Knowledge Discovery and Interestingness Measures: A Survey

Robert J. Hilderman and Howard J. Hamilton

Department of Computer Science, University of Regina

Regina, Saskatchewan, Canada S4S 0A2

E-mail: {hilder,hamilton}@cs.uregina.ca

Abstract

Knowledge discovery in databases, also known as data mining, is the efficient discovery of previously unknown, valid, novel, potentially useful, and understandable patterns in large databases. It encompasses many different techniques and algorithms which differ in the kinds of data that can be analyzed and the form of knowledge representation used to convey the discovered knowledge. An important problem in the area of data mining is the development of effective measures of interestingness for ranking the discovered knowledge. In this report, we provide a general overview of the more successful and widely known data mining techniques and algorithms, and survey seventeen interestingness measures from the literature that have been successfully employed in data mining applications.

1 Introduction

Knowledge discovery in databases, also known as data mining, is the efficient discovery of previously unknown, valid, novel, potentially useful, and understandable patterns in large databases [19, 15]. Ultimately, the knowledge that we seek to discover describes patterns in the data as opposed to knowledge about the data itself. Patterns in the data can be represented in many different forms, including classification rules, association rules, clusters, sequential patterns, time series, contingency tables, summaries obtained using some hierarchical or taxonomic structure, and others. Typically, the number of patterns generated is very large, but only a few of these patterns are likely to be of any interest to the domain expert analyzing the data. The reason for this is that many of the patterns are either irrelevant or obvious, and do not provide new knowledge [49]. To increase the utility, relevance, and usefulness of the discovered patterns, techniques are required to reduce the number of patterns that need to be considered. Techniques which satisfy this goal are broadly referred to as *interestingness measures*.

This report is organized as follows. In Section 2, we present a general overview of classical data mining techniques and algorithms. In Section 3, we present a survey of seventeen interestingness measures that have been successfully employed in data mining applications. We conclude in Section 4 with a summary table classifying the seventeen interestingness measures described in Section 3.

2 Data Mining Techniques and Algorithms

Data mining encompasses many different techniques and algorithms. These differ in the kinds of data that can be analyzed and the kinds of knowledge representation used to convey the discovered knowledge. Here we describe some of the more successful and widely known techniques and algorithms.

2.1 Classification

Classification is perhaps the most commonly applied data mining technique. Early examples of classification techniques from the literature include Mitchell's VSA [45, 46], Quinlan's ID3 [53], Michalski et al. AQ15 [43], and Clark and Niblett's CN2 [11]. VSA induces a single classification rule from two complementary

trees (a specialization tree and a generalization tree) that converge on a common node containing the rule. ID3 induces a decision tree. An object is classified by descending the tree until a branch leads to a leaf node containing the decision. AQ15 induces a set of decision rules. An object is classified by selecting the most preferred decision rule according to user-defined criteria. CN2 induces a decision list. An object is classified by selecting the best rule according to user-defined accuracy and statistical significance criteria.

Later examples of classification techniques from the literature include Zhang and Michalski's FCLS [71], Gur-Ali and Wallace's PrIL [24], Mehta et al. SLIQ [42], and CLOUDS [7]. FCLS induces a weighted threshold rule. The threshold determines the number of conditions which must be satisfied in a valid rule. An object is classified by generalizing and specializing examples until the number of incorrectly classified examples is below some user-defined error rate. PrIL induces decision rules in a manner similar to those induced by ID3. However, the rules induced by PrIL are associated with a minimum correct classification threshold and confidence level. When a rule cannot meet the minimum correct classification threshold, objects cannot be classified according to that rule. SLIQ induces a decision tree built using the Minimum Description Length principle [57]. It is similar to other decision tree classifiers except that it is capable of handling large disk-resident datasets (i.e., all of the data cannot fit into memory). CLOUDS induces a decision tree similar to the manner used by SLIQ except that a more computationally efficient method is used to determine the splitting points at each node.

Other examples of classification techniques from the literature include C4.5/C5.0 [52], KID3 [51], parallel ID3 [17], and SPRINT [60]. C4.5/C5.0 is an industrial-quality descendant of ID3 that has seen widespread use in the research community. KID3 induces exact decision rules (i.e., those that are always correct) and strong decision rules (i.e., those that are almost always correct). An efficient parallel technique is used that accesses the data only once to generate all exact rules. Parallel ID3 uses a distributed tree construction technique to induce decision trees. SPRINT is a parallel version of the SLIQ algorithm that uses different and more memory efficient data structures to induce a decision tree.

2.2 Association

Association is another of the commonly applied data mining techniques. The problem is typically examined in the context of discovering buying patterns from retail sales transactions, and is commonly referred to as market basket analysis. Market basket analysis was originally introduced in [2] and has since been studied extensively.

Much of the literature focuses on the Apriori algorithm [3] and its descendants containing various refinements. Apriori extracts the set of frequent itemsets from the set of candidate itemsets generated. A frequent itemset is an itemset whose support is greater then some user-defined minimum and a candidate itemset is an itemset whose support has yet to be determined.It has an important property that if any subset of a candidate itemset is not a frequent itemset, then the candidate itemset is also not a frequent itemset.

Refinements to Apriori include Partition [59], DHP [50], sampling [65], DIC [9], and parallel Apriori [4]. Partition reads the database at most two times to generate all significant association rules while generating no false negatives. It is also inherently parallel in nature and can be parallelized with minimal communication and synchronization between nodes. DHP is a hash-based algorithm for generating candidate itemsets that reduces the number of candidate 2-itemsets by an order of magnitude. Pruning the candidate 2-itemsets significantly reduces the number of frequent k-itemsets that need to be considered when $k > 2$. Sampling is used to take a random sample from a database to find all association rules that are probably valid in the entire database. A second pass of the database is used to verify the support for each potential association rule. DIC partitions the database into blocks. When scanning the first block it counts only 1-itemsets. When scanning the k-th block, it counts 2-, 3-, 4-, ..., k-itemsets. Usually it can finish counting all the itemsets in two passes over the data. Parallel Apriori is a parallel version of Apriori that exhibits excellent scaleup behaviour and requires only minimal additional overhead compared to serial Apriori.

Other literature focuses on alternative approaches for discovery of association rules. These approaches include Hybrid Distribution [28], Itemset Clustering [69], Share Measures [30], and Q2 [10]. Hybrid Distribution is a parallel algorithm that improves upon parallel Apriori by dynamically partitioning the candidate itemsets to achieve superior load balancing across the nodes. More association rules can then be generated more quickly in a single pass over the database. Scaleup is near linear and in general it utilizes memory

more efficiently. Itemset Clustering approximates the set of potentially maximal frequent itemsets and then uses an efficient lattice traversal technique to generate clusters of frequent itemsets in a single pass over the database. Share measures are used to more accurately indicate the financial impact of an itemset by not only considering the co-occurrence of items in an itemset, but by also considering the quantity and value of the items purchased. Q2 obtains performance improvements of more than an order of magnitude over Apriori by computing and pruning the frequent Boolean itemsets before searching for valid association rules. Once this is done, association rules can be found with a single pass over the database.

2.3 Clustering

Identifying objects that share some distinguishing characteristics is also a frequently used data mining technique. Known as clustering, there are numerous techniques described in the literature.

Early examples of clustering from the literature include Michalski's CLUSTER/2 [44] and Fisher's COBWEB [18]. CLUSTER/2 finds a disjoint clustering of objects that optimizes user-defined parameters regarding the number of clusters required and clustering quality criteria. It uses an efficient path-rank-ordered search procedure to limit the number of nodes visited in the search tree. COBWEB further increases efficiency by using an incremental approach that organizes data in a way that maximizes its inference abilities by identifying data dependencies involving important attributes.

More recent examples from the literature include Ng and Han's CLARANS [47], Zhang et al. BIRCH [72], Ester et al. DBSCAN [14], Wang et al. STING [66], and Agrawal et al. CLIQUE [1]. CLARANS is an extension of the k-medoids approach developed by Kaufman and Rousseeuw in [32]. It is based upon a randomized search algorithm with user-defined parameters that control the length and quality of the search. BIRCH incrementally and dynamically evaluates data to generate the best quality clusters possible given user-defined time constraints and available memory. A single pass over the database is usually enough to find high quality clusters. DBSCAN is a density-based approach that utilizes user-defined parameters for controlling the density of the discovered clusters. This approach allows adjacent regions of sufficiently high density to be connected to form clusters of arbitrary shape and is able to differentiate noise in regions of low density. STING models the search space as a hierarchical structure of rectangular cells corresponding to

different levels of resolution. Each cell at a high level is partitioned to form a number of smaller cells in the next lower level. Statistical information is associated with each cell to facilitate querying and incremental updates. CLIQUE is a density-based approach that has the ability to find clusters in subspaces of high dimensional data. The search space is partitioned into equal-sized units. Discovered clusters are unions of adjacent high-density units.

2.4 Correlation

Statistically-oriented in nature, correlation has seen increasing use as a data mining technique. Although the analysis of multi-dimensional categorical data is possible and described extensively in the literature [56, 16, 22], the most commonly employed method is that of two-dimensional contingency table analysis of categorical data using the chi-square statistic as a measure of significance.

Recent examples from the literature include the work of Sanjeev and Zytkow [58], Knobbe and Adrian [35], Zemobowicz and Zytkow [70], Brin et al. [8], and Liu et al. [37]. Sanjeev and Zytkow analyze contingency tables to discover students who are poorly prepared for university level course and at risk of dropping out. Knobbe and Adrian analyze contingency tables to discover simple associations between single attributes that can be easily visualized in a bar graph. Zembowicz and Zytkow analyze contingency tables using the 49er data mining system. 49er examines each pair of attributes in a contingency table and applies statistical tests of significance and strength to quantify the discovered patterns. Brin et al. analyze contingency tables to generate dependence rules that identify statistical dependence in both the presence and absence of items in itemsets. Liu et al. analyze contingency tables to discover unexpected and interesting patterns that have a low level of support and a high level of confidence.

2.5 Other Techniques

Other data mining techniques search for patterns in sequences and time series. The problem of mining for patterns in sequences was introduced in [6] and [63]. The search for sequences of events that occur in a particular order and within a particular time interval is described in [40] and [39]. A logic for expressing temporal patterns defined over categorical data as a means for discovering patterns in sequences is described

in [48]. Recent approaches for the discovery of patterns in sequences are described in [67], [25], and [68].

The problem of mining for patterns in time series has received a considerable amount of attention recently. An approach that queries the Fourier series representation of a sequence is described in [54]. A framework is described in [55] where models containing high-dimensional time series data are learned so that data values can be forecast for the immediate future. An extended representation of time series that allows accurate classification and clustering through a relevance feedback mechanism is described in [33]. A method for mining segment-wise periodicity in time series data is described in [29]. The problem of finding rules relating patterns in a time series to other patterns in the same time series, or to patterns in another time series are described in [12].

3 Interestingness Measures

One problem area in the field of knowledge discovery in databases is the development of interestingness measures for ranking the usefulness and utility of discovered patterns. In this section, we survey and describe interestingness measures proposed in the literature that have been successfully employed in data mining applications.

3.1 Piatetsky-Shapiro's Rule-Interest Function

The *rule-interest function* [51] is used to quantify the correlation between attributes in a simple classification rule. A *simple classification rule* is one where the left- and right-hand sides of the logical implication $X \rightarrow Y$ correspond to a single attribute. The rule-interest function is given by:

$$RI = |X \cap Y| - \frac{|X||Y|}{N},$$

where N is the total number of tuples, $|X|$ and $|Y|$ are the number of tuples satisfying conditions X and Y, respectively, $|X \cap Y|$ is the number of tuples satisfying $X \rightarrow Y$, and $|X||Y|/N$ is the number of tuples expected if X and Y were independent (i.e., not associated).

When $RI = 0$, then X and Y are statistically independent and the rule is not interesting. When $RI > 0$

$(RI < 0)$, then X is positively (negatively) correlated to Y. The significance of the correlation between X and Y can be determined using the chi-square test for a 2×2 contingency table. Those rules which do not exceed a pre-determined minimum significance threshold are determined to be the most interesting.

3.2 Smyth and Goodman's J-Measure

The *J-measure* [62] is the average information content of a probabilistic classification rule and is used to find the best rules relating discrete-valued attributes. A *probabilistic classification rule* is a logical implication $X \rightarrow Y$ with some probability p, where the left- and right-hand sides correspond to a single attribute. The right-hand side is restricted to simple single-valued assignment expressions, while the left-hand side may be a conjunction of these simple expressions. The J-measure is given by:

$$J(x;y) = p(y)\left[p(x|y)\log\left(\frac{p(x|y)}{p(x)}\right) + (1-p(x|y))\log\left(\frac{1-p(x|y))}{(1-p(x))}\right)\right],$$

where $p(y)$, $p(x)$, and $p(x|y)$ are the probabilities of occurrence of y, x, and x given y, respectively, and the term inside the square brackets is the relative (or cross) entropy. Relative entropy is the similarity (or goodness of fit) of two probability distributions.

High values for $J(x;y)$ are desirable, but are not necessarily associated with the best rule. For example, rare conditions may be associated with the highest values for $J(x;y)$ (i.e., where a particular y is highly unlikely), but the resulting rule is insufficiently general to provide any new information. Consequently, analysis may be required in which the accuracy of a rule is traded for some level of generality or goodness-of-fit.

3.3 Major and Mangano's Rule Refinement

Rule refinement [38] is a strategy used to induce interesting classification rules from a database of classification rules. The strategy consists of three phases: identifying potentially interesting rules, identifying technically interesting rules, and removing rules that are not genuinely interesting. *Potentially interesting rules* are those that satisfy specified confidence, coverage, and simplicity (i.e., rule length) criteria, or are closely related to rules that do (closely related rules are those that may be specializations/generalizations of rules in the

rule lattice). *Technically interesting rules* are selected from the potentially interesting rules according to simplicity and statistical significance (i.e., chi-square test) criteria. While the selection of potentially and technically interesting rules is strictly algorithmic, removing rules that are not genuinely interesting is a manual task performed by a domain expert. This task involves keeping the simplest and/or most general rules that adequately describe the data and removing other similar rules.

3.4 Agrawal and Srikant's Itemset Measures

The *itemset measures* [2, 5] are used to identify frequently occurring association rules from sets of items in large databases. An association rule is a logical implication $X \rightarrow Y$, where the left- and right-hand sides correspond to a set of attributes, and X and Y are disjoint. The association rule $X \rightarrow Y$ holds in a transaction set D with *confidence* c, if $c\%$ of the transactions in D that contain X, also contain Y. The association rule $X \rightarrow Y$ has *support* s in transaction set D, if $s\%$ of the transactions in D contain $X \cup Y$. From these definitions, we see that confidence corresponds to the strength of a rule, while support corresponds to statistical significance. Those rules which exceed a predetermined minimum threshold for support and confidence are considered to be interesting. Syntactic constraints can also be used to restrict the items that can appear in the left- or right-hand side of a rule [2, 64].

3.5 Klemettinen et al. Rule Templates

Rule templates [34] are an extension of the syntactic constraints described in [2] and are used to describe a pattern for those attributes that can appear in the left- or right-hand side of an association rule. A rule template is given by:

$$A_1, A_2, \ldots, A_k \rightarrow A_m,$$

where each A_i is either an attribute name, a class name (a class hierarchy is used to map database values to a taxonomy of classes), or an expression $C+$ or $C*$. In the expressions $C+$ and $C*$, C is a class name and $C+$ and $C*$ correspond to one or more, and zero or more, instances of the class C, respectively. An induced rule matches the pattern specified in a rule template if it can be considered to be an instance of the pattern. Rule templates may be either inclusive or restrictive. An *inclusive rule template* specifies desirable

rules that are considered to be interesting, while a *restrictive rule template* specifies undesirable rules that are considered to be uninteresting. Rule pruning can be done by setting support, confidence, and rule size thresholds.

3.6 Matheus and Piatetsky-Shapiro's Projected Savings

Projected savings [41] is a measure that assesses the financial impact of cost deviations from some normative or expected values. Projected savings is given by:

$$PS = PI * SP,$$

where PI is the projected impact and SP is the savings percentage. The *projected impact* is given by:

$$PI = PD * PF,$$

where PD is the difference between the current average cost and the normative or expected cost for some product or service, and IF is the *impact factor* (which may be viewed as the number of units sold). The *savings percentage* is a domain expert specified value of the percentage decrease in deviation that would result following some relevant intervention strategy. The interestingness of a deviation is directly related to the projected savings achievable as a result of this strategy.

3.7 Hamilton and Fudger's I-Measures

The *I-measures* [26] are used to quantify the significance of discovered knowledge, presented in the form of generalized relations or summaries, based upon the structure of concept hierarchies associated with the attributes in the original ungeneralized relation. The I_1 measure considers the number of non-ANY, non-leaf nodes in a summary and is given by:

$$I_1 = \sum_v c(t(v)),$$

where v is an attribute value, $t(v)$ is the concept hierarchy associated with the attribute containing v, and $c(t(v))$ is a function that returns 1 if v is a non-ANY, non-leaf concept, and 0 otherwise. The I_2 measure considers the depth and weighted height for all nodes in a summary is given by:

$$I_2 = \sum_v (k)d(v,t(v)) + (1-k)wh(v,t(v)),$$

where k specifies the relative significance of the depth of a concept versus the weighted depth (e.g., $k = 0.5$ indicates the distance from the root node is as significant as the distance from the leaf nodes), v is an attribute value, $t(v)$ is the concept hierarchy associated with the attribute containing v, $d(v,t(v))$ is the depth of v in concept hierarchy $t(v)$, and $wh(v,t(v))$ is the weighted height of v in concept hierarchy $t(v)$. The depth $d(v,t(v))$ of v in concept hierarchy $t(v)$ is defined so that the depth of the root node is zero and the depth of any other concept is one more than the depth of its parent. The weighted height $wh(v,t(v))$ of v in concept hierarchy $t(v)$ is a function of the number of leaf concepts it has as descendants and the sum of the distances to each of its descendants. Summaries associated with higher values of I_1 and I_2 are considered more interesting.

3.8 Silbershatz and Tuzhilin's Interestingness

Interestingness [61] determines the extent to which a soft belief is changed as a result of encountering new evidence (i.e., discovered knowledge). A soft belief is one that a user is willing to change as new evidence is encountered. Interestingness within the context of soft beliefs is given by:

$$I = \sum_\alpha \frac{p(\alpha|E,\varepsilon) - p(\alpha|\varepsilon)|}{p(\alpha|\varepsilon)},$$

where α is a belief, E is new evidence, ε is the previous evidence supporting belief α, $p(\alpha|\varepsilon)$ is the confidence in belief α, and $p(\alpha|E,\varepsilon)$ is the new confidence in belief α given the new evidence E. Summation is over all beliefs. Bayes Theorem is used to determine the new confidence and is given by:

$$p(\alpha|E,\varepsilon) = \frac{p(E|\alpha,\varepsilon)p(\alpha|\varepsilon)}{p(E|\alpha,\varepsilon)p(\alpha|\varepsilon) + p(E|\neg\alpha,\varepsilon)p(\neg\alpha|\varepsilon)},$$

Positive (negative) evidence strengthens (weakens) the belief.

3.9 Kamber and Shinghal's Interestingness

Interestingness [31] determines the interestingness of classification rules based upon necessity and sufficiency. There are two kinds of classification rules: discriminant and characteristic. A *discriminant rule*, $e \rightarrow h$, where e is evidence and h is a hypothesis, summarizes the conditions sufficient to distinguish one class from another. *Sufficiency* is given by:

$$S(e \rightarrow h) = \frac{p(e|h)}{p(e|\neg h)}.$$

A *characteristic rule*, $h \rightarrow e$, summaries the conditions necessary for membership in a class. *Necessity* is given by:

$$N(e \rightarrow h) = \frac{p(\neg e|h)}{p(\neg e|\neg h)}.$$

Necessity and sufficiency can be used to assess the interestingness of the characteristic rule $h \rightarrow e$, as follows:

$$IC^{++} = \begin{cases} (1 - N(e \rightarrow h)) \times p(h), 0 \leq N(e \rightarrow h) < 1 \\ 0, \text{otherwise} \end{cases},$$

$$IC^{+-} = \begin{cases} (1 - S(e \rightarrow h)) \times p(h), 0 \leq S(e \rightarrow h) < 1 \\ 0, \text{otherwise} \end{cases},$$

$$IC^{-+} = \begin{cases} (1 - 1/N(e \rightarrow h)) \times p(\neg h), 1 < N(e \rightarrow h) < \infty \\ 0, \text{otherwise} \end{cases},$$

and

$$IC^{--} = \begin{cases} (1 - 1/S(e \rightarrow h)) \times p(\neg h), 1 < S(e \rightarrow h) < \infty \\ 0, \text{otherwise} \end{cases},$$

where $++$, $+-$, $-+$, and $--$ correspond to the characteristic rules $h \rightarrow e$, $h \rightarrow \neg e$, $\neg h \rightarrow e$, and $\neg h \rightarrow \neg e$, respectively. Interestingness values for each of the measures lies in $[0, 1]$, where 0 and 1 represent the minimum and maximum possible interestingness, respectively.

3.10 Hamilton et al. Credibility

Credibility [27] determines the extent to which a classification (i.e., generalized relation or summary) provides decisions for all or nearly all possible values of the condition attributes, based upon adequately supported evidence. Credibility is given by:

$$Cred_E(C) = Q_E(C) \times min(I/M, 1),$$

where E is an equivalence class, C is a classification, $Q_E(C)$ is the quality of classification C, I is the actual number of instance supporting the equivalence class E, M is the minimum number of instances required for a credible classification, and $min(I/M, 1)$ is a factor that ensures a proportional weight is associated to equivalence classes not supported by an adequate number of instances. The quality function $Q_E(C)$ is given by:

$$Q_E(C) = \beta \times p(E) \times |p(F|E) - p(F)|,$$

where β is a normalizaton factor to ensure that $Q_E(C)$ is always within the interval $[0, 1]$, $P(E)$ is the probability of equivalence class E, $P(F|E)$ is the conditional probability of the occurrence of the concept F (i.e., the decision attribute) given that E has occurred, and $P(F)$ is the probability of concept F. The normalization factor is given by:

$$\beta = \frac{1}{2p(F)(1 - p(F))}.$$

3.11 Liu et al. General Impressions

A *general impression* [36] is used to evaluate the importance of classification rules by comparing discovered rules to an approximate or vague description of what is considered to be interesting. So, a general impression is a kind of specification language. There are two types of general impressions that can be specified: Type 1 and Type 2. A *Type 1 general impression* is a rule of the form $A_1\ OP_1, A_2\ OP_2, \ldots, A_x\ OP_x \rightarrow C_j$, where each $A_i\ OP_i$ is called an impression term, each A_i is an attribute, each OP_i is an impression descriptor from the set $\{<, >, \ll, |, []\}$, and C_j is a class. The $<$ ($>$) impression descriptor means smaller (larger) attribute values are more likely to lead to inclusion in class C_j, $\ll$ means some range of attribute values are more

likely to lead to inclusion in class C_j, | means some relationship exists between an attribute and class C_j but the nature of this relationship is not exactly known, and [] means that some subset of the possible values for an attribute are more likely to lead to inclusion in class C_j. A Type 2 general impression is specified when there is more confidence that the combination of impression terms will lead to inclusion in class C_j. A *Type 2 general impression* is a rule of the form $A_1\ OP_1, A_2\ OP_2, \ldots, A_k\ OP_k \& A_m\ OP_m, A_n\ OP_n, \ldots, A_x\ OP_x \rightarrow C_j$, where the part to the left (right) of the & symbol is called the *core* (*supplement*). The core must always exist, otherwise the general impression should be specified as Type 1. If the supplement exists, then the rule is called a maximal impression. In a *maximal impression*, the general impression is that the impression terms in the core and any subset of those in the supplement are more likely to lead to inclusion in class C_j. If the supplement does not exist, then the rule is called an exact impression. In an *exact impression*, the general impression is that the impression terms in the core are more likely to lead to inclusion in class C_j. The specified general impressions are matched against the rules generated, and ranked to identify those that are most valid.

3.12 Gago and Bento's Distance Metric

The *distance metric* [21] measures the distance between two rules and is used to determine the rules that provide the highest coverage for the given data. The distance metric is given by:

$$D(R_i, R_j) = \begin{cases} (\frac{DA(R_i,R_j)+2DV(R_i,R_j)-2EV(R_i,R_j)}{N(R_i)+N(R_j)}, NO(R_i, R_j) = 0 \\ 2, \text{otherwise} \end{cases},$$

where R_i and R_j are rule i and j, respectively, $DA(R_i, R_j)$ is the sum of the number of attributes in R_i not in R_j and the number of attributes in R_j not in R_i, $DV(R_i, R_j)$ is the number of attributes in R_i and R_j that have slightly overlapping values in the range conditions (slightly overlapping means less than 66% of the range), $EV(R_i, R_j)$ is the number of attributes in R_i and R_j that have overlapping values in the range conditions (overlapping means more than 66% of the range), $N(R_i)$ and $N(R_j)$ are the number of attributes in R_i and R_j, respectively, and $NO(R_i, R_j)$ is the number of attributes in R_i and R_j with nonoverlapping values. The distance metric returns a value on $[-1, 1]$ or 2. Values near -1 and 1 indicate a strong and slight

overlap, respectively, while the value 2 indicates no overlap. The rules with the highest average distance to the other rules are considered to be the most interesting.

3.13 Freitas' Surprisingness

Surprisingness [20] is a measure that determines the interestingness of discovered knowledge via the explicit detection of occurrences of Simpson's paradox. Simpson's paradox is described, as follows. Let X_1 and X_2 be two mutually exclusive and exhaustive populations partitioned according to the value of a binary event attribute E, where E_1 and E_2 are the values of E in X_1 and X_2, respectively. Let $P(E_1)$ and $P(E_2)$ be the probabilities of events E_1 and E_2 in X_1 and X_2, respectively. Now let X_1 and X_2 be further partitioned according to the value of a categorical attribute having m distinct values (i.e., event E_i is partitioned into events E_{ij}, $i = 1, 2$ and $j = 1, 2, \ldots, m$). Then let $P(E_{1j})$ and $P(E_{2j})$ be the probabilities for events E_{1j} and E_{2j} in X_1 and X_2, respectively. Assuming that $P(E_1) > P(E_2)$ $(P(E_1) < P(E_2))$, *Simpson's paradox* occurs when $P(E_{1j}) \leq P(E_{2j})$ $(P(E_{1j}) \geq P(E_{2j}))$ for all $j = 1, 2, \ldots, m$. That is, although event E_1 (E_2) has a higher (lower) overall probability of occurring, the probability of occurrence of each E_{1j} (E_{2j}) in E_1 (E_2) can actually be lower (higher) than each E_{2j} (E_{1j}) in E_2 (E_1).

3.14 Gray and Orlowska's Interestingness

Interestingness [23] is used to evaluate the strength of associations between sets of items in retail transactions (i.e., association rules). While support and confidence have been shown to be useful for characterizing association rules, interestingness contains a discrimination component that gives an indication of the independence of the antecedent and consequent. Interestingness is given by:

$$I = \left(\left(\frac{P(X \cap Y)}{P(X) \times P(Y)} \right)^k - 1 \right) \times (P(X) \times P(Y))^m,$$

where $P(X \cap Y)$ is the confidence, $P(X) \times P(Y)$ is the support, $\frac{P(X \cap Y)}{P(X) \cap P(Y)}$ is the discrimination, and k and m are parameters to weight the relative importance of the discrimination and support components, respectively. Higher values of interestingness are considered more interesting.

3.15 Dong and Li's Interestingness

Interestingness [13] is used to evaluate the importance of an association rule by considering its unexpectedness in terms of other association rules in its neighborhood. The *neighborhood* of an association rule consists of all association rules within a given distance. The distance metric is given by:

$$D(R_1, R_2) = \delta_1 \times |(X_1 \cup Y_1)\Theta(X_2 \cup Y_2)| + \delta_2 \times |X_1 \Theta X_2| + \delta_3 \times |Y_1 \Theta Y_2|,$$

where $R_1 = X_1 \rightarrow Y_1$, $R_2 = X_2 \rightarrow Y_2$, δ_1, δ_2, and δ_3 are parameters to weight the relative importance of all three terms, and Θ is an operator denoting the symmetric difference between X and Y (i.e., $(X-Y) \cup (Y-X)$). An *r-neighborhood* of a rule is given by the set:

$$N(R_0, r) = \{R | D(R, R_0) \leq r, \text{R a potential rule}\}$$

and is used to define the interestingness of a rule. Two types of interestingness are: unexpected confidence and isolated. *Unexpected confidence interestingness* is given by:

$$UCI = \begin{cases} 1, \text{if} ||c(R_0) - ac(R_0, r)| - sc(R_0, r)| > t_1 \\ 0, \text{otherwise}, \end{cases},$$

where $c(R_0)$ is th confidence of R_0, $ac(R_0, r)$ and $sc(R_0, r)$ are the average confidence and standard deviation of the rules in the set $M \cap N(R_0, r) - \{R_0\}$ (M is the set of rules satisfying the minimum support and confidence), and t_1 is a threshold. *Isolated interestingness* is given by:

$$II = \begin{cases} 1, \text{if} |N(R_0, r)| - |M \cap N(R_0, r)| > t_2 \\ 0, \text{otherwise}, \end{cases},$$

where $|N(R_0, r)|$ is the number of potential rules in an r-neighborhood, $|M \cap N(R_0, r)|$ is the number of rules generated from the neighborhood, and t_2 is a threshold.

3.16 Liu et al. Reliable Exceptions

A *reliable exception* [37] is a weak rule having relatively small support and relatively high confidence. Reliable exceptions can be induced, as follows. First, use rule induction to generate the strong rules (or some predetermined number of the strongest rules according to some threshold). Reliable exceptions will be evaluated with respect to these strong rules. Second, using contingency table analysis, identify significant deviations between the actual and expected frequency of occurrence of attribute-value and class pairs. Third, specify a deviation threshold. For positive (negative) deviations, any deviation greater than (less than) the threshold is considered outstanding. Fourth, get al. l instances containing the attribute-value and class pairs of the outstanding negative deviations (i.e., all instances satisfying the rule $X \rightarrow c$, where X is an attribute-value and c is a class. Fifth, calculate the difference between the confidence of the rule $X \rightarrow c$ for the selected instances and the whole dataset. Now the confidence for the selected instances is always 1. So, a large difference (i.e., near 1) implies that the confidence on the whole dataset is low (i.e., near 0), and thus, a reliable exception has been discovered.

3.17 Zhong et al. Peculiarity

Peculiarity [73] is used to determine the extent to which one data object differs from other similar data objects. The peculiarity factor is given by:

$$PF(x_i) = \sum_{j=1}^{n} \sqrt{N(x_i, x_j)},$$

where x_i and x_j are attribute values, n is the number of different attribute values, and $N(x_i, x_j)$ is the conceptual distance between x_i and x_j. The conceptual difference is given by:

$$N(x_i, x_j) = |x_i - x_j|.$$

4 Conclusion

Here we classify the interestingness measures from the previous section according to the three criteria shown in Table 1. The *Representation* column describes the general form of the knowledge representation expected by each measure, the *Foundation* column describes the general nature of the fundamental calculation or methodology for each measure (i.e., utilitarian, probabilistic, syntactic, distance), the *Scope* column describes the number of rules covered by each interestingness value generated by each measure (i.e., a single rule or the whole rule set), and the *Class* column describes the the class of each measure (i.e., objective or subjective). *Objective measures* are based upon the structure of the discovered patterns, while *subjective measures* are based upon user beliefs or biases regarding relationships in the data.

Table 1: Classification of Interestingness Measures

Interestingness Measure	*Representation*	*Foundation*	*Scope*	*Class*
Piatetsky-Shapiro's Rule-Interest Function	classification rules	probabilistic	single rule	objective
Smyth and Goodman's J-Measure	classification rules	probabilistic	single rule	objective
Major and Mangano's Rule Refinement	classification rules	probabilistic	single rule	objective
Agrawal and Srikant's Itemset Measures	association rules	probabilistic	single rule	objective
Klemettinen et al. Rule Templates	association rules	syntactic	single rule	subjective
Matheus and Piatetsky-Shapiro's Projected Savings	summaries	utilitarian	single rule	subjective
Hamilton and Fudger's I-Measures	generalized relations	distance	rule set	objective
Silbershatz and Tuzhilin's Interestingness	format-independent	probabilistic	rule set	subjective
Kamber and Shinghal's Interestingness	classification rules	probabilistic	single rule	objective
Hamilton et al. Credibility	generalized relations	probabilistic	rule set	objective
Liu et al. General Impressions	classification rules	syntactic	single rule	subjective
Gago and Bento's Distance Metric	classification rules	distance	rule set	objective
Freitas' Surprisingness	format-independent	probabilistic	rule set	objective
Gray and Orlowska's Interestingness	association rules	probabilistic	single rule	objective
Dong and Li's Interestingness	association rules	distance	single rule	subjective
Liu et al. Reliable Exceptions	association rules	probabilistic	single rule	objective
Zhong et al. Peculiarity	association rules	distance	single rule	objective

5 Acknowledgements

We acknowledge the support of the Institute for Robotics and Intelligent Systems, the Networks of Centres of Excellence Program of the Government of Canada, the Natural Sciences and Engineering Research Council (NSERC), and the participation of PRECARN Associates, Inc.

References

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'98)*, pages 94–105, June 1998.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on the Management of Data (SIGMOD'93)*, pages 207–216, Washington, D.C., May 1993.

[3] R. Agrawal, H. Mannila, R.Srikant, H.Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328, Menlo Park, California, 1996. AAAI Press/MIT Press.

[4] R. Agrawal and J.C. Shafer. Parallel mining of association rules. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):962–969, December 1996.

[5] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Databases (VLDB'94)*, pages 487–499, Santiago, Chile, September 1994.

[6] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the 11th International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan, March 1995.

[7] K. Alsabti, S. Ranka, and V. Singh. Clouds: a decision tree classifier for large datasets. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 2–8, New York, New York, August 1998.

[8] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'97)*, pages 265–276, May 1997.

[9] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'97)*, pages 255–264, May 1997.

[10] O. Buchter and R. Wirth. Discovery of association rules over ordinal data: a new and faster algorithm and its application to basket analysis. In X. Wu, R. Kotagiri, and K. Korb, editors, *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, pages 36–47, Melbourne, Australia, April 1998.

[11] P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3:261–283, 1989.

[12] G. Das, K.-I. Lin, H. Mannila, G. Renganathan, and P. Smyth. Rule discovery from times series. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 16–22, New York, New York, August 1998.

[13] G. Dong and J. Li. Interestingness of discovered association rules in terms of neighborhood-based unexpectedness. In X. Wu, R. Kotagiri, and K. Korb, editors, *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, pages 72–86, Melbourne, Australia, April 1998.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231, Portland, Oregon, August 1996.

[15] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI/MIT Press, 1996.

[16] S.E. Fienberg. *The analysis of cross-classified categorical data.* MIT Press, 1978.

[17] D.J. Fifield. Distributed tree construction from large datasets. Master's thesis, Australian National University, 1992.

[18] D.H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.

[19] W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus. Knowledge discovery in databases: An overview. In *Knowledge Discovery in Databases*, pages 1–27. AAAI/MIT Press, 1991.

[20] A.A. Freitas. On objective measures of rule surprisingness. In J. Zytkow and M. Quafafou, editors, *Proceedings of the Second European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'98)*, pages 1–9, Nantes, France, September 1998.

[21] P. Gago and C. Bentos. A metric for selection of the most promising rules. In J. Zytkow and M. Quafafou, editors, *Proceedings of the Second European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'98)*, pages 19–27, Nantes, France, September 1998.

[22] L.A. Goodman. *The analysis of cross-classified data having ordered categories.* Harvard University Press, 1984.

[23] B. Gray and M.E. Orlowska. Ccaiia: clustering categorical attributes into interesting association rules. In X. Wu, R. Kotagiri, and K. Korb, editors, *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, pages 132–143, Melbourne, Australia, April 1998.

[24] F.O. Gur-Ali and W.A. Wallace. Are we losing accuracy while gaining confidence in induced rules - an assessment of PrIL. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, pages 9–14, Montreal, Canada, August 1995.

[25] V. Guralnik, D. Wijesekera, and J. Srivastava. Pattern directed mining of sequence data. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 51–57, New York, New York, August 1998.

[26] H.J. Hamilton and D.F. Fudger. Estimating DBLearn's potential for knowledge discovery in databases. *Computational Intelligence*, 11(2):280–296, 1995.

[27] H.J. Hamilton, N. Shan, and W. Ziarko. Machine learning of credible classifications. In A. Sattar, editor, *Proceedings of the Tenth Australian Conference on Artificial Intelligence (AI'97)*, pages 330–339, Perth, Australia, November/December 1997. Springer Verlag.

[28] E.-H. Han, G. Karypis, and V. Kumar. Scalable parallel data mining for association rules. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'97)*, pages 277–288, May 1997.

[29] J. Han, W. Ging, and Y. Yin. Mining segment-wise periodic patterns in time-related databases. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 214–218, New York, New York, August 1998.

[30] R.J. Hilderman, C.L. Carter, H.J. Hamilton, and N. Cercone. Mining market basket data using share measures and characterized itemsets. In X. Wu, R. Kotagiri, and K. Korb, editors, *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, pages 159–173, Melbourne, Australia, April 1998.

[31] M. Kamber and R. Shinghal. Evaluating the interestingness of characteristic rules. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 263–266, Portland, Oregon, August 1996.

[32] L. Kaufman and P.J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis.* Wiley and Sons, 1978.

[33] E.J. Keogh and M.J. Pazzani. An enhanced representation of time series which allows fast and accurate classification, clustering, and relevance feedback. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 239–243, New York, New York, August 1998.

[34] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. In N.R. Adam, B.K. Bhargava, and Y. Yesha, editors,

Proceedings of the Third International Conference on Information and Knowledge Management, pages 401–407, Gaitersburg, Maryland, 1994.

[35] A.J. Knobbe and P.W. Adrians. Analyzing binary associations. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 311–314, Portland, Oregon, August 1996.

[36] B. Liu, W. Hsu, and S. Chen. Using general impressions to analyze discovered classification rules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*, pages 31–36, Newport Beach, California, August 1997.

[37] H. Liu, H. Lu, L. Feng, and F. Hussain. Efficient search of reliable exceptions. In N. Zhong and L. Zhou, editors, *Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, pages 194–203, Beijing, China, April 1999.

[38] J.A. Major and J.J Mangano. Selecting among rules induced from a hurricane database. In *Knowledge Discovery in Databases: Papers from the 1993 Workshop*, pages 28–41, Menlo Park, California, 1993. AAAI Press. WS-93-02.

[39] H. Mannila and H. Toivonen. Discovering generalized episodes using minimal occurrences. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 146–151, Portland, Oregon, August 1996.

[40] H. Mannila, H. Toivonen, and A.I. Verkamo. Discovering frequent episodes in sequences. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, pages 210–215, Montreal, Canada, August 1995.

[41] C.J. Matheus and G. Piatetsky-Shapiro. Selecting and reporting what is interesting: The kefir application to healthcare data. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 401–419, Menlo Park, California, 1996. AAAI Press/MIT Press.

[42] M. Mehta, R. Agrawal, and J. Rissanen. Sliq: a fast scalable classifier for data mining. In *Proceedings of the Fifth International Conference on Extending Database Technology (EDBT'96)*, pages 18–32, Avignon, France, March 1996.

[43] R.S. Michalski, I. Mozetic, J. Hong, and N. Lavrac. The multi-purpose incremental learning system aq15 and its testing application to three medical domains. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 1041–1045, 1986.

[44] R.S. Michalski and R.E. Stepp. Learning from observation: conceptual clustering. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach*, pages 331–363. Tioga Publishing Company, 1983.

[45] T.M. Mitchell. *Version Spaces: An Approach to Concept Learning.* PhD thesis, Stanford University, 1978.

[46] T.M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.

[47] R.T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Databases (VLDB'94)*, pages 144–155, Santiago, Chile, September 1994.

[48] B. Padmanabhan and A. Tuzhilin. Pattern discovery in temporal databases: a temporal logic approach. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 351–354, Portland, Oregon, August 1996.

[49] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 94–100, New York, New York, August 1998.

[50] J.S. Park, M.-S. Chen, and P.S. Yu. An effective hash-based algorithm for mining association rules. *SIGMOD Record*, 25(2):175–186, 1995.

[51] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

[52] J. R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1993.

[53] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[54] D. Rafiei and A. Mendelzon. Similarity-based queries for time series data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'97)*, pages 13–23, May 1997.

[55] R.B. Rao, S. Rickard, and F. Coetzee. Time series forecasting from high-dimensional data with multiple adaptive layers. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 319–323, New York, New York, August 1998.

[56] H.T. Reynolds. *The analysis of cross-classifications*. Free Press, 1977.

[57] J. Rissanen. *Stochastic complexity in statistical inquiry*. World Scientific Publishing Company, 1989.

[58] A.P. Sanjeev and J. Zytkow. Discovering enrollment knowledge in university databases. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, pages 246–251, Montreal, Canada, August 1995.

[59] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21th International Conference on Very Large Databases (VLDB'95)*, pages 432–444, Zurich, Switzerland, September 1995.

[60] J. Shafer, R. Agrawal, and M. Mehta. Sprint: a scalable parallel classifier for data mining. In *Proceedings of the 22nd International Conference on Very Large Databases (VLDB'96)*, pages 544–555, Mumbay, India, September 1996.

[61] A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, pages 275–281, Montreal, Canada, August 1995.

[62] P. Smyth and R.M. Goodman. Rule induction using information theory. In *Knowledge Discovery in Databases*, pages 159–176. AAAI/MIT Press, 1991.

[63] R. Srikant and R. Agrawal. Mining sequential patterns: generalization and performance improvements. In *Proceedings of the Fifth International Conference on Extending Database Technology (EDBT'96)*, Avignon, France, March 1996.

[64] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*, pages 67–73, Newport Beach, California, August 1997.

[65] H. Toivonen. Sampling large databases for finding association rules. In *Proceedings of the 22nd International Conference on Very Large Databases (VLDB'96)*, pages 134–145, Mumbay, India, September 1996.

[66] W. Wang, J. Yang, and R. Muntz. Sting: a statistical information grid approach to spatial data mining. In *Proceedings of the 23nd International Conference on Very Large Databases (VLDB'97)*, pages 186–195, Athens, Greece, September 1997.

[67] G.M. Weiss and H. Hirsh. Learning to predict rare events in event sequences. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 359–363, New York, New York, August 1998.

[68] M.J. Zaki, N. Lesh, and M. Ogihara. PlanMine: sequence mining for plan failures. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 369–373, New York, New York, August 1998.

[69] M.J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*, pages 283–286, Newport Beach, California, August 1997.

[70] R. Zembowicz and J. Zytkow. From contingency tables to various forms of knowledge in databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 329–349, Menlo Park, California, 1996. AAAI Press/MIT Press.

[71] J. Zhang and R.S. Michalski. An integration of rule induction and exemplar-based learning for graded concepts. *Machine Learning*, 21:235–267, 1995.

[72] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'96)*, pages 103–114, June 1996.

[73] N. Zhong, Y.Y. Yao, and S. Ohsuga. Peculiarity-oriented multi-database mining. In J. Zytkow and J. Rauch, editors, *Proceedings of the Third European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'99)*, pages 136–146, Prague, Czech Republic, September 1999.