarti, S., Sarawagi, S., and Dom, B. (1998). Mining surprising patterns using temporal description length. In *Proceedings of the 24th International Conference on Very Large Databases*, pages 606–617. Morgan Kaufmann Publishers Inc.

Chatfield, C. (1996). *The Analysis of Time Series – An Introduction.* Chapman and Hall/CRC.

Chatfield, C. (2001). *Time-Series Forecasting.* Chapman and Hall/CRC.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* John Wiley & Sons, New York.

Cox, D. and Stuart, A. (1955). Some quick sign tests for trend in location and dispersion. *Biometrika*, 42:80–95.

Dong, G., Han, J., and Lakshmanan, L. (2003). Online mining of changes from data streams - research problems and preliminary results. In *Proceedings of the ACM SIGMOD Workshop on Management and Processing of Data Streams.*

Dong, G. and Li, J. (1999). Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* AAAI Press and MIT Press, Menlo Park and Cambridge,MA,USA.

Fayyad, U. M., Piatetsky-Shapiro, G., and Uthurusamy, R. (2003). Summary from the KDD-03 panel: Data mining – the next 10 years. *SIGKDD Explorations Newsletter*, 5(2):191–196.

Fogel, D. B. (1997). The advantages of evolutionary computation. In Lundh, D., Olsson, B., and Narayanan, A., editors, *Bio-Computing and Emergent Computation.* World Scientific Press, Singapore.

Gilbert, R. (1987). *Statistical Methods for Environmental Pollution Monitoring.* Van Nostrand Reinhold, New York.

Goethals, B. and Zaki, M. J. (2004). Advances in frequent itemset mining implementations: report on FIMI'03. *SIGKDD Explorations Newsletter*, 6(1):109–117.

Goldin, D. Q. and Kanellakis, P. C. (1995). On similarity queries for time-series data: Constraint specification and implementation. In *Proceedings of the 1st International Conference on Principles and Practice of Constraint Programming*, pages 137–153. Springer-Verlag.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation. *Data Mining and Knowledge Discovery*, 8:53–87.

Helsel, D. and Hirsch, R. (1992). *Statistical Methods in Water Resources.* Elsevier, Amsterdam.

Hilderman, R. J. and Hamilton, H. J. (1999). Knowledge discovery and interestingness measures: A survey. Technical report, Department of Computer Science, University of Regina.

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for association rule mining - a general survey and comparison. *SIGKDD Explorations Newsletter*, 2(2):1–58.

Hussain, F., Liu, H., Suzuki, E., and Lu, H. (2000). Exception rule mining with a relative interestingness measure. In *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 86–97. Springer-Verlag.

James, W. (1890). *The Principles of Psychology.* Holt.

Kimball, R. (1996). *Data Warehouse Toolkit: Practical Techniques for Building High Dimensional Data Warehouses.* John Wiley & Sons.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

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

Liu, B., Hsu, W., and Chen, S. (1997). Using general impressions to analyze discovered classification rules. In *Proceedings of the 3rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 31–36.

Liu, B., Hsu, W., and Ma, Y. (1999). Pruning and summarizing the discovered associations. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 125–134. ACM Press.

Liu, B., Hsu, W., and Ma, Y. (2001a). Discovering the set of fundamental rule changes. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 335–340.

Liu, B., Ma, Y., and Lee, R. (2001b). Analyzing the interestingness of association rules from the temporal dimension. In *Proceedings of the IEEE International Conference on Data Mining*, pages 377–384. IEEE Computer Society.

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

Montgomery, D. and Runger, G. (2002). *Applied Statistics and Probability for Engineers.* John Wiley & Sons.

Ng, R. T., Lakshmanan, L. V. S., Han, J., and Pang, A. (1998). Exploratory mining and pruning optimizations of constrained associations rules. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 13–24.

NIST (2004). NIST/SEMATECH e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (1999). *Discrete-time Signal Processing.* Pearson.

Padmanabhan, B. and Tuzhilin, A. (1999). Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27.

Padmanabhan, B. and Tuzhilin, A. (2000). Small is beautiful: discovering the minimal set of unexpected patterns. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 54–63.

Padmanabhan, B. and Tuzhilin, A. (2002). Knowledge refinement based on the discovery of unexpected patterns in data mining. *Decision Support Systems*, 33(3):309–321.

Pei, J. and Han, J. (2002). Constrained frequent pattern mining: a pattern-growth view. *SIGKDD Explorations Newsletter*, 4(1):31–39.

Piatesky-Shapiro, G. and Matheus, C. J. (1994). The interestingness of deviations. In *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases*, pages 25–36.

Piatetsky-Shapiro, G. (2000). Knowledge discovery in databases: 10 years after. *SIGKDD Explorations Newsletter*, 1(2):59–61.

Rafiei, D. and Mendelzon, A. (1997). Similarity-based queries for time series data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 13–25. ACM Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.

Shah, D., Lakshmanan, L. V. S., Ramamritham, K., and Sudarshan, S. (1999). Interestingness and pruning of mined patterns. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*.

Sheskin, D. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures, Third Edition*. Chapman and Hall.

Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974.

Silverstein, C., Brin, S., and Motwani, R. (1998). Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2(1):39–68.

Spiliopoulou, M. and Baron, S. (2002). Monitoring the results of the kdd process: An overview of pattern evolution. In Meij, J. M., editor, *Dealing with the Data Flood: Mining data, text and multimedia*, chapter 6, pages 845–863. STT Netherlands Study Center for Technology Trends, Den Haag, Netherlands.

Srikant, R., Vu, Q., and Agrawal, R. (1997). Mining association rules with item constraints. In Heckerman, D., Mannila, H., Pregibon, D., and Uthurusamy, R., editors, *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pages 67–73. AAAI Press.

Suzuki, E. and Zytkow, J. M. (2000). Unified algorithm for undirected discovery of exception rules. In *Proceedings 4th European Conference on Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Computer Science, page 169. Springer-Verlag.

Tan, P.-N. and Kumar, V. (2000). Interestingness measures for association patterns: A perspective. Technical report, Department of Computer Science, University of Minnesota.

Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313.

Tjaden, G. (1996). Measuring the information age business. *Technology Analysis and Strategic Management*, 8(3):233–246.

Tjaden, G. (1997). Its the information, stupid! (not the computers). Technical report, Georgia Institute of Technology.

Wang, K., Jiang, Y., and Lakshmanan, L. V. S. (2003). Mining unexpected rules by pushing user dynamics. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 246–255.

Watterson, K. (1995). A data miner's tools. *BYTE Magazine*, October 1995.

Webb, G. I. (2000). Efficient search for association rules. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 99–107.

Widmer, G. and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101.

Wilson, R. A. and Keil, F. C. (1999). *The MIT Encyclopedia of the Cognitive Sciences*. Bradford Book.

Yager, R. R. (2000). A hierarchical document retrieval language. *Information Retrieval*, 3(4):357–377.

Yue, S., Pilon, P., and Caradias, G. (2002). Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrologics*, 259:254–271.

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

Zaki, M. J. (2004). Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9(3):223–248.

Zaki, M. J. and Hsiao, C.-J. (2002). CHARM: An efficient algorithm for closed itemset mining. In *Proceedings of the 2nd SIAM International Conference on Data Mining*, Arlington, VA. SIAM.

Zhang, X., Dong, G., and Kotagiri, R. (2000). Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 310–314.

Selbständigkeitserklärung

Hiermit versichere ich, daß ich die vorliegende Diplomarbeit selbständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht habe, was aus Arbeiten anderer unverändert oder mit Abänderung entnommen wurde.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form für keine andere Prüfung vorgelegt.

Magdeburg, den 24. April 2005

Mirko Böttcher