The Knowledge Engineering Review, Vol. 20:1, 39–61. © 2005, Cambridge University Press
doi:10.1017/S0269888905000408 Printed in the United Kingdom

A survey of interestingness measures for knowledge discovery

KEN MCGARRY

School of Computing and Technology, Informatics Building, University of Sunderland, St Peters Campus, St Peters Way, Sunderland SR6 ODD, UK;
E-mail: ken.mcgarry@sunderland.ac.uk

Abstract

It is a well-known fact that the data mining process can generate many hundreds and often thousands of patterns from data. The task for the data miner then becomes one of determining the most useful patterns from those that are trivial or are already well known to the organization. It is therefore necessary to filter out those patterns through the use of some measure of the patterns actual worth. This article presents a review of the available literature on the various measures devised for evaluating and ranking the discovered patterns produced by the data mining process. These so-called interestingness measures are generally divided into two categories: objective measures based on the statistical strengths or properties of the discovered patterns and subjective measures that are derived from the user's beliefs or expectations of their particular problem domain. We evaluate the strengths and weaknesses of the various interestingness measures with respect to the level of user integration within the discovery process.

1 Introduction

Data mining and knowledge discovery from databases has received much attention in recent years. However, the majority of this work has concentrated on producing accurate models with little consideration for the ranking and evaluation of the value of patterns. Recent work has to some extent addressed this problem through the development of various forms of interestingness measures (Fayyad & Stolorz, 1997; Hilderman & Hamilton, 2003). It is the purpose of this article to provide an overview of previous work and to highlight the latest research and future directions.

For those managers involved with making decisions, the complexity and level of detail required for making accurate and timely choices is now more difficult than ever. The scale of the data problem overwhelms most of the traditional approaches to data analysis. The hidden knowledge locked away in corporate data stores has a great deal of potential that can theoretically be uncovered by data mining. Data mining is an area of intense activity, both as a research topic for developing new algorithms and techniques but also as an application area where real benefits are to be made (Hand *et al*, 2001). Interestingly, data mining is much more than number crunching as, in addition to intelligent techniques and statistics, it also encompasses elements from human–computer interaction, cognitive science, databases, information technology and information retrieval (Pazzani, 2000).

The application of intelligent techniques has meant that much information previously hidden or buried within the sheer scale of the data can be revealed. Many areas such as fraud detection, fault prediction in telecommunications networks, health care etc. have all benefited from the application of data mining (Fayyad *et al*, 1996a).

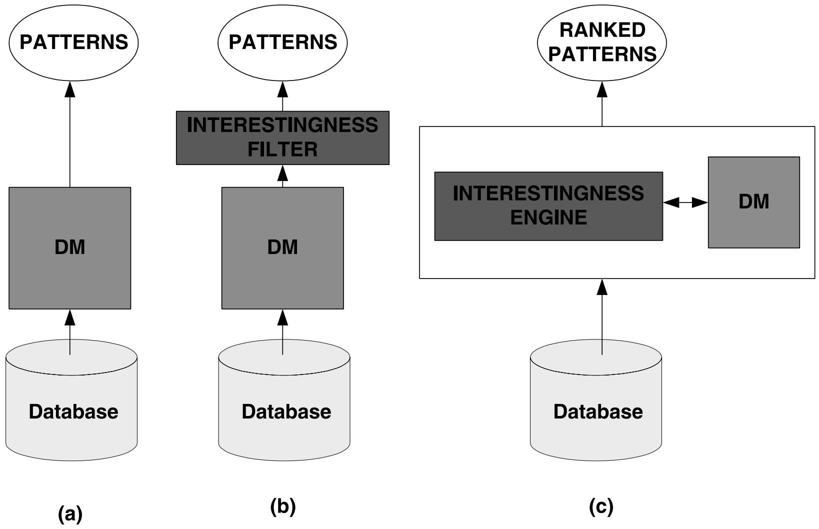

Figure 1 Techniques for knowledge discovery. (a) shows that all patterns produced by the data mining process are passed to the user. (b) shows how the search for interesting patterns occurs as post-processing effort. The preferred method would be to integrate the search for interesting patterns within the data mining algorithm as in (c)

In Figure 1, three main techniques of pattern assessment are shown. The approach taken in Figure 1(a) is a simple data dredging approach that is cumbersome and time consuming as every pattern generated is presented to the user. In Figure 1(b) a filter is used as a post-processor to sift out the interesting patterns according to some criteria or guidelines. In Figure 1(c) a much more intelligent scheme is presented whereby the data mining algorithms are working in conjunction with the pattern assessment element to provide a much more dynamic approach to knowledge discovery. This typically involves an element of feedback as the search for interesting patterns proceeds. This is where the challenge to the data mining community lies and involves the development of interestingness measures and the integration of the user within the system.

In this paper we give an overview of the continued development of interestingness measures and how they have evolved into the two main techniques of objective and subjective measures of interest. The objective approach uses the statistical strength or characteristics of the patterns to assess their degree of interestingness. Subjective techniques incorporate the subjective knowledge of the user into the assessment strategy. The remainder of this paper is structured as follows. Section 2 reviews several important aspects of data mining with regard to the automated discovery of knowledge and in particular we describe the structure and composition of *patterns*. Section 3 describes in a general way the motivation, goals and problems of developing interestingness measures. Section 4 concentrates on objective interestingness measures. Section 5 reviews subjective measures and describes how a user can define their domain knowledge. In Section 6, the focus is on evaluating the interestingness measures developed specifically for association rules and we also provide a brief review of association rule operation. Section 7 presents the conclusions and the likely trends for future research directions.

2 Data mining and knowledge discovery

One of the most insightful definitions of data mining states that to be truly successful data mining shouldbe '*the non-trivial process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge from databases*' that is used to make crucial business decisions (Fayyad *et al*, 1996a).

- Non-trivial: rather than simple computations complex processing is required to uncover the patterns that are buried in the data.

Table 1 Knowledge representation techniques

Technique	Data types	Understandable	Model
Propositional rules	Categorical	Yes	Classification
First-order rules	Categorical	Yes	Classification
Association rules	Categorical	Yes	Association/co-occurrence
Decision trees	Categorical	Yes	Classification
Fuzzy logic	Continuous	Yes	Classification
Clustering	Categorical, continuous	Yes	Self-organizing
Neural networks	Continuous	No	Classification, self-organizing
Bayesian networks	Categorical	Yes	Classification

- Valid: the discovered patterns should hold true for new data.
- Novel: the discovered patterns should be new to the organization.
- Useful: the organization should be able to act upon these patterns to become more profitable, efficient etc.
- Comprehensible: the new patterns should be understandable to the user and add to their knowledge.

There are clear overlaps between statistics and data mining, Glymour *et al.* (1997) and Hand (1999) provide some insights into this. Other issues include the interfacing of data mining algorithms to transactional database systems (Chen *et al*, 1996; Lavington *et al*, 1999). Matheus *et al.* (1993) describes the challenges facing the user when designing a system for knowledge discovery and compares the results from three systems: Explora, CoverStory and the KDD Workbench. Work by St Amant & Cohen (1998) examines the issues from exploratory data analysis viewpoint. A useful collection of knowledge discovery papers edited by Piatetsky-Shapiro & Frawley (1991) covers many aspects of interestingness. A further volume was produced that covers many later advances in data mining and knowledge discovery (Fayyad *et al*, 1996b).

2.1 What is a Pattern?

There appears to be no standard definition of the term ‘pattern’. However, Frawley *et al.* (1991) provide a working definition that is quite general and broad enough to be applied to several sources of information.

> *‘Given a set of facts (data) F, a Language L, and some measure of certainty C, a pattern S is a statement S in L that describes relationships among a subset Fs of F with certainty C, such that S is simpler (in some sense) than the enumeration of all facts in Fs.’*

This definition is general enough to apply to any approach that produces patterns that are a generalization or summary of the initial data set. There are many different data mining algorithms with different objectives, different outcomes and with different representation techniques (see Table 1).

One unifying metric is that of similarity or dissimilarity between patterns. In fact many data mining techniques such as clustering depend on similarity measures between objects. For example, $d(i, j)$ is a metric that denotes the dissimilarity of two objects but must satisfy three conditions (Hand *et al*, 2001):

1. $d(i, j) \geq 0$ for all i and j, and $d(i, j) = 0$ if and only if $i = j$;
2. $d(i, j) = d(i, j)$ for all i and j; and

Rule 4:(172/1, lift 1.5)
Clump Thickness ≤ 6
Bland Chromatin ≤ 2
$\Rightarrow$ class 2 [0.989]

Rule 5: (112/2, lift 2.7)
Bare Nuclei > 2
Bland Chromatin > 3
$\Rightarrow$ class 4 [0.974]

Figure 2 Example of C5 rules derived from the Breast Cancer data set

3. $d(i, j) \leq d(i, k) + d(k, j)$ for all i, j and k.

Patterns are typically local in nature, i.e. a given pattern such as a rule would typically cover only a few of the records or instances in a database. However, it is necessary to define the syntax and semantics of the particular patterns we are investigating. The majority of data mining/machine learning type patterns are rule based in nature with a well-defined structure. Rules derived from decision trees and association rules generated from databases are the most common forms encountered (Berger & Tuzhilin, 1998).

Rule-based patterns are composed of a number of primitive patterns (antecedents) connected together with logical connectives that imply, if true, the target class (consequent). Figure 2 shows an example of propositional rules derived from the C5 algorithm (see Quinlan, 1992). The rules each have two antecedents that must be satisfied in order for the consequent to be true. The C5 rules have additional information such as *lift* and *confidence* factors attached.

Other conditions such as temporal and spatial factors will make a significant impact upon the analysis undertaken. For a survey of temporal data mining techniques see Roddick & Spilioppoulou (2001).

2.2 User involvement and automation

Data mining and knowledge discovery is generally regarded as a process with the user involved at every stage, from the initial data pre-processing to the final analysis. The user is important because it is human judgement that ultimately determines if a discovered pattern is useful knowledge or not. However, as Gaines (1996) points out, this is not the easiest metric or criterion to implement in a data mining system because the most interesting discoveries are those that are *unforeseen* and *surprising*, i.e. we all know a novel pattern when we see one but find it difficult to provide the guidance necessary to discover one.

However, there are many situations where an automated discovery system would be beneficial, and only involving the user when assessing the interesting patterns. The characteristics necessary for an automated discovery system have been identified by Zytkow (1993) to be:

- to select the most relevant parameters and variables;
- to guide the data mining algorithms in selecting the most salient parameters and in searching for an optimum solution;
- to identify and filter the most meaningful results to the user;
- to be able to identify useful target concepts.

A possible semi-automatic arrangement is shown in Figure 3, based on the concept of templates (Silberschatz & Tuzhilin, 1996a). The Data Monitoring and Discovery Triggering (DMDT) system limits user involvement to providing feedback to guide the system as it searches. Pattern templates have been used in the past to specify interesting association rules but the form used by Silberschatz is first-order logic, IF..THEN type rules (see Klemettinen *et al*, 1994). The DMDT system is intended to scale up on large industrial data sets that may be composed of many databases. Over

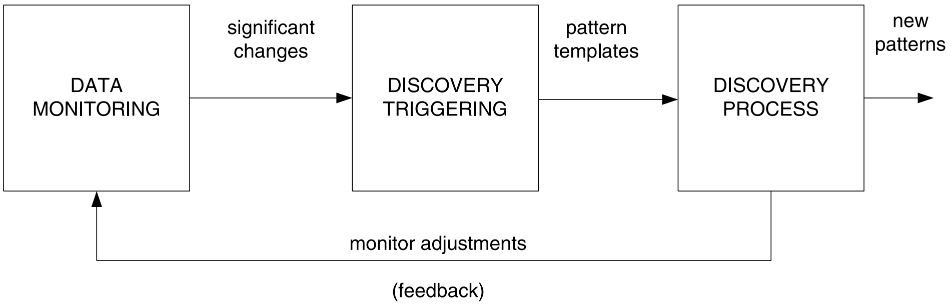

Figure 3 User-based feedback within the knowledge discovery cycle (from Silberschatz & Tuzhilin, 1996a)

a period of time, the pattern templates defining the interesting rules will change as the data changes and will thus *trigger* new discoveries.

There is a vast body of related work on scientific discovery that can be used by data miners, for example the seminal work by Lenat on the use of heuristics for discovery provides useful insights into the automated discovery process (Lenat, 1983; Lenat & Brown, 1984). The AM system developed by Lenat & Brown (1984) used 43 heuristics dedicated to assessing concepts; a concept was deemed to be interesting if it is similar to a related concept that had been previously classed as interesting. Many such systems have been developed over the years; for example, Zytkow (1993) provides an interesting overview of the similarities and differences between *learning* and *discovery* by automated systems. Kulkarni & Simon (1988) developed the REDUCE system that investigated the process of scientific discovery. For a good overview of such systems, see the article by Langely *et al.* (1983) in the collection edited by Michalski and Carbonell.

Automated discovery systems that incorporate domain knowledge to assess the novelty of discovered patterns, such as the system proposed by Ludwig & Livingstone (2000), generally use this knowledge to create a prior model that can be revised with new information. The authors define novelty as '*a hypothesis H is novel, with respect to a set of beliefs B, if and only if H is not derivable from B*'. Therefore, if a new pattern contradicts a set of beliefs then such a pattern should be surprising and novel. Therefore, their system places an important role for background knowledge in the discovery process. The authors continued their work to develop an agenda and justification-based system for knowledge discovery within biosciences (Livingstone *et al*, 2003). The system is able to estimate the plausibility of the user's tasks and will suggest modifications using heuristics. The modified concepts and hypotheses are examined to identify novel and interesting relationships. Related work by Colton & Bundy (2000) relates to using empirical evidence for discovering the most interesting patterns. The authors compared five automated mathematical discovery systems with regard to how they identified interesting concepts. Colton and Bundy noted that the systems appeared to have certain interestingness criteria in common such as novelty, plausibility and non-triviality when assessing the worth of conjectures and using models, truth-maintenance and understand ability when assessing concepts.

An appreciation of the problems inherent with automating a user-centric process is tackled by Zhong *et al.* (2001). The authors believe that only a human-centered system can discover knowledge and that support for keeping the human in the loop should be maintained. This assumption is supported by the *Conceptual Knowledge Discovery in Databases* (CKDD) system by Correia *et al.* (2003). The CKDD system is highly visual in nature and uses graph structures to display information and enables the user to test out a hypothesis by detailed inspection of the graph structures. Experts are able to modify the values linking the nodes and hence integrate domain knowledge directly into the system. The CKDD system was used to investigate flight activity at a major European airport using spatial and temporal data.

Hilario & Kalousis (2000) tackle the issues of model selection for an automated system. They build individual profiles for different learning techniques for a variety of algorithms such as C5, neural networks and rule induction programs. Hilario and Kalousis are effectively defining model-based, interestingness measures based on knowledge representation, efficiency, resilience and implementation details.

An example of an automated data mining system is by Merzbacher (2002) who uses the AM approach to generate trivia questions from a database. Klösgen (1995) developed the Explora system to generate and assess hypothesis by using objective interestingness measures. Explora allows the user to constrain the search space by creating a hierarchically organized set of hypotheses. The chosen hypothesis will examine the patterns produced by the search effort. The user can specify the extent of the *usefulness* of a particular pattern according to their specific goals. This information is then used by Explora to refine the original search. However, the dialog between user and system is made tractable only through efficient computation and 'crude' granularity tradeoffs. Brachman (1993) describes *data archaeology* as essentially a human oriented task that requires special support since the majority of the goals cannot be specified in advance. Brachman also notes that any hypothesis may need to be refined given sufficient evidence and notes that current SQL and other batch-based, data management tools cannot directly support this. Their IMAC system provides the user with the necessary tools to interactively support the user.

3 Overview of interestingness measures for knowledge discovery

The majority of the data mining literature concentrates on the discovery of accurate and comprehensible patterns. Whilst this approach will provide the user with a degree of confidence regarding the discoveries it falls far short of the notion of 'knowledge discovery'. The essential task for data mining algorithms is to search for patterns that are 'surprising' in addition to being accurate and comprehensible.

For example, the Wal-Mart (possibly apocryphal) data mining story of how they used association rules on their market basket data to reveal that there was a strong correlation with the purchase of nappies and beer, i.e.

$$\text{nappies} \Rightarrow \text{beer.}$$

What exactly did Wal-Mart discover? These purchases were typically small numbers of items but nappies and beer did co-occur quite regularly. Checking their store loyalty card details highlighted that these were young fathers who were probably sent out of the house by their partners to purchase the nappies but in the meantime decided to treat themselves to some beer. Wal-Mart took the obvious action and placed beer at the ends of the baby section aisles and increased their profits. The story is intended to highlight that an interesting pattern must also be actionable or useful to be valuable. But what would have happened if the association rules had discovered the following correlation?

$$\text{nappies} \Rightarrow \text{babyfood}$$

We can reasonably expect that the sales of baby food and nappies occur together frequently. Therefore, an association rule representing this pattern would be accurate and undoubtedly would have a high degree of confidence it would not be classed as an interesting pattern. It is certainly not a surprising discovery to make, nor would it be actionable.

Figure 4 presents a taxonomy of interestingness measures, showing how the two main groups of subjective and objective are composed. Objective criteria such as rule coverage, rule complexity, rule confidence and rule completeness are often used as a measure of the interestingness of the discovered patterns. Subjective criteria such as unexpectedness, actionability and novelty are harder to define because they are domain specific. Although a pattern may be unexpected, this does not mean it is useful. This could be a simple case of the data mining algorithm detecting noise or

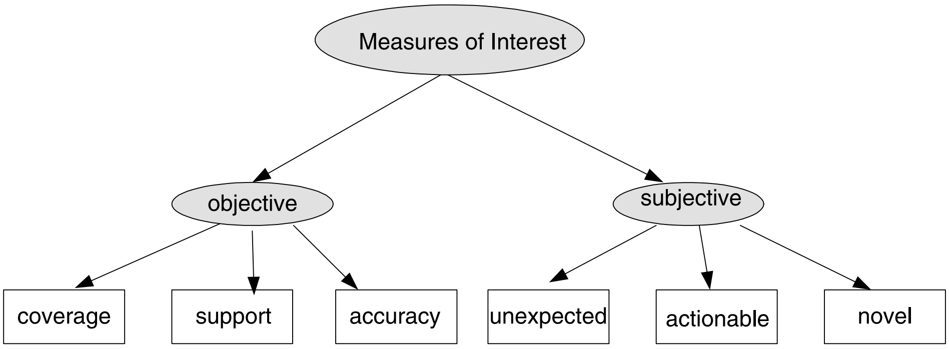

Figure 4 Taxonomy of interestingness measures

outliers. Even if a pattern is deemed to be unexpected and valid, it may not be possible to act upon it. Actionability ultimately defines if a pattern is truly useful to an organization and is explained further in Section 5.2.

3.1 The use of deviations and similarity measures

Early work by Piatetsky-Shapiro & Matheus (1994) led to grouping of deviations from normative expectations by means of *utility functions* within the KEFIR system. The utility functions were quite easy to define, as the system was intended to save money in a health-care application. The interestingness measures were based on the *actionability* of a particular pattern by measuring the savings anticipatedfrom taking a specific action. The system then recommended to the user the most cost-effective approach to take. Further refinements of the *deviation* approach by Anand *et al.* (1998) led to improved performance in providing a methodology for the support of cross-sales in a commercial domain. It should be noted that domain knowledge was used in both systems and played an important role to determine the effectiveness of the deviation measures, i.e. they were not domain independent.

The *Context Sensitive Similarity Discovery* (CSSD) method was developed by Roussinov & Zhao (2003) for detecting the semantic relationships within a Web mining application. The CSSD method automates the discovery of concepts from the analysis of online text messages. A similarity network expresses the inter-relationships between topics as a weighted graph. This is constructed through a hierarchial, agglomerating clustering process. The CSSD technique uses techniques from information retrieval as acquiring vocabularies, indexing and stemming must be performed to create the concepts. The clustering, or rather grouping of the concepts, can be performed manually; not surprisingly, the systems performance was worse when domain knowledge was from a user unfamiliar with the topics and hence the sensitivity to the *context* of the messages.

Roddick & Rice (2001) place the emphasis on simple thresholds based on pattern behavior on temporal sports data. The thresholds are modified over time and are context dependent; the intention is to provide a dynamic model of the situation and not to flag expected activity as interesting. The model takes into account the significance of an event and the possibility of it actually occurring. This enables a particular sportsman or team to be monitored for under achievement, which is of interest in this particular domain. However, Roddick and Rice do point out that if thresholds do exist *a priori*, then they would be better employed in guiding data mining algorithms to produce *only the most* interesting rules rather than post-processing the entire rule set.

4 Objective interestingness measures

Piatetsky-Shapiro & Frawley (1991) proposed three rule interestingness (RI) measures to objectively evaluate the values of patterns. The measures effectively quantify the correlation between the antecedents and the consequent:

1. RI $=0$ if $|A\&B|=|A||B|\backslash N$;
2. RI monotonically increases with $|A\&B|$ when other parameters are fixed;
3. RI monotonically decreases with $|A|$ or $|B|$ when other parameters are fixed;

where N is the total number of patterns, A and B are the tuples that satisfy A and B when RI $=0$, A and B are independent and the rule is not interesting. Major & Mangano (1995) proposed a fourth measure:

4. RI monotonically increases with $|A|$ with fixed confidence factor $Cf > Cf_0$.

Major and Mangano used a *dominance* factor that ranked the rules according to their place in a hierarchy that listed the rules as potentially interesting, technically interesting and genuinely interesting rules. To achieve this, Major & Mangano (1995) devised criteria such as performance, simplicity and significance, which they then applied to a Hurricane database. By applying these criteria the authors were able to identify ten genuinely interesting rules from an initial list of 161 rules. The author's developed a three-stage process. Stage one identifies potentially interesting (PI) rules directly from the data, based on set-points defined by the user. In stage two technically interesting (TI) rules are identified from the PI rules based on simplicity and statistical criteria. The TI rules are further pruned in stage three to reveal genuinely interesting (GI) rules. The GI rules are those that are simple, general and have no redundancy; in addition, further recall to an expert is taken at this stage. Thus, a mixture of objective measures acts as a preprocessor so as to build a list of interesting rules for partial elimination by an expert.

All four measures were designed to operate with the rule quality factors of coverage, completeness and confidence. However, several factors are assumed to be fixed but, as Freitas (1999) points out, in practice they often do not remain fixed.

Several systems have been developed that make use of these fundamental metrics, such as the Bank loan system of Ikizler & Gvenir (2001). This system ranked the importance of rules extracted from the C4.5 algorithm. Gago & Bento (n.d.) derive a metric for computing the distance between two rules, which are again are derived from the C4.5 algorithm. This metric was intended to select the best predictive rules rather than to analyze the interestingness of the rules *per se*. The two systems demonstrate the simplicity of developing distance metrics; however, they also illustrate that further analysis is required to qualify statistically strong but obvious patterns.

Work by McGarry used these techniques to evaluate the worth of rules extracted from RBF neural networks to discover their internal operation (McGarry *et al*, 2001a,b; McGarry & Malone, 2004). In this work the authors exploited the local nature of the RBF function, which is usually a Gaussian hypersphere and responds (theoretically) to only a small fraction of the input space. This enables simple, easy to comprehend rules to be extracted. However, the rules showed a high degree of redundancy when the RBF networks were trained on data sets with highly nonlinear or overlapping classes. This meant that the local response assumption was violated as the RBF networks tended to cover more and more of the input space as their basis functions overlapped. Eliminating redundant rules from the viewpoint of using them as a classifier meant a loss of accuracy; therefore, there was a trade-off between comprehensibility and accuracy. On the other hand, this approach provided insights into where a data set gave a neural network particular problems identifying which part of the input space required additional RBF hidden units.

The original measures proposed by Piatesky-Shapiro have been expanded by Kamber & Shinghal (1996) to reflect the difference between discriminant and *characteristic rules*. See the work by Han *et al.* (1993) for a discussion on characteristic rules.

Patterns or rules are usually more interesting if they *cannot* be predicted from existing knowledge. An objective measure using relative entropy to identify exception rules was developed by Hussain *et al.* (2000). Their technique seeks those rules that contradict a set of rules with strong support and confidence and are *exceptions* to existing knowledge and therefore interesting. The

Table 2 Exception rule structure

Format	Description
$A \rightarrow X$	Common sense rule (strong pattern)
$A,\ B \rightarrow \neg X$	Exception rule (weak pattern)
$B \rightarrow \neg X$	Reference rule (low support/confidence)

Table 3 Value of interesting patterns

Measure	Unexpected	Expected	Novel
Actionable	Quite interesting	Interesting	Very interesting
Unactionable	Interesting	Not interesting	Not interesting

authors are aware that exception rules typically have low levels of support and are computationally expensive to extract from large databases. However, exception rules often have strong levels of confidence similar to the *commonsense* rules. Hussain *et al.* used the contradictory approach to derive a set of rules; Table 2 presents the structure of an exception rule.

The interestingness (I) of a rule is defined by its support (S), confidence (C) and knowledge of common sense rules (κ) and is given by $I=f(S,\ C,\ \kappa)$. The relative entropy measure given by $D(p(x)||q(x))$ computes the differences in probability distributions when $q(x)$ is $p(x)$.

Furthermore, in Table 3 Hussain *et al.* attempted to rank the importance of a discovered rule based on whether it is actionable or not and if it is unexpected. The most valuable discovered patterns are those that are actionable and unexpected. However, these are subjective views, although *actionable* rules must be *reliable* in the sense that they can be acted upon. Such rules must be statistically significant, which is dependent upon the confidence and support levels.

Silberschatz & Tuzhilin (1996b) view the identification of interesting patterns as those more likely to be unexpected and actionable. A pattern is unexpected if it *surprises* the user and actionable if the user can *act* on this new information. We discuss these concepts in greater detail in the next section.

Graphical methods for interpreting data mining patterns have been used in several systems; these typically allow the overall structure of a data set to be visible without causing the user to be overwhelmed by the mathematical details. Buntine (1996) described a probabilistic graphical model that provides a framework where Bayesian networks, Markov models and influence diagrams can all be used to represent a knowledge discovery process to highlight where the dependencies and irrelevant information reside. Work by DiFore (2000) involved the development of a measure based on the relative frequencies of patterns that could be easily visualized by the user using the well-known star schema from the field of data warehousing. Notable work by Dykes & Mountain (2003) provides an overview of visualizing spatio-temporal patterns by assessing the geographical relevance of a pattern.

Romao *et al.* (2004) have used a genetic algorithm (GA) to optimize expert beliefs to rank the interestingness of fuzzy prediction rules. The authors have modified the approach of Liu *et al.* (1999), which involved the domain experts to define their knowledge in terms of *general impressions*, that are in the form of fuzzy rules. Romao *et al.* use the GA to dynamically maintain and search populations of rule sets for the most interesting rules rather than act as a post-processor. The rules identified by the GA compared favorably with the rules selected by the domain expert.

There are numerous ways of testing the statistical strength of a pattern. A principled approach was taken by Hilderman & Hamilton (1999, 2003), who used a heuristic-based method for

Table 4 Objective measures of interest

Measure	Description	Form
Shannon entropy	Relative entropy measure from information theory, calculates average information content (Shannon, 1948).	$-\sum_{i=1}^{m} P_i \,\mathrm{Log}_2 P_i$
Lorenz	Curve from statistics, calculates the probabilities associated with the data (Lorenz, 1905).	$q \sum_{i=1}^{m} (m - i + 1)P_i$
Gini	Inequality measure based on the Lorenz curve, uses the ratio of the Lorenz curve and the total area (Hilderman & Hamilton, 2001).	$\frac{q \sum_{i=1}^{m} q \sum_{j=1}^{m} \mid P_i - P_j \mid}{2}$
Kullback–Leiber	Uses the Shannon measure and a distance measure to calculate the difference between actual distribution and a uniform distribution (Kullback & Leiber, 1951).	$\log_2 m - \sum_{i=1}^{m} P_i \,\mathrm{Log}_2 \frac{\mathrm{P_i}}{\mathrm{q}}$
Atkinson	Inequality measure from economics, measures population distribution (Atkinson, 1970).	$\max(P_i)$

determining the usefulness of data mining patterns. Table 4 describes several measures of interest; Hilderman & Hamilton (2001) performed a detailed analysis of 16 such measures.

Although, there are many different measures available from a wide variety of application areas, it is relatively straightforward to devise a measure specifically for a given domain. Freitas (1998, 1999) recognized this and proposed that future interestingness measures could benefit by taking into account certain criteria.

- Disjunct size. The number of attributes selected by the induction algorithms prove to be of interest. Rules with fewer antecedents are generally perceived as being easier to understand and comprehend. However, those rules consisting of a large number of antecedents may be worth examining as they may refer either to noise or a special case.
- Misclassification costs. These will vary from application to application as incorrect classifications can be more costly depending on the domain (e.g, a false negative is more damaging than a false positive in Cancer detection).
- Class distribution. Some classes may be under represented by the available data and may lead to poor classification accuracies etc.
- Attribute ranking. Certain attributes may be better at discriminating between the classes than others. So discovering attributes present in a rule that were thought previously not to be important is probably worth investigating further.

The statistical strength of a rule or pattern is not always a reliable indication of novelty or interestingness. Those relationships with strong correlations usually produce rules that are well known, obvious or trivial. We must therefore seek another method to detect interesting patterns; the next section evaluates how we can devise measures that are domain dependent to subjectively rank our patterns.

5 Subjective interestingness measures

As stated in the last section, objective interestingness measures may not highlight the most important patterns produced by the data mining system. Subjective techniques generally operate by comparing the beliefs of a user against the patterns discovered by the data mining algorithm. The measures usually determine if a pattern is 'actionable' and/or 'unexpected'. Actionability refers to the organizations ability to do something useful with the discovered pattern. Therefore, a pattern can be said to be interesting if it is both unexpected and actionable. This is clearly a highly

Pattern 1: IF ($Age > 60$) $\wedge$ ($Salary = High$)
THEN Loan = approved
Pattern 2: IF ($Age < 60$) $\wedge$ ($Salary = Average$) $\wedge$ ($Record = poor$)
THEN Loan = not approved
Pattern 3: IF ($Age < 60$) $\wedge$ ($Salary = Low$)
THEN Loan = approved

Figure 5 Example rules

subjective view of the patterns as actionability is dependent not only on the problem domain but also on the user's objectives at a given point in time (see Silberschatz & Tuzhilin, 1996b).

5.1 Defining the knowledge/set of beliefs

There are techniques for devising belief systems and these typically involve a knowledge acquisition exercise from the domain experts. Other techniques use inductive learning from data and some also refine an existing set of beliefs through machine learning. Silberschatz & Tuzhilin (1995) view subjective interesting patterns as those more likely to be *both* unexpected and actionable. A number of techniques have been used in the past to define subjective domain knowledge.

- Probabilistic measures. Bayesian approaches have been used to enable conditional probabilities to be used. Silbershatz devised a system of 'hard' and 'soft' beliefs (see Silberschatz & Tuzhilin, 1995). The soft beliefs could be revised and updated in the light of new evidence. However, any new patterns that contradict hard beliefs would not result in modification of beliefs but would tend to indicate that these patterns are interesting.
- Syntactic distance measure. This measure is based on the degree of distance between the new patterns and a set of beliefs. More specifically, if the rule consequents are the same but the antecedents are greatly different then the system may have uncovered some interesting patterns (see Liu *et al*, 1999).
- Logical contradiction. Although devised for association rules, this technique uses the statistical strength (objective) of a rule, i.e. the confidence and support to determine if a pattern can be classed as unexpected or not (see Padmanabhan & Tuzhilin, 1999, 2002). Any conflict with between the degrees of user suggested by the user's and those actually generated by the association rules.

The work of Liu *et al.* (1999) is an important contribution that tackles the problem of acquiring the domain knowledge that is to be used by a fuzzy pattern matching algorithm. The user is expected to describe their knowledge in terms of fuzzy sets and linguistic variables. Fuzzy logic is an ideal technique to manage the vagueness and uncertainty involving human language terms. For example, a data mining algorithm produces the patterns as shown in Figure 5.

However, one of the patterns expected and defined by the user is as follows:

User Defined Pattern
IF (Age>50) $\wedge$ ($Salary$=Low)
THEN Loan=not approved

We can now rank the discovered patterns using our domain knowledge to either represent the closest match or alternatively to show the most unexpected patterns. Table 5 shows the three patterns ranked on this basis and the mismatch between the expert's beliefs and the learnt rule causes the slight difference in ranking.

Pattern 2 discovered by the data mining system conforms to the expert's belief and is very similar. Pattern 3 contradicts the user's belief and is therefore unexpected. Unexpected patterns can be analyzed in a variety of ways, such as rules containing unexpected attributes, unexpected

Table 5 Pattern rankings based on

Patterns ranked by similarity	Patterns ranked by unexpectedness
Pattern 2	Pattern 3
Pattern 3	Pattern 2
Pattern 1	Pattern 1

attribute values, unexpected consequents or patterns with overlapping attributes but with different consequents.

5.2 The role of actionability in knowledge discovery

Actionability requires further explanation, for example the projected savings measure of Piatetsky-Shapiro & Matheus (1994) is used to forecast the percentage savings in a medical domain:

$$PS = PI \times SP. \tag{1}$$

Here *PI* is the projected impact, *SP* is the savings percentage and *PI* is calculated by

$$PI = PD \times PF \tag{2}$$

where *PD* is the difference between the current average cost and the expected cost for a product or service, and *PF* is the impact factor that relates to increased profit. The savings percentage is gained from a domain expert and is a value of the percentage decrease in deviation that would occur following some relevant intervention strategy. For our overall scheme, automating the knowledge discovery process based on unexpected and actionable patterns presents a problem as actionable patterns are rather difficult to define formally in advance. Silberschatz & Tuzhilin (1996b) identified that this would entail dividing the pattern space into a set of equivalence classes and associating an action to each.

Another example, describing the feedback from students on the courses they have attended at a University, provides details on actionable and unexpected patterns (see Silberschatz & Tuzhilin, 1995). If a typical course should have a 60–90% response rate then the University would be surprised by a course with only an 8% response. This knowledge could be actionable in the future if the problem lay with shortcomings on the part of the faculty staff but would obviously be too late for the students who responded.

However, a number of considerations need to be observed.

- Constraints: what conditions or context must be true for the pattern to hold?
- Lifetime: how long will the pattern be useful for?
- Effort: what do we have to do to act upon the pattern?
- Side-effects: are there any anticipated side-effects?
- Impact: are there any changes to current practices?
- Immediacy: how soon can we act upon it?

Sahar (1999, 2001, 2002) has taken a different approach to the discovery of interesting patterns by eliminating non-interesting association rules. Rather than getting the user to define their entire knowledge of a domain, they are asked to identify several non-interesting rules, generated by the Apriori algorithm. The rule discovery system will semi-automatically eliminate those rules that appear to be similar. Several passes are required but Sahar calculates that the 5th pass will eliminate

Table 6 Sahar's subjective measures of interest

Rule type	Example
True-not-interesting	<*husband*>→<*married*>
Not-true-interesting	<*male*>→<*married*>
Not-true-not-interesting	<*pregnant*>→<*female*>
True-interesting/interesting	<*male*>→<*parent*>

50% of the rules. This approach avoids the difficult and time-consuming process of getting the domain expert to define why a particular rule is interesting. Sahar has classified the rules according to their validity and potential to be interesting. Table 6 illustrates the scheme. For example, the first rule states that a *husband* is a person who is *married*, which is clearly true but not interesting. The second rule, however, states that *males* are *married*, which is not true but interesting because we may wish to target all married males for a particular campaign. The last rule has discovered a strong co-occurrence between certain males who are fathers and therefore may be of interest for a particular marketing campaign.

The elimination process was performed on three real-world databases: transactions from a grocery store, a World Wide Web log and an adult census database with promising results. However, increasing the level of domain knowledge will improve the chances of discovering useful patterns but make the data mining process less general and therefore difficult to apply to other domains. The main advantage of Sahar's method is that it can be viewed as way of determining the *minimum* amount of domain knowledge required to seek out interesting rules and is therefore an important contribution.

5.3 The role of unexpectedness in knowledge discovery

If an unexpected pattern is identified by the data mining system then we may be surprised as this may contradict some previously held belief concerning our problem domain. Work by Padmanabhan & Tuzhilin (1998, 1999, 2002) has explored the interestingness problem through the notion of unexpectedness. An unexpected pattern is likely to conflict with some previously held belief (see Padmanabhan & Tuzhilin, 1998).

5.3.1 Paradox detection

A paradox by its very nature is surprising and the occurrence of Simpson's paradox has been identified by data mining researchers in the past but only to avoid generating spurious or invalid patterns (see Glymour *et al*., 1997). However, Fabris & Freitas (1999) have turned the notion of paradox detection into a goal for knowledge discovery. The effect of Simpson's paradox was noted in 1899 by Karl Pearson and has always been viewed as a problem. This is a phenomena whereby an event increases the probability in the super-population but also decreases the probability in the sub-populations comprising the data (see Pearl, 2000). The effects of the paradox can also be experienced in the opposite direction, i.e. the sub-populations can seemingly have the opposite effect to the super-population.

The authors describe the characteristics of Simpson's paradox, which occurred when determining the Tuberculosis mortality rates in New York and Richmond. Overall, the mortality rate of New York was lower than Richmond's but the opposite occurred when the populations were divided into racial groups of white and non-white. Partitioning the data meant that Richmond's mortality rate was now lower in both categories. Another example evoking the same surprise would be if we observed that the effects of a drug were beneficial to patients as whole, yet if we broke down the effects of the drug based on sex, we found that for both male and female patients found the drug was harmful. Pearl has placed the paradox in a probabilistic context:

$$P(E|C) > P(E|\neg C), \tag{3}$$

$$P(E|C, F) < P(E|\neg C, F), \tag{4}$$

$$P(E|C, \neg F) < P(E|\neg C, \neg F), \tag{5}$$

where C is the *cause* when taking the drug, E is the effect of recovery, F denotes a female patient and $\neg F$ denotes a male (Pearl, 2000). Therefore, Equation (3) identifies that the drug is beneficial to the overall population but Equation (4) indicates that the drug is harmful to females and Equation (5) indicates the drug is harmful to males.

Fabris & Freitas (1999) developed an algorithm that could detect Simpson's paradox and then calculate its magnitude, using the ranking as an indication of the degree of surprisingness. Fabris and Freitas point out that monothetic algorithms such as decision trees and rule induction programs particularly suffer from this effect. The authors conducted a series of experiments on several machine learning data sets and observed the occurrence of 13 instances of the paradox. This work is particularly useful because it deals with *surprising* patterns in a principled way.

5.3.2 Peculiarity rules and small support values

Zhong & Yao (2003) place a slightly different interpretation on the issue of unexpectedness by placing it in the context of *peculiarity* of the patterns. This involves pre-processing the data set to seek out those classes that are under-represented or have small support values. The authors compare their method against association and exception rules. The authors define the Peculiarity Factor or interestingness measure, $PF(x_{ij})$,

$$PF(x_{ij}) = \sum_{k=1}^{n} N(x_{ij}, x_{kj})^{a} \tag{6}$$

where N is the distance and a is a user-defined parameter.

Zhong and Yao identify the problem that unexpected but novel patterns may exist in rules with small support, i.e. a small number of tuples. It is difficult for standard association rules to identify such patterns, which rely on strong levels of support. For example, the authors provide the following rule that describes grocery store sales.

Example rule

```
meat.sales(low)∧vegetable.sales(low)∧fruit.sales(low)⇒
turnover(veryLow)
```

This particular rule covers information for one day's transactions, and would have been missed by traditional association rule mining due to the requirement to have the support and confidence levels set at particular levels. The criterion necessary for forming peculiarity rules is to seek patterns that are very different from other objects in the data set and the database should contain few instances of them. Zhong provides a fuzzy logic framework for the expression and interpretation of the peculiarity rules.

6 Interestingness measures for association rules

In this section we concentrate on interestingness measures for association rules, as they are so important and widespread within the data mining community. Association rules have received a great deal of attention since their development by Agrawal *et al.* (1993). They are an extremely important and widely used data mining technique that is computationally efficient. The idea for association rules originated in market basket analysis where purchases of grocery items are observed for their co-occurrence with particular products. The correlations between particular items can help the organization determine marketing strategies, sales promotion, stock planning and control. A survey comparing various algorithms and their computational complexity was conducted by Hipp *et al.* (2000). Agrawal and colleagues improved the computational efficiency of generating association rules (Agrawal & Srikant, 1995; Agrawal *et al.*, 1996). Research work continues, and Coenen & Leng (2001) have optimized the generation of association rules through the ordering of the itemsets that outperforms the Apriori algorithm. Related work by Han *et al.* (1993) optimized the search process through improved data structures such as the FPtree that avoids the generation of candidate sets, which are essential to efficient association rule creation. Inokuchi *et al.* (2003) have extended the Apriori algorithm to extract and search sub-graph structures efficiently, leading to substantial reductions in computation time.

Association rules are composed of implication expressions of the form $X \rightarrow Y$, where X and Y are itemsets; for example, $\{Fries\} \rightarrow \{Cola,\ Burgers\}$. There are several rule evaluation measures such as *Support*, which is the fraction of transactions that contain both X and Y. Another metric, the *Confidence*, measures how often items in Y appear in transactions that contain X. The *Expected Confidence* is simply the number of times the item of interest appears within the total number of transactions (e.g. Cola). The *Lift* factor is a ratio of *Confidence/Expected Confidence* and measures the overall rule strength.

Support indicates the significance of a rule; any rules with very low support values are uncommon, and probably represent outliers or very small numbers of transactions that are unlikely to be interesting or profitable. Confidence is important because it reflects how reliable is the prediction made by the rule. Generally, we can say that rules with high confidence values are more predominant in the total number of transactions. We can also say that confidence is an estimate of the conditional probability of a particular item appearing with another. For a general introduction to association rule interestingness, Brijs *et al.* (2003) is recommended. For example, Table 7 shows a fictitious set of market basket transactions.

Based on the frequencies of the products we can now derive support and confidence measures for our original rule,

$$\begin{aligned}
&Support = 5000/100\,000 = 5.0\% \\
&Confidence = 5000/30\,000 = 16.6\% \\
&Expected\ Confidence = 10\,000/100\,000 = 10\% \\
&Lift = 16.6/10.0 = 1.66
\end{aligned}$$

$$\{Fries\} \rightarrow \{Cola,\ Burgers(c16.6\%,\ s5.0\%)\}$$

which states that when customers buy fries, they also purchase cola and burgers. Ramaswamy *et al.* (1998) developed the objective concepts of *lift* to determine the importance of each association rule. The difficulty was that the *confidence* was not sufficient to determine the baseline frequency of the consequence. Liu *et al.* (2000, 2003) proposed subjective criteria to identify the most interesting rules from market basket data. The work of Liu *et al.* concentrates on modelling unexpectedness through an iterative process that gets the user to define certain aspects of their knowledge. To this end they have simplified the knowledge acquisition process by the creation of a specification language that can be applied to an analysis system that identifies rules that are conforming rules, unexpected conditions, unexpected consequences and rules that contain both types of unexpected knowledge. The user is allowed to specify their knowledge as loosely or precisely as they are able, thus enabling a certain amount of *vagueness* to be built into the system. The authors have defined the system to check for specific matches of the generated association rules against the predefined

Table 7 Example transaction data

Number	Transaction composition
100 000	Total transactions
40 000	Transactions contain Cola
30 000	Transactions contain Fries
20 000	Transactions contain Burgers
10 000	Transactions contain Burgers and Cola
5 000	Transactions contain Fries, Burgers and Cola

Table 8 Novel approaches to interestingness measures

Technique	Type	Automated
Via not interesting (Sahar, 1999)	Subjective	Domain expert required
Thresholds (Roddick & Rice, 2001)	Objective	Some expertise required
Paradox detection (Fabris & Freitas, 1999)	Hybrid	Some expertise required
Multi-criteria (Freitas, 1999)	Objective	No expert required
Bayesian techniques (Jaroszewicz & Simovici, 2004)	Hybrid	Domain expert required
dFr (Malone *et al*, 2004)	Objective	No expert required

knowledge of the user. If U is the rule set defining the knowledge of the user, A is the set of association rules generated from the data:

$$\begin{gathered}
confm_{ij} = L_{ij} * R_{ij} \\
unexpConseq_{ij} = \begin{cases} 0, & L_{ij} - R_{ij} \leq 0 \\ L_{ij} - R_{ij}, & L_{ij} - R_{ij} > 0 \end{cases} \\
unexpCond_{ij} = \begin{cases} 0, & R_{ij} - L_{ij} \leq 0 \\ R_{ij} - L_{ij}, & R_{ij} - L_{ij} > 0 \end{cases} \\
bsUnexp_{iju} = 1 - max(confm_{ij}, unexpConseq_{ij}, unexpCond_{ij})
\end{gathered} \tag{7}$$

where $confm_{ij}$ refers to confirmatory rules, $unexpConseq_{ij}$ refers to unexpected consequences, $unexpCond_{ij}$ refers to unexpected conditions, L_{ij} and R_{ij} are the degrees of condition and consequent match of rule A_i against U_j, and L_{ij}–R_{ij} calculates the unexpected consequent match since rules with high L_{ij} and low R_{ij} are ranked higher.

The search for interesting subsets of rules was investigated by Bayardo & Agrawal (1999). The authors developed an optimized algorithm for searching for the most interesting rules by integrating a variety of measures such as lift, support, confidence, entropy, laplace values, gini and chi-squared measures. The rule discovery process was interactive, which allowed the user to interrogate and compare the set of rules against the several measures. Partial ordering of the rules allowed the characteristics of specific subsets to be revealed, which otherwise would have been missed by using support and confidence levels alone.

Brin *et al.* (1997a,b) used the chi-squared measure to enable association rules to identify the significance of correlations rather than confidence and were thus able to generalize. Theoretically, this avoids the directional implication of confidence that treats the presence or absence of a value differently. The authors encountered scaling problems because the chi-squared measure is an approximation and requires certain conditions, i.e. variables must lie in certain ranges that association rules breach.

An extensive review of several measures was conducted by Tan *et al.* (2002) who examined 20 interestingness measures with regard to the internal data structures used by association rules and the computational and scaling effects observed. The authors made several conclusions regarding the key properties and applicability of the measures given the possible domain types encountered by association rules. Gray & Orlowska (1998) take the approach of clustering the attributes to association rules to tackle the so-called granularity problem. This is normally managed by assuming a taxonomy between the attributes, so that a general rule such as *Hardware⇒Software* would subsume *Keyboard⇒Software*. Without a taxonomy, there is always a danger that items at a finer granularity may be missed because of low support. However, the authors take a novel approach using a bottom-up, hierarchial clustering algorithm. The advantage is that there is no need to define a taxonomy that limits the number of expected associations. Since the authors only tackle a synthetic data set, the results indicate that the technique may not scale up for large numbers of variables.

The problems encountered by association rules when faced with temporal data were tackled by Ramaswamy *et al.* (1998), who devised a technique for identifying interesting association rules generated from time series data. The authors implement a *calendar algebra* capable of defining and manipulating time intervals, otherwise implementing the concept of time within association rules often entails a combinatorial explosion of the variables. It is possible to view the slicing and dicing of the data at different granularities of time and the authors note that such rules are more useful for detecting trends and patterns over time.

A method of mining interesting itemsets was identified by Mielikainen (2003) who computed the minimum frequency required that was independent of the data set used. Although, the set-based computations were highly expensive (calculating set intersections), they were able to identify values for the necessary parameters to create the *maximal* and *closed* set representations. The use of Bayesian networks to provide useful domain knowledge was investigated by Jaroszewicz & Simovici (2004) who used the networks to evaluate the interestingness of frequent itemsets. In this case the interestingness criterion was defined as the absolute difference between the support derived from the data and the calculated values from the Bayesian network. The method suggested by the authors requires a two-stage approach that initially finds all frequent itemsets using a minimum support level that is then compared against the frequent itemsets derived from the Bayesian network. A third stage involved a hierarchial and topological interestingness evaluation that enabled the expert knowledge held in the Bayesian network to be updated. Although the updating of the network was not the most sophisticated available, as it is possible to update both structure and parameters, the system appeared to be capable of updating the network in light of new discoveries.

An interestingness-based measure was applied by Wang *et al.* (1998) to address the problem of the combinatorial explosion encountered when defining intervals to enable association rules to manage numeric data. The authors considered the *J-measure* of Smyth & Goodman (1992) and a measure based on the minimum confidence. The *J-measure* is particularly suited to discriminating between the *prior* probability and the *posterior* probability of an event. The most useful rules are therefore those with a high degree of dissimilarity; each interval has its own interestingness ranking that is modified after merging. The work by the authors is notable as most association rules are intended to operate on categorical data rather than numerical data.

In the next two sections we discuss variations on the association rule theme.

6.1 Ratio rules

Since their inception, association rules have spawned several variations such as the ratio rules developed by Korn *et al.* (2000). Ratio rules are a technique that employs eigensystem analysis to calculate correlations between values of attributes; the eigensystem analysis reveals the axes of greatest variation and thus the most important correlations. These are presented as a ratio rule.

Example rule:

```
Customers typically spend 10:30:50 dollars on shirt:pants:shoes
```

The rule states that:
if the customer spends $10 on a shirt, then the customer is likely to spend $30 on pants and $50 on shoes.

The interestingness measure proposed by Korn to assess the quality of the ratio rules is described as a 'guessing error' (GE). The GE refers to the estimation of calculating missing values in the data matrix; the values are in fact known but are left out deliberatively to validate the rule generation process. The GE can be defined formally as

$$GE = \sqrt{\frac{1}{NM}\sum_{i}^{N}\sum_{j}^{M}(x_{ij} - x_{ij})^2} \tag{8}$$

where N is the number of customers, M is the number of products, $\hat{x}_{ij}$ is the estimated 'missing' value and x_{ij} is the actual value.

A ratio rule with a low value for GE implies that it has identified a novel pattern in the data with high confidence. Ratio rules address the issue of reconstructing missing/hidden values as well as being able to perform *what-if* type scenarios, where an item and corresponding value is required for a given set of antecedents. The technique appears to scale up quite well when faced with large data sets and is useful for predicting attribute trends, given empirical data. However, the process does not incorporate either spatial or temporal elements and would be unsuitable for the analysis of such data.

6.2 *Differential ratio rules*

Korns ratio rules were developed further by Malone *et al.* (2004) by adding a temporal element to them in the form of *differential ratio rules* (dFr) capable of detecting interesting patterns in spatio-temporal data. The dFr technique incorporates all aspects of the spatio-temporal patterns and was used to detect changes in digitized images of two-dimensional protein gels within a time series. Specific proteins were flagged as interesting and ranked according to the amount they had altered. Particular alterations include morphological variations, the absence/presence of proteins over time and spatial variations.

Differential ratio data mining is used to measure variation of a given object in terms of the pairwise ratios of the elements describing the data over time. Consider two variables x and y as elements of a given object. Calculation of a single differential ratio (herein, differential ratio, or *dFr*, will be referred to as the measure of difference calculated by this process) between two time points t and $t+1$ is given by

$$\log\left\{\frac{(x_t/y_t)}{(x_{t+1}/y_{i+1})}\right\} = dFr_t \tag{9}$$

where $x \leq y$. When this is not the case, that is $y<x$, the variables are inverted to ensure the measures remain consistent. Since it is the magnitude of difference in ratios we are looking for, i.e. how they increase or decrease together, we are not concerned with maintaining the two variables juxtaposition as numerator and denominator. When considering the instance of $y<x$, then the following is used:

$$\log\left\{\frac{(y_t/x_t)}{(y_{t+1}/x_{i+1})}\right\} = dFr_t \tag{10}$$

Such a calculation would be performed for a time series ($t=1, \ldots, n$) and for all pairs of variables that were deemed as representative of the dataset. For a single pair of variables, this describes the variance that occurs over time for a given object. This can be summarized for a series of differential ratios (*dFr*) for a given time series for variables x and y in the form

$$Object{:}x,\ y[dFr_t,\ \mathrm{dFr}_{t+1}, \ldots, dFr_{t+n}].$$

Using the definition of *interestingness*, for each dFr_t extracted, the following can be said about the ratio between variable x and y over time point t and $t+1$:

- $dFr_t \sim 0$, ratio has remained constant;
- $dFr_t < 0$, ratio of difference has decreased over time;
- $dFr_t > 0$, ratio of difference has increased over time.

That is, a positive *dFr* value indicates that the two variables values are growing further apart in terms of the two ratios over time. A negative value is the opposite of this, i.e. the two variables values are becoming closer together in terms of the two ratios over time. A value of around zero indicates that the ratios between the variables has barely altered over time; exactly zero means no difference at all. The magnitude of the measure also has a proportional meaning since the greater the value the more change has occurred. For instance, a larger positive dFr_t value means a larger difference in ratios over time compared with a smaller value.

7 Novel areas for future research

Objective measures are generally chosen for their properties of similarity/dissimilarity, diversity, dispersion and inequality. Although there is no ‘correct’ or ‘wrong’ measure to use, a careful analysis of the problem domain will reveal the key characteristics to be tested for deviations from the norm. Many techniques make use of the chi-squared test, which can reveal the level of independence of the attributes. This approach can often reduce the number of patterns to be considered by the user. However, the additional factors identified by Freitas (1999) should be taken into account (disjunct size, attribute interestingness, misclassification costs and class imbalance) when analyzing ‘interesting’ patterns.

The challenge is to address the subjective/objective dichotomy by techniques that can reduce the knowledge acquisition bottleneck. The approach taken by Sahar should be further explored; although it does not entirely remove the expert, it is perhaps the most promising of the subjective approaches. The evaluation of ‘hybrid’ interestingness measures such as paradox detection should yield good results and should be investigated further (see Fabris & Freitas, 1999). Paradox detection lies midway between the subjective and objective measures; it relies on the notion of *surprise* to discover interesting patterns. Goal-directed versus data-directed analysis is intimately related to the use of subjective and objective measures.

Detecting change over time is a particularly difficult task, especially if other factors such as spatial or other characteristics are present (Roddick & Spilioppoulou, 2001; Dykes & Mountain, 2003). It may also be argued that interestingness is a temporal phenomena anyway, since many patterns are compared with what has happened in the past. If prior knowledge of an event or pattern is available then it is possible to incorporate these beliefs directly into the algorithms to guide the knowledge discovery process. Even a limited amount of knowledge in the form of simple thresholds can provide sufficient background information for a goal-driven approach to discovery.

Bayesian belief networks (BN) are becoming a popular choice in many data mining application areas, especially bioinformatics. The BN model is particularly well suited for representing and managing the difficulties involved with missing, noisy and uncertain data. These networks can also incorporate subjective human beliefs that can be later refined and updated in the light of new evidence, which are aspects that are difficult to formalize using a traditional frequentist approach to probabilities.

8 Conclusion

This paper has reviewed and evaluated the current research literature on the various techniques for determining the interestingness of patterns discovered by the data mining process. The characteristics of several goal- and data-driven measures have been discussed. The main disadvantage of the subjective or user-driven approach is that it constrains the discovery process to seek only what the user can anticipate or hypothesize, i.e. it cannot discover unexpected or unforeseen patterns because it is entirely goal driven. Alternatively, objective measures or data-driven measures tend to concentrate on finding patterns through statistical strength or correlations. Such patterns may not be interesting to the user as a strong pattern may imply knowledge that is already well known to the organization. We have discussed the issues regarding the degree of user involvement within the knowledge discovery process. The planned level of autonomy will have a critical impact on how the interestingness measures are applied and therefore the overall efficiency and usefulness of the discovery system. A major research question is how to combine both objective and subjective measures into a unified measure. The unification is necessary as developing subjective measures is quite expensive because of the costs associated with the knowledge acquisition process. It is likely that ontologies and other semantic technologies will play an increasingly important role in bridging this gap, by providing the level of detail required to develop robust belief systems. Without a sufficiently rich belief system it is difficult for subjective measures to accurately model the user domain. Recent work in paradox detection has seen the development of interestingness measures that are part way between objective and subjective measures, and perhaps they may provide the foundations for general purpose automated discovery systems.

Acknowledgement

The author is grateful to the two reviewers who provided many useful comments for improving this paper.

References

Agrawal, R, Imielinski, T and Swami, A, 1993, Mining association rules between sets of items in large databases. In *SIGMOD-93*, pp. 207–216.

Agrawal, R, Mannila, H, Srikant, R, Toivonen, H and Verkamo, A, 1996, Fast discovery of association rules. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P and Uthursamy, R. (eds.), *Advances in Knowledge Discovery and Data Mining*. AAAI Press.

Agrawal, R and Srikant, R, 1995, Mining sequential patterns. In *Proceedings of the International Conference on Data Engineering (ICDE), Taipei, Taiwan.*

Anand, S., Patrick, A., Hughes, J and Bell, D, 1998, A, datamining methodology for cross-sales. *Knowledge-based Systems* **10**, 449–461.

Atkinson, A, 1970, On the measurement of inequality. *Journal of Economic Theory* **2**, 244–263.

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

Berger, G and Tuzhilin, A, 1998, Discovering unexpected patterns in temporal data using temporal logic. In Etzon, O, Jajodia, S and Sripada. S (eds.), *Temporal Databases Research and Practice (Lecture Notes in Computer Science, 1399)*, pp. 281–309.

Brachman, R, 1993, Integrated support for data archaeology. *International Journal of Intelligent and Cooperative Information Systems* **2**(2), 159–185.

Brijs, T., Vanhoof, K and Wets, G, 2003, Defining interestingness for association rules. *International Journal of Information Theories and Applications* **10**(4), 370–376.

Brin, S, Motwani, R and Silverstein, C, 1997a, Beyond market baskets: generalizing association rules to correlations. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 265–276.

Brin, S, Motwani, R, Ullman, J and Tsur, S, 1997b, Dynamic itemset counting and implication rules for market basket data. In J Peckham, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13–15, Tucson, AZ*, pp. 255–264.

Buntine, W, 1996, Graphical models for discovering knowledge. In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthursamy, R (eds.), *Advances in Knowledge Discovery and Data Mining*. AAAI Press, pp. 59–82.

Chen, M, Han, J and Yu, P, 1996, Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering* **8**(6), 866–883.

Coenen, F and Leng, P, 2001, Optimising association rule algorithms using itemset ordering. In *Proceedings of the 21st Annual International Conference of the British Computer Society Specialist Group on Expert Systems*. Cambridge: Springer, pp. 53–66.

Colton, S and Bundy, A, 2000, On the notion of interestingness in automated mathematical discovery. *International Journal of Human-Computer Studies* **53**, 351–375.

Correia, J, Stumme, G, Wille, R. &Wille, U, 2003, Conceptual knowledge discovery—a human centered approach. *Applied Artificial Intelligence* **17**(3), 281–302.

DiFore, F, 2000, Visualizing interestingness. In *Proceedings of the 3rd International Conference on Data Mining*, pp. 147–156.

Dykes, J and Mountain, D, 2003, Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and applications. *Computational Statistics and Data Analysis* **43**, 581–603.

Fabris, C and Freitas, A, 1999, Discovering surprising patterns by detecting occurrences of simpson's paradox. In *Development in Intelligent Systems XVI (Proc. Expert Systems 99, The 19th SGES International Conference on Knowledge-based Systems and Applied Artificial Intelligence), Cambridge, UK*, pp. 148–160.

Fayyad, U, Piatetsky-Shapiro, G and Smyth, P, 1996a, From data mining to knowledge discovery: an overview. In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthursamy, R (eds.), *Advances in Knowledge Discovery and Data Mining*. AAAI Press, pp. 1–34.

Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthursamy, R (eds.) 1996b, *Knowledge Discovery in Databases*. AAAI Press/MIT Press.

Fayyad, U and Stolorz, P, 1997, Data mining and KDD: promise and challenges. *Future Generation Computer Systems* **13**(2–3), 99–115.

Frawley, WJ, Piatetsky-Shapiro, G and Maltheus, CJ, 1991, *Knowledge Discovery in Databases: An Overview*. AAAI Press/MIT Press.

Freitas, A, 1998, On objective measures of rule surprisingness. In *Principles of Data Mining and Knowledge Discovery: Proceedings of the 2nd European Symposium, Nantes, France (Lecture Notes in Artificial Intelligence, 1510)*, pp. 1–9.

Freitas, A, 1999, On rule interestingness measures. *Knowledge-based Systems* **12**(5–6), 309–315.

Gago, P and Bento, C, A metric for selection of the most promising rules. In *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-98), Nantes, France*, pp. 19–27.

Gaines, B, 1996, Transforming rules and trees. In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthursamy, R (eds.), *Advances in Knowledge Discovery and Data Mining*. AAAI Press, pp. 205–226.

Glymour, C, Madigan, D, Pregibon, D and Smyth, P, 1997, Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery* **1**(1), 11–28.

Gray, B and Orlowska, M, 1998, CCAIIA: clustering categorical attributes into interesting association rules. In *Proceedings of the 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 132–143.

Han, J, Cai, Y and Cercone, N, 1993, Data driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Engineering* **5**(1), 29–40.

Hand, D, 1999, Statistics and data mining: intersecting disciplines. *ACM SIGKDD Explorations* **1**(1), 16–19.

Hand, D, Mannila, H and Smyth, P, 2001, *Principles of Data Mining*. MIT Press.

Hilario, M and Kalousis, A, 2000, Building algorithm profiles for prior model selection in knowledge discovery systems. *Engineering Intelligent Systems* **8**(2), 77–87.

Hilderman, R and Hamilton, H, 1999, Heuristic measures of interestingness. In *Proceedings of the 3rd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'99), Prague, Czech Republic (Lecture Notes in Computer Science)*, pp. 232–241.

Hilderman, R and Hamilton, H, 2001, *Knowledge Discovery and Measures of Interest*. Kluwer Academic Publishers.

Hilderman, R and Hamilton, H, 2003, Measuring the interestingness of discovered knowledge: a principled approach. *Intelligent Data Analysis* **7**(4), 347–382.

Hipp, J, Guntzer, U and Nakhaeizadeh, G, 2000, Algorithms for association rule mining—a general comparison. *SIGKDD Explorations* **2**, 58–64.

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*, pp. 86–97.

Ikizler, N and Gvenir, HA, 2001, Mining interesting rules in bank loans data. In Acan, A, Aybay, I and Salamah, M (eds.), *Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN 2001)*, pp. 238–246.

Inokuchi, A, Washio, T and Motoda, H, 2003, Complete mining of frequent patterns from graphs: mining graph data. *Machine Learning* **50**, 321–354.

Jaroszewicz, S and Simovici, DA, 2004, Interestingness of frequent itemsets using Bayesian networks as background knowledge. In *Proceedings of the 2004, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA*, pp. 178–186.

Kamber, M and Shinghal, R, 1996, Evaluating the interestingness of characteristic rules. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 263–266.

Klemettinen, M, Mannila, H, Ronkainen, P, Toivonen, H and Verkamo, A, 1994, Finding interesting rules from large sets of discovered association rules. In Adam, N, Bhargava, B and Yesha, Y (eds.), *Proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM'94)*. ACM Press, pp. 401–407.

Klösgen, W, 1995, Efficient discovery of interesting statements in databases. *Journal of Intelligent Information Systems* **4**, 53–69.

Korn, F, Labrinidis, A, Kotidis, Y and Faloutsos, C, 2000, Quantifiable data mining using ratio rules. *The VLDB Journal* **8**, 254–266.

Kulkarni, D and Simon, H, 1988, The process of scientific discovery: the strategy of experimentation. *Cognitive Science* **12**, 139–175.

Kullback, S and Leiber, R, 1951, On information and sufficiency. *Ann. Math. Stat.* **4**, 99–111.

Langely, P, Zytkow, J, Simon, HA and Bradshaw, GL, 1983, The search for regularity: four aspects of scientific discovery. In Michalski, RS and Carbonell, J (eds.), *Machine Learning: an AI Approach*. Tioga Publishing Company.

Lavington, S, Dewhirst, N and Freitas, A, 1999, Interfacing knowledge discovery algorithms to large database management systems. *Information and Software Technology* **41**, 605–617.

Lenat, D, 1983, The role of heuristics in learning by discovery: three case studies. In Michalski, RS and Carbonell, J (eds.), *Machine Learning: an AI Approach*. Tioga Publishing Company.

Lenat, D and Brown, J, 1984, Why AM and Eurisko appear to work. *Artificial Intelligence* **23**, 269–294.

Liu, B, Hsu, W, Chen, S and Ma, Y, 2000, Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems* **15**(5), 47–55.

Liu, B, Hsu, W, Mun, L and Lee, HY, 1999, Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering* **11**(6), 817–832.

Liu, B, Ma, Y and Wong, C, 2003, Scoring the data using association rules. *Applied Intelligence* **18**, 119–135.

Livingstone, G, Rosenberg, J and Buchanan, B, 2003, An agenda and justification based framework for discovery systems. *Knowledge and Information Systems* **5**(2), 133–161.

Lorenz, M, 1905, Methods of measuring the concentration of wealth. *Quart. Publ. American Statist. Assoc.* **9**, 209–219.

Ludwig, J and Livingstone, G, 2000, What's new? Using prior models as a measure of novelty in knowledge discovery. In *Proceedings of the 12th IEEE Conference on Tools with Artificial Intelligence*, pp. 86–89.

Major, J and Mangano, J, 1995, Selecting among rules induced from a hurricane database. *Journal of Intelligent Information Systems* **4**(1), 39–52.

Malone, J, McGarry, K and Bowerman, C, 2004, Performing trend analysis on spatio-temporal proteomics data using differential ratio rules. In *Proceedings of the 6th EPSRC Conference on Postgraduate Research in Electronics, Photonics, Communications and Software, University of Hertfordshire, UK*, pp. 103–105.

Matheus, C, Chan, P and Piatetsky-Shapiro, G, 1993, Systems for knowledge discovery in databases. *IEEE Transactions on Knowledge and Data Engineering* **5**(6), 903–913.

McGarry, K and Malone, J, 2004, Analysis of rules discovered by the data mining process. In Lofti, A and Garibaldi, J (eds.), *Applications and Science in Soft Computing Series: Advances in Soft Computing*. Springer, pp. 219–224.

McGarry, K, Wermter, S and MacIntyre, J, 2001a, The extraction and comparison of knowledge from local function networks. *International Journal of Computational Intelligence and Applications* **1**(4), 369–382.

McGarry, K, Wermter, S and MacIntyre, J, 2001b, Knowledge extraction from local function networks. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence, 2, Seattle, WA*, pp. 765–770.

Merzbacher, M, 2002, *Automatic Generation of Trivia Questions (Lecture Notes in Computer Science, 2366)*, pp. 123–130.

Mielikainen, T, 2003, Finding all occurring sets of interest. In Boulicaut, J and Dzeroski, S (eds.), *Proceedings of the 2nd International Workshop on Knowledge Discovery in Inductive Databases*, pp. 97–106.

Padmanabhan, B and Tuzhilin, A, 1998, A belief driven method for discovering unexpected patterns. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pp. 94–100.

Padmanabhan, B and Tuzhilin, A, 1999, Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems* **27**, 303–318.

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

Pazzani, M, 2000, Knowledge discovery from data? *IEEE Intelligent Systems* **15**(2), 10–13.

Pearl, J, 2000, *Causality*. Cambridge: Cambridge University Press.

Piatetsky-Shapiro, G and Frawley, WJ (eds.) 1991, *Knowledge Discovery in Databases*. AAAI Press/MIT Press.

Piatetsky-Shapiro, G and Matheus, CJ, 1994, The interestingness of deviations. In *Proceedings of AAAI Workshop on Knowledge Discovery in Databases*.

Quinlan, J. R, 1992 *C4.5: Programs for Machine Learning*. San Francisco: Kaufmann.

Ramaswamy, S, Mahajan, S and Silberschatz, A, 1998, On the discovery of interesting patterns in association rules. In *Proceedings of the 24th International Conference on Very Large Data Bases*. San Francisco: Kaufmann, pp. 368–379.

Roddick, J and Rice, S, 2001, What's interesting about cricket?—On thresholds and anticipation in discovered rules. *SIGKDD Explorations* **3**(1), 1–5.

Roddick, J and Spilioppoulou, M, 2001a, A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering* **14**(4), 750–767.

Romao, W, Freitas, A and Gimenes, I, 2004, Discovering interesting knowledge from a science and technology database with a genetic algorithm. *Applied Soft Computing* **4**(2), 121–137.

Roussinov, D and Zhao, J, 2003, Automatic discovery of similarity relationships through web mining. *Decision Support Systems* **35**, 149–166.

Sahar, S, 1999, Interestingness via what is not interesting. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, San Diego, CA*, pp. 332–336.

Sahar, S, 2001, Interestingness preprocessing. In *Proceedings of the 2001, IEEE International Conference on Data Mining, San Jose, CA*, pp. 489–496.

Sahar, S, 2002, On incorporating subjective interestingness into the mining process. In *Proceedings of the 2002, IEEE International Conference on Data Mining (ICDM 2002), Maebashi City, Japan*, pp. 681–684.

Shannon, C, 1948, A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423.

Silberschatz, A and Tuzhilin, A, 1995, On subjective measures of interestingness in knowledge discovery. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, pp. 275–281.

Silberschatz, A and Tuzhilin, A, 1996a, User-assisted knowledge discovery: how much should the user be involved. In *ACM-SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*.

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

Smyth, P and Goodman, R, 1992, An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering* **4**(4), 301–316.

St Amant, R and Cohen, P, 1998, Intelligent support for exploratory data analysis. *Journal of Computational and Graphical Statistics* **7**(4), 545–558.

Tan, P, Kumar, V and Srivastava, J, 2002, Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, pp. 32–41.

Wang, K, Tay, S and Liu, B, 1998, Interestingness-based interval merger for numeric association rules. In Agrawal, R, Stolorz, PE and Piatetsky-Shapiro, G (eds.), *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, KDD*. AAAI Press, pp. 121–128.

Zhong, N, Liu, C and Ohsuga, S, 2001, Dynamically organizing KDD processes. *International Journal of Pattern Recognition and Artificial Intelligence* **15**(3), 451–473.

Zhong, N and Yao, Y, 2003, Peculiarity oriented multidatabase mining. *IEE Transactions on Knowledge and Data Engineering* **15**(4), 952–960.

Zytkow, J, 1993, Introduction: cognitive autonomy in machine learning. *Machine Learning* **12**, 7–16.