Discovering significant rules

Geoffrey I. Webb
Faculty of Information Technology
Monash University
Clayton, Vic, Australia
webb@infotech.monash.edu

ABSTRACT

In many applications, association rules will only be interesting if they represent non-trivial correlations between all constituent items. Numerous techniques have been developed that seek to avoid false discoveries. However, while all provide useful solutions to aspects of this problem, none provides a generic solution that is both flexible enough to accommodate varying definitions of true and false discoveries and powerful enough to provide strict control over the risk of false discoveries. This paper presents generic techniques that allow definitions of true and false discoveries to be specified in terms of arbitrary statistical hypothesis tests and which provide strict control over the experimentwise risk of false discoveries.

Categories and Subject Descriptors: H.2.8 [**Database Management**] Database Applications: data mining

General Terms: Algorithms, Performance, Reliability, Experimentation

Keywords: Association rules, Rule discovery, Statistics

1. INTRODUCTION

Association rule discovery [1] finds collections of items that co-occur frequently in data. In many applications, such rules will only be interesting if they represent non-trivial correlations between all constituent items. For the purposes of this paper we will call such associations *significant rules* and all remaining associations *false discoveries*. Many techniques have been developed that seek to avoid false discoveries [1, 3, 4, 5, 10, 12, 16, 18, 20, 22, 25, 23, 28, 29]. This paper builds upon this body of previous work, presenting two generic techniques that both allow definitions of true and false discoveries to be specified in terms of arbitrary statistical hypothesis tests, while providing strict control over the risk of false discoveries. We show that each has relative strengths and weaknesses and provide analyses of these. We substantiate the need for strict control over the risk of false discoveries, showing that on some real-world tasks there is

potential for all 'discoveries' to be false unless appropriate safeguards are employed.

2. PROBLEM STATEMENT

We consider the problem of finding rules from data $D = \langle t_1, t_2, \ldots, t_n \rangle$, where each *transaction* or *record* $t_i \subseteq I$ and $I = \{item_1, item_2, \ldots item_m\}$ is the set of *items* of which transactions are composed. For *market-basket* data items are atomic forms and for *attribute-value* data items have the form $a_i = v_{i,j}$ where a_i represents an attribute and $v_{i,j}$ a value of a_i. For attribute value data, no transaction t_i, $1 \le i \le n$ may contain two items $a_i = v_{i,j}$ and $a_i = v_{i,k}$, $j \ne k$. That is, each transaction may contain at most one value for each attribute. Rules take the form $x \rightarrow y$ where $x \subset I$ and $y \in I$. Note that we limit the consequent y to a single value. While many association rule techniques allow multiple values in the consequent y, the techniques we present generalize directly to multiple-value consequents and a single rule with multiple elements in the consequent can be represented by multiple rules with single elements in the consequent.

We are potentially interested in a number of properties of a rule $x \rightarrow y$ relative to D, and these properties vary from application to application. In this paper we utilize *support* [1], *confidence* [1], *lift* [16] and *leverage* [22], defined as follows:

$$sup(x \rightarrow y, D) = |\{i : x \subseteq t_i \wedge y \in t_i\}| \qquad (1)$$
$$conf(x \rightarrow y, D) = sup(x \rightarrow y, D) / |\{i : x \subseteq t_i\}| \qquad (2)$$
$$lift(x \rightarrow y, D) = conf(x \rightarrow y, D) / (|\{i : y \in t_i\}| / n) \qquad (3)$$
$$lev(x \rightarrow y, D) = sup(x \rightarrow y, D) - $$
$$|\{i : x \subseteq t_i\}| \times |\{i : y \in t_i\}| / n \qquad (4)$$

Note that the parameters $x \rightarrow y$ and D will be omitted from these functions where they can be determined from the context.

The original association rule task [1] was to find all rules $x \rightarrow y$ such that $sup \ge minsup$ and $conf \ge minconf$, where $minsup$ and $minconf$ are user-specified constraints.

Typically, rules will only be interesting if they represent non-trivial correlations between items. Relatively high values of $minsup$ and $minconf$ usually deliver rules for which x and y are correlated when applied to the sparse data typical of market-basket analysis [20]. However, as will be demonstrated in the experiments below, this is not the case for dense data such as typical attribute-value data. Also, there is a serious problem that x may contain items that are independent of y, and hence potentially misleading. To illustrate this problem consider a rule $\{pregnant\} \rightarrow$

oedema that represents a relationship between pregnancy and oedema. Now consider $\{pregnant, female\} \rightarrow oedema$. Assuming that all cases of pregnancy will be female, this will represent an equally strong correlation to the first rule, but in most contexts *redundant rules* such as this will not be useful so long as the first rule together with the further rule $\{pregnant\} \rightarrow female$ are known. For a different kind of example consider another term *dataminer* that we will assume is in no way related to *oedema*. In this case $\{pregnant, dataminer\} \rightarrow oedema$ should represent as strong a correlation between the antecedent and consequent as the first rule, the only difference being a reduction in support and random differences in confidence resulting from sampling effects. However, in most contexts *unproductive rules* such as this will be of no interest so long as the first rule is known.

Such redundant and unproductive rules represent rules $x \rightarrow y$ for which there exists $z \in x$ such that $P(y \mid x) = P(y \mid x\backslash\{z\})$, or, in other words, for which z and y are conditionally independent given $x\backslash\{z\}$.

Apriori, one might expect there to be very large numbers of unproductive rules, as from every single productive rule $x \rightarrow y$ many unproductive rules can be generated by inserting into x any arbitrary collections of unrelated items.

Note that for the rest of this paper we will assume that the analytic task of interest is to identify positive rules and hence we will regard as false discoveries any negative rules such as $\{male\} \rightarrow pregnant$ [*confidence=0.0*]. This seems a reasonable assumption in the context of rule discovery using a minimum support constraint, as strong, and hence presumably interesting, negative associations will have low support and hence are excluded from consideration by the imposition of a minimum-support constraint. However, the techniques that we advance generalize directly to different definitions of false discoveries, requiring only the use of different constraints and statistical tests.

3. TECHNIQUES FOR PREVENTING FALSE DISCOVERIES

Clearly it is desirable to avoid false discoveries and numerous techniques have successfully tackled aspects of this problem.

Non-redundant rule techniques [3, 28] identify and discard rules such as $\{pregnant, female\} \rightarrow oedema$. Specifically, they discard rules $x \rightarrow y$ for which $\exists z{\in}x : sup(x \rightarrow y) = sup(x\backslash\{z\} \rightarrow y)$.

A more powerful filter is provided by a *minimum improvement* constraint [5]. The improvement of rule $x \rightarrow y$ is defined as

$$imp(x \rightarrow y) = conf(x \rightarrow y) - \max_{z \subset x}(conf(z \rightarrow y)) \quad (5)$$

It represents the difference between the confidence of a rule and the highest confidence of any of its generalizations. A minimum improvement constraint is justified in contexts where only positive associations are of interest. In that case an association is unlikely to be of interest unless it represents a stronger correlation than any of its generalizations. A redundant rule will have improvement no greater than 0.0, as for it to be redundant it must have a generalization with identical confidence. Thus, a minimum improvement constraint is stronger than a non-redundant rule constraint as it rejects all redundant rules as well as many unproductive

rules. The only unproductive rules that such a minimum-improvement constraint will fail to reject are those for which random sampling effects happen to result in raised confidence for a rule by chance. However, almost 50% of rules may fall into this category, because confidence is as likely to be raised as lowered through sampling effects and is unlikely to exactly represent the true probability of the consequent given the antecedent unless that probability is 1.0. If the minimum improvement constraint is set high enough to exclude the majority of these cases, it is also likely to exclude many productive rules.

An even stronger filter is represented by the use of statistical hypothesis tests, either to test for independence between x and y [10, 17, 18, 29] or to test for unproductive rules [25]. We will focus here on the test for unproductive rules, as it most directly addresses the issue at hand, but the key points also apply to the other approaches. For the sake of computational efficiency this test compares $x \rightarrow y$ only against the global frequency of y and against each of its immediate generalizations $x\backslash\{z\} \rightarrow y$, where $z \in x$. We do not test against all generalizations as there are $2^{|x|} - 1$ of these, and hence for large antecedents the computation would be infeasible.

This test rejects a rule $x \rightarrow y$ if $p \leq \alpha$ for a Fisher exact test [2] for improvement with respect to any of the rule's immediate generalizations $x\backslash\{z\} \rightarrow y$ and with respect to $\{\} \rightarrow y$. The p value for the test for improvement with respect to an immediate generalization $x\backslash\{z\} \rightarrow y$ can be calculated as follows. Let $a = |\{i : x \subseteq t_i \wedge y \in t_i\}|$ (the support for the rule), $b = |\{i : x \subseteq t_i \wedge y \notin t_i\}|$ (the number of transactions that contain x but not y), $c = |\{i : (x\backslash\{z\}) \subseteq t_i \wedge z \notin t_i \wedge y \in t_i\}|$ (the number of transactions that contain y and all the x values other than z but not z) and $d = |\{i : (x\backslash\{z\}) \subseteq t_i \wedge z \notin t_i \wedge y \notin t_i\}|$ (the number of transactions that contain all the x values other than z but neither y nor z).

$$p = \sum_{i=0}^{min(b,c)} \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!(a+i)!(b-i)!(c-i)!(d+i)!}. \quad (6)$$

Here, $q!$ denotes the factorial of q. By convention $\alpha = 0.05$ is used. Equation (6) is also used to calculate the p value for the test for improvement with respect to $\{\} \rightarrow y$, using the same a and b, but with $c = |\{i : x \not\subseteq t_i \wedge y \in t_i\}|$ (the number of transactions that contain y but not x) and $d = |\{i : x \not\subseteq t_i \wedge y \notin t_i\}|$ (the number of transactions that contain neither x nor y).

The use of this statistical test avoids the problem of setting an appropriate minimum improvement constraint, as it rejects all rules for which there is insufficient evidence that improvement is greater than zero. However, it still suffers from a very serious problem known as the *multiple comparisons* or *multiple tests* problem. For each rule considered, the risk of it being accepted if it is not productive is no more than α. Now consider a typical market basket task for which more than 10^{22} potential rules might be considered. If $\alpha = 0.05$ and none of these rules is productive it is still possible that as many as 5×10^{20} rules might be accepted. Clearly this is an undesirable state of affairs.

An alternative approach is to use shrinkage estimates, or Bayesian smoothing, to provide conservative estimates of the true probability of a set of items [12, 23, 26]. These approaches can be very effective at reducing the overestimates of measures such as support or confidence that can occur for

rules with low support. Their use can reduce type-1 error with respect to minimum support or confidence (or similar) constraints. However, they do not allow for the number of alternative rules under consideration, and hence do not address the multiple tests problem. Nor do they provide a general mechanism for assessing arbitrary criteria for defining false discoveries.

One solution that has been proposed is randomization tests [20]. Under this approach the data are randomized to establish the null hypothesis to be tested (for example, that x and y are independent). The rule discovery software is run under multiple such randomizations and settings are identified such that rules would be discovered for no more than α proportion of the runs. If the software is subsequently run on the non-randomized data with these settings the probability that any rule discovered satisfies the null hypothesis is no more than α. Unfortunately, however, this approach does not solve the problem at hand, as it requires that a single randomization of the data establish all null hypotheses that are required. Consider the example of $\{pregnant, dataminer\} \rightarrow oedema$. To be able to reject this rule we would need to randomize the data so as to make $dataminer$ independent of $pregnant$ and $oedema$ while retaining the existing correlations between the latter two values. However, we also need to test whether $pregnant$ is conditionally independent of $oedema$ given $dataminer$, and for this purpose we need to randomize the data to make $pregnant$ independent of $dataminer$ and $oedema$ while not altering any correlations between this second pair. Clearly it is not possible to perform a single randomization that satisfies both these requirements.

4. THE WITHIN-SEARCH APPROACH

The classical statistical solution to the multiple tests problem is to employ a procedure such as the well-known Bonferroni adjustment that makes explicit allowance for the number of hypotheses tested and either seeks to control the experimentwise risk of false discoveries (the risk that any false discovery will occur) [14] or the false discovery rate (the expected proportion of discoveries that are false discoveries) [6]. In the current paper we address only control of the experimentwise risk of false discoveries, but it would be straightforward to extend the techniques to control of the false discovery rate by simply substituting an adjustment for the latter in place of the former.

The Bonferroni adjustment replaces α in the hypothesis tests with $\alpha' = \alpha/r$, where r is the number of tests performed. This ensures that the experimentwise risk of false discoveries is no more than α. This adjustment provides strict control over the experimentwise risk of false discoveries, even if the hypothesis tests are correlated with one another. This feature is important in the association rule discovery context, as many rules considered are likely to be closely related to one another and hence the hypothesis tests applied are likely to be strongly correlated.

More powerful alternatives exist to the Bonferroni adjustment [24], such as the Holm procedure [15]. The Holm procedure requires that all hypothesis tests be evaluated and their p values ordered from lowest p_1 to highest p_r. The adjusted α is then $\alpha' = \max(p_i : \forall 1 \leq j \leq i, p_j \leq \alpha / (r - j + 1))$. All such alternatives to the Bonferroni adjustment require that all tests be evaluated before the adjusted significance level is determined. Clearly this is infeasible during associa-

tion rule discovery, as exploration of the large search spaces involved is only feasible if efficient pruning is able to avoid explicit consideration of the majority of rules.

Before we can apply a Bonferroni adjustment we need an upper bound on the number of hypothesis tests in the search space. For market-basket data it is straightforward to determine the size of the search space. Recall that m is the total number of items and assume that x must contain at least one item and that there is an upper bound $maxx$ on the number of items it may contain. There are m possible values for y, and for each y value there are $m-1$ items from which up to $maxx$ x values are selected.

$$s = m \times \sum_{i=1}^{maxx} C_i^{m-1} \qquad (7)$$

where C_i^{m-1} is the number of combinations containing i out of $m-1$ items. So, for example, with the Retail dataset, used below, the number of items is 16,470 and hence with x limited to no more than 5 items the size of the rule space is 1.66×10^{23}.

For attribute-value data the situation is a little more complex, as no rule containing more than one item for a single attribute can be productive. Examples of such rules include $\{sex{=}male, sex{=}female\} \rightarrow occupation{=}dataminer$ and $\{sex{=}male, occupation{=}dataminer\} \rightarrow sex{=}female$. A tighter bound can be calculated if all such rules are excluded from the calculation. To calculate the total s we must first be able to calculate the number of combinations of values of a given subset of i attributes, $atts$. To do so we first order the attributes in arbitrary order from 1 to i and refer to the individual attributes using this order as $att_1 \dots att_i$. We use intermediate values $c_{att,j,k}$ that each represent the total number of combinations of up to j items, where items contain only values for attributes $att_1 \dots att_k$.

$$c_{att,j,k} = \begin{cases} \#att_k, & j=1, k=1 \\ 0, & j>1, k=1 \\ c_{att,1,k-1} + \#att_k, & j=1, k>1 \\ c_{att,j,k-1} + \#att_k \times c_{att,j-1,k-1}, & \text{otherwise} \end{cases}$$
$$(8)$$

where $\#att_j$ is the number of values for attribute att_j. An upper bound on the number of hypothesis tests in the search space can then be calculated as follows, where each z represents the use of an attribute in the role of consequent.

$$s = \sum_{z \in a} \left(\#z \times \sum_{j=1}^{maxx} c_{a \setminus \{z\}, m, j} \right) \qquad (9)$$

It is possible that the Bonferroni adjustment has been overlooked by the association rule discovery community because it has been assumed that the required adjustments are so large that the resulting adjusted significance levels (for example $\alpha' = 0.05 / 1.66 \times 10^{23} = 3.01 \times 10^{-25}$) will be so low that no rules will be discovered. However, as this paper will show, this turns out not to be the case.

This approach is called a *within-search* approach, as statistical tests, with an appropriate Bonferroni adjustment, are applied to rules as they are encountered during the search process.

5. HOLDOUT EVALUATION

Before investigating the application of a within-search Bonferroni adjustment to find significant rules, we should
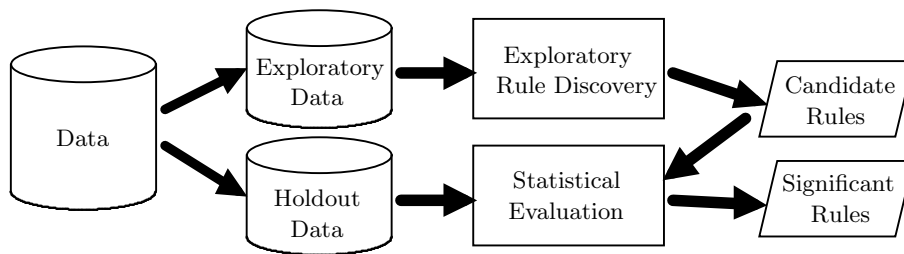
Figure 1: Holdout evaluation process

also consider another alternative. Rather than applying statistical tests during the rule discovery process, we could partition our data into exploratory and holdout sets, discover candidate rules using the exploratory data and then test those rules using the holdout data, accepting only those rules that pass relevant statistical tests for significance. This process is illustrated in Figure 1. It will be necessary to correct for multiple tests, but only with respect to the number of rules found in the exploratory stage, not the full size of the search space considered. As the former is likely to be much smaller than the latter, the adjustment will be much smaller. Further, because only a constrained number of rules will be tested, it becomes feasible to employ a more powerful alternative to the Bonferroni adjustment, such as the Holm procedure. Note that, unlike most similar procedures, the Holm procedure is also safe in the face of correlated hypothesis tests. If it were desired to control the false discovery rate rather than the experimentwise risk of error, the Benjamini-Yekutieli procedure [7], which is likewise safe, could also be used.

The use of holdout evaluation in this way is similar to established holdout evaluation methodology in machine learning, except that whereas it is used there to obtain unbiased estimates of properties of a single model, such as its error, here it is being used to perform unbiased hypothesis tests on multiple models.

There are a number of reasons to believe apriori that holdout evaluation might be more powerful than applying a Bonferroni adjustment during rule discovery. First, more powerful adjustments such as the Holm procedure can be applied. Second, the scale of the adjustments should be far smaller, as there should be far fewer rules discovered than there are rules in the search space. For example, if 10,000 rules are found then the adjusted significance level for a Bonferroni adjustment would be 5.00×10^{-6}, irrespective of the size of the search space from which the 10,000 rules were discovered. On the other hand, however, the power of the tests must be reduced by the use of less data both for the initial rule discovery and also for the statistical tests.

6. EXPERIMENTS

These experiments seek to answer the following questions:

1. Is there a need for significance tests during association rule discovery?

2. Do the within-search and holdout-evaluation processes successfully limit false discoveries in practice?

3. How powerful are the within-search and holdout-evaluation processes? That is, what is the risk that they will fail to find true rules?

4. How do the techniques perform on real-world data?

A pre-release of version 3.1 of the well-known *Magnum Opus* rule discovery software was employed [26]. This software implements *k*-optimal rule discovery [27], whereby the user specifies a maximum number of rules to be discovered *k* together with a measure to optimize such as *support*, *confidence*, *lift* or *leverage* and any other constraints to impose, such as that the rules must be non-redundant or productive. It also supports the application of a Fisher exact test as described in Section 4 and holdout-evaluation. Using this software the within-search technique can be evaluated simply by calculating and imposing the appropriate adjusted significance level.

6.1 Experiment 1

The first experiment sought to investigate questions 1 and 2. Random data were generated containing 10,000 transactions, each containing values for 100 binary variables, with each value being equiprobable. As each value was generated randomly without reference to any other variable, all variables are independent of one another and all rules are false discoveries. 100 such data sets were generated. Magnum Opus was applied to each data set using each of the following set of parameters.

Non-redundant: find the 1000 non-redundant rules with the highest leverage.

Productive: find the 1000 productive rules with the highest leverage.

Significance=0.05: find the 1000 rules with the highest leverage that pass a significance test at the 0.05 significance level.

Bonferroni: find the 1000 rules that pass a significance test at the 1.77×10^{-08} significance level that results from applying a Bonferroni correction to a raw significance level of 0.05 with a search space of 2.82×10^{06} rules.

Non-redundant+holdout: find the 1000 non-redundant rules with the highest leverage from half the data and then validate the rules using the remaining holdout data.

Productive+holdout: find the 1000 productive rules with the highest leverage from half the data and then validate the rules using the remaining holdout data.

For all settings the maximum antecedent (x) size was set to the default value of 4.

The non-redundant, productive and significance=0.05 treatments all resulted in discovery of 1000 rules for every dataset. Table 1 shows the minimum, mean and maximum support, confidence and leverage for each of these treatments. As can be seen, some rules had substantial support, confidence and leverage. For this task there were almost no differences in the rules discovered by the non-redundant

Table 1: Support, confidence and leverage of rules found from random data

| Treatment | — support — | | | — confidence — | | | — leverage — | | |
	min	mean	max	min	mean	max	min	mean	max
Non-redundant	320	950	2,688	0.490	0.537	0.618	0.0044	0.0050	0.0116
Productive	320	950	2,688	0.490	0.537	0.618	0.0044	0.0050	0.0116
Significance=0.05	320	860	2,688	0.489	0.537	0.618	0.0042	0.0050	0.0116

and productive approaches because almost all rules with the highest leverage were productive.

These results appear to support an affirmative answer to question 1. It seems clear that there is a strong risk of false discoveries unless appropriate allowance is made for the multiple-tests problem.

No rules were found for any dataset under the Bonferroni or either holdout treatment. It might come as a surprise that no rules were found under any of these treatments, whereas the adjustments are supposed to restrict the number of analyses for which any false discoveries are made to no more than 5% and hence one might have expected rules to have been found for up to 5 of the datasets under each of these treatments. With the Bonferroni adjustment this is possibly not so surprising as it is much more strict than the Holm procedure which also guarantees strict control over the experimentwise risk of false discoveries, and hence must be less likely to make false discoveries. The reason the Holm procedure makes no false discoveries may relate to the inter-dependencies between the rules. Both the Holm and Bonferroni procedures control against the most disadvantageous form of relationship between the hypothesis tests which occurs when the null hypotheses are mutually exclusive. In practice, however, many of the hypothesis tests for this rule discovery task will be closely related to one another, and may even be equivalent, as for example for the rules *pregnant → oedema* and *oedema → pregnant*. In this circumstance the probability of any false discovery occurring is greatly reduced, although if one occurs it is likely several will occur.

These results provide support for an affirmative answer for question 2. Both the within-search and holdout approaches can control the risks of false discoveries.

6.2 Experiment 2

The second experiment sought to investigate question 3. Random data were generated comprising 10,000 transactions each containing values for 20 binary variables X55, Y55, X60, Y60, X65, Y65, X70, Y70, X75, Y75, X80, Y80, X85, Y85, X90, Y90, X95, Y95, X100, Y100. Each of the X values was randomly generated with each value being equiprobable. The probability of $Yv=1$ was v% if $Xv=1$, otherwise 100-v% and the probability of $Yv=0$ was v% if $Xv=0$, otherwise 100-v%. For example, the probability of Y55=1 was 55% if X55=1, otherwise 45%. These data give rise to 40 true discoveries, $X55 = 0 → Y55 = 0$, $X55 = 1 → Y55 = 1$, $Y55 = 0 → X55 = 0$, $Y55 = 1 → X55 = 1$ and so on. Any other rules found represent false discoveries. The varying confidence levels of the rules (from 0.55 through to 1.00) represent different levels of challenge to a discovery system. 100 such random datasets were generated and all six treatments used in the previous experiment were applied to each.

All treatments found all true rules relating to X and Y65 and higher for all data sets. Only significance=0.05

and Bonferroni found the remaining 8 rules for any dataset and they consistently did so for all datasets. All of non-redundant, productive and significance=0.05 found as many false discoveries as required to fill a quota of 1000 rules. The reason that non-redundant and productive failed to find the 55 and 60 level rules was that they found so many higher leverage rules that these true rules did not fit within the quota of 1000.

Neither the Bonferroni nor either of the holdout treatments made any false discoveries. The reason the holdout treatments did not find the 55 or 60 level rules was that those rules were not found during the exploratory rule discovery stage. The highest p-value for any of the rules found under the Bonferroni treatment was 6.19×10^{-26}. This would have been rejected if the search space had contained more than 8.095×10^{23} rules, which would occur, for example, if there were 100 pairs of variables and x were allowed to contain up to 15 values (rule space size = 5.79×10^{27}, adjusted $p = 1.73 \times 10^{-30}$.

These results show that the relative power of the within-search and holdout approaches will vary depending upon the size of the search space and on the capacity under holdout evaluation of the search technique applied during the exploratory stage to find the true discoveries.

6.3 Experiment 3

The final experiment investigates question 4, how do the techniques perform on real-world data. The same six treatments were used as for the previous experiments. Experiments were conducted using eight of the largest attribute-value datasets from the UCI machine learning [21] and KDD [13] repositories together with the BMS-WebView-1 [30] and Retail [9] datasets. These datasets are described in Table 2. We first found for each dataset the minimum even value for minimum-support that results in fewer than 10,000 non-redundant rules given a minimum-confidence of 0.75. This is listed in the minsup column of Table 2. Each treatment was then applied to each dataset five times, once with each maximum limit *maxx* on the size of x from 1 to 5. All runs used minimum-confidence=0.75 and the appropriate minimum-support, except for the holdout treatments which only use half the data for rule discovery and for which the minimum-support was therefore halved.

Table 3 presents the number of rules found by each technique for each data set and setting of *maxx*. The meanings of the columns are as follows:

Dataset: The dataset.

maxx: The maximum number of items allowed in x.

NR: The number of non-redundant rules 'discovered.'

Prod: The number of productive rules 'discovered.'

0.05: The number of rules 'discovered' that passed an unadjusted significance test at the 0.05 level.

Within-Search Rule Space: The number of rules in the search space. The within-search technique used a significance level of 0.05 divided by this value.

Table 2: Data sets

Dataset	Transactions/Records	Items	Minsup	Description
BMS-WebView-1	59,602	497	50	E-commerce clickstream data
Covtype	581,012	125	577,684	Geographic forest vegetation data
IPUMS LA 99	88,443	1,883	54,660	Census data
KDDCup98	52,256	4,244	47,748	Mailing list profitability data
Letter Recognition	20,000	74	1,186	Image recognition data
Mush	8,124	127	1,730	Biological data
Retail	88,162	16,470	28	Retail market-basket data
Shuttle	58,000	34	1,418	Space shuttle mission data
Splice Junction	3,177	243	152	Gene sequence data
TICDATA 2000	5,822	709	5,730	Insurance policy holder data proc

Table 3: Number of rules found under each treatment

Dataset	$|x|$	NR	Prod	0.05	Within-Search		Holdout-NR		Holdout-Prod		WS$\overline{HP}$	$\overline{WS}$HP
					Rule Space	Disc	Cand	Disc	Cand	Disc		
BMS-WebView-1	1	3	3	3	2.47×10^5	3	3	3	3	3	0	0
BMS-WebView-1	2	1075	1075	1068	6.13×10^7	941	2398	2076	2397	2076	146	1281
BMS-WebView-1	3	4530	4496	4084	1.01×10^{10}	841	12873	1839	12550	1846	170	1175
BMS-WebView-1	4	6600	6210	4885	1.25×10^{12}	721	20877	1755	17840	1781	116	1176
BMS-WebView-1	5	7173	6414	4915	1.23×10^{14}	594	22135	1747	18179	1780	77	1263
Covtype	1	234	0	0	1.78×10^4	0	236	0	0	0	0	0
Covtype	2	1380	0	0	1.17×10^6	0	1379	0	0	0	0	0
Covtype	3	4096	0	0	4.98×10^7	0	4103	0	0	0	0	0
Covtype	4	7576	0	0	1.55×10^9	0	7588	0	0	0	0	0
Covtype	5	9988	0	0	3.80×10^{10}	0	10018	0	0	0	0	0
IPUMS LA 99	1	329	156	143	3.08×10^6	133	329	136	157	136	1	4
IPUMS LA 99	2	1992	401	323	2.23×10^9	263	1962	261	412	266	3	6
IPUMS LA 99	3	5312	542	377	9.60×10^{11}	288	5209	290	586	297	3	12
IPUMS LA 99	4	8447	581	383	2.77×10^{14}	286	8315	290	655	297	3	14
IPUMS LA 99	5	9990	582	383	5.73×10^{16}	282	9857	288	668	297	3	18
KDDCup98	1	265	138	50	1.50×10^8	32	269	34	128	34	0	2
KDDCup98	2	1580	399	80	4.39×10^{11}	33	1582	38	363	41	0	8
KDDCup98	3	4646	715	105	7.49×10^{14}	30	4608	40	641	42	0	12
KDDCup98	4	8157	906	120	8.76×10^{17}	30	8037	40	795	40	0	10
KDDCup98	5	9984	958	124	7.66×10^{20}	30	9863	40	854	40	0	10
Letter Recognition	1	20	20	20	4.66×10^3	20	17	17	17	17	4	1
Letter Recognition	2	693	554	509	1.34×10^5	416	595	359	473	366	126	76
Letter Recognition	3	3696	2265	1810	2.30×10^6	939	3285	828	2055	852	390	303
Letter Recognition	4	7826	3861	2739	2.68×10^7	1029	6607	958	3454	998	452	421
Letter Recognition	5	9950	4455	2981	2.27×10^8	975	7978	952	3842	1005	422	452
Mush	1	223	142	135	1.52×10^4	126	223	126	137	127	1	2
Mush	2	1674	630	601	8.74×10^5	516	1674	494	630	519	25	28
Mush	3	4885	1292	1224	3.12×10^7	1041	4832	963	1261	1027	73	59
Mush	4	7751	1665	1581	7.85×10^8	1302	7657	1221	1611	1303	95	96
Mush	5	9070	1750	1662	1.48×10^{10}	1317	8957	1266	1682	1351	100	134
Retail	1	263	263	263	2.71×10^8	65	447	85	447	85	8	28
Retail	2	4722	4538	2613	2.23×10^{12}	62	6052	107	5853	107	5	50
Retail	3	8808	7688	3280	1.23×10^{16}	55	10641	98	9220	102	4	51
Retail	4	9723	8149	3315	5.05×10^{19}	52	11556	98	9653	101	4	53
Retail	5	9887	8199	3315	1.66×10^{23}	51	11656	97	9675	101	4	54
Shuttle	1	49	44	44	1.03×10^3	43	48	43	44	43	1	1
Shuttle	2	745	529	479	1.46×10^4	407	750	396	530	403	14	10
Shuttle	3	3910	1823	1322	1.19×10^5	878	3895	795	1809	819	82	23
Shuttle	4	8260	2529	1713	6.29×10^5	985	8138	877	2457	923	94	32
Shuttle	5	9999	2630	1771	2.28×10^6	964	9760	876	2555	929	77	42
Splice Junction	1	17	17	17	5.81×10^4	17	16	16	16	16	1	0
Splice Junction	2	1545	1045	394	6.88×10^6	62	1535	62	763	70	6	14
Splice Junction	3	7354	4524	1894	5.33×10^8	120	7067	134	4220	145	20	45
Splice Junction	4	9261	5399	2241	3.04×10^{10}	92	8758	133	5135	148	10	66
Splice Junction	5	9463	5465	2251	1.36×10^{12}	75	8937	132	5208	148	4	77
TICDATA 2000	1	368	160	136	4.67×10^5	32	366	30	110	46	2	16
TICDATA 2000	2	2200	328	248	1.56×10^8	24	2246	30	166	46	2	24
TICDATA 2000	3	5736	392	280	3.42×10^{10}	22	6150	30	166	46	2	26
TICDATA 2000	4	8584	392	280	5.53×10^{12}	22	9478	30	166	46	2	26
TICDATA 2000	5	9384	392	280	7.02×10^{14}	22	10438	30	166	46	2	26

Within-Search Disc: The number of rules 'discovered' that passed the adjusted significance test. This is abbreviated as WS, below.

Holdout-NR Cand: The number of non-redundant candidate rules generated from the exploratory data under the holdout approach.

Holdout-NR Disc: The number of those candidate rules that passed the subsequent holdout evaluation. This is abbreviated as HN, below.

Holdout-Prod Cand: The number of productive candidate rules generated from the exploratory data under the holdout approach.

Holdout-Prod Disc: The number of those candidate rules that passed the subsequent holdout evaluation. This is abbreviated as HP, below.

WS$\overline{\text{HP}}$: The number of rules 'discovered' by the within-search technique but not passed by holdout evaluation on productive rules.

$\overline{\text{WS}}$HP: The number of rules passed by holdout evaluation on productive rules but not 'discovered' by the within-search technique.

The relative numbers of rules discovered for each dataset and $maxx$ by within-search adjusted significance tests and by each of the holdout evaluation techniques are plotted in Figure 2.

6.3.1 Observations

A number of points are worth highlighting. First, the number of non-redundant rules that are not productive and the number of productive rules that do not pass an unadjusted significance test is in many cases extremely high. In the most extreme case, Covtype, none of the non-redundant rules is productive. This is due to a peculiarity of this particular dataset which uses 40 mutually exclusive binary variables **ST01** to **ST40** to represent which one of 40 soil types predominates in an area. Thus, the most frequent attribute-values are values of 0 for individual **ST??** variables and the most frequent itemsets are sets of these values. Because they are mutually exclusive, for any two of these variables w and z, $P(w=0 \mid z=1) = 1.0$. It follows that, so long as $P(z=1) > 0.0$, $P(w=0 \mid z=0) < P(w=0)$. Hence, all associations between these variables must be unproductive. The fact that all the top 9,988 non-redundant associations for this dataset represent *negative* associations highlights the dangers of data mining without both a clear definition of what constitutes a false discovery and sound mechanisms for detecting and rejecting such false discoveries. (Indeed, it turns out that all of the 197,183,686 highest support associations for this data are associations between these variables and hence negative associations.)

Holdout evaluation with productive rules usually finds slightly more rules than Holdout evaluation with non-redundant rules. This is because the size of the correction for multiple tests that is performed during holdout evaluation is smaller, as there are fewer productive than non-redundant rules. The only circumstance in which the use of non-redundant rules could result in more discoveries is if some rules that were unproductive on the exploratory data turned out to be significantly productive on the holdout data. This rarely occurred in the experiments, a total of four rules being discovered by holdout with non-redundant but not holdout with productive. Only one of these rules is productive with respect to the full data, having by chance

proved unproductive with respect to the exploratory sample. The remaining three rules were unproductive with respect to both the full data and the exploratory data, but by chance turned out to be productive on the holdout data. This illustrates a potential flaw in the holdout method. Consider random data such as that used in Experiment 1, above, where a dataset is divided into exploratory and holdout sets. If one were to take the rules that represented the strongest negative correlations with respect to the exploratory data, the probability of those rules representing positive correlations on the holdout data would be increased, as one would be actively selecting for rules for which, by chance, more of the transactions containing both the antecedent and consequent were selected into the holdout data than the training data. It is clearly necessary to avoid any such confounding selection process under the holdout evaluation strategy.

Overall, holdout evaluation with productive rules found more rules than within-search significance testing, with the relative performance being more favorable to within-search when the size of the rule space was smaller (small $maxx$ or fewer items) and more favorable to holdout evaluation as the size of the search space increased. This is because of the extremely small significance levels that are employed with large search spaces. The total number of rules discovered by WS often decreased as the search space increased. Such decreases occurred less for holdout evaluation and when they did, the decreases were smaller. Note, however, that these relative results are in part an artifact of the experimental design. If minimum support had been set lower, then WS would have found more rules, as the rules found under the current setting would not have been affected and further rules with lower support could have potentially passed the adjusted significance test. However, increasing minimum support could have reduced the number of rules found by holdout evaluation as it would have increased the number of candidate rules and hence lowered the adjusted significance level that rules had to pass.

It is interesting to consider the causes of holdout evaluation failing to find a rule that a within-search correction finds. WS made 16,238 discoveries, where this is the sum of the WS column of Table 3, and multiply counts some rules because a single rule may be discovered repeatedly at different settings of $maxx$. Of these, 2,554 were not discovered by HP, as shown in column WS$\overline{\text{HP}}$ of Table 3. In 533 cases the rule was found during the exploratory stage but failed holdout evaluation. In the remaining 2,021 cases the rule was not found during the exploratory stage. This latter case occurred primarily because rules failed to attain the required minimum support with respect to the smaller sample of data available for the initial rule discovery stage when holdout evaluation is performed.

It is also interesting to observe that in some cases huge numbers of additional rules were found during the exploratory stage relative to those found from the full data set, the most extreme case being BMS-Webview-1, for which approximately three times the number of candidate rules were found relative to discovery from the full dataset. This illustrates the disadvantage of working with the smaller samples inherent in the holdout approach.

7. FUTURE RESEARCH

The effectiveness of holdout evaluation with productive rules relative to that of holdout evaluation with non-
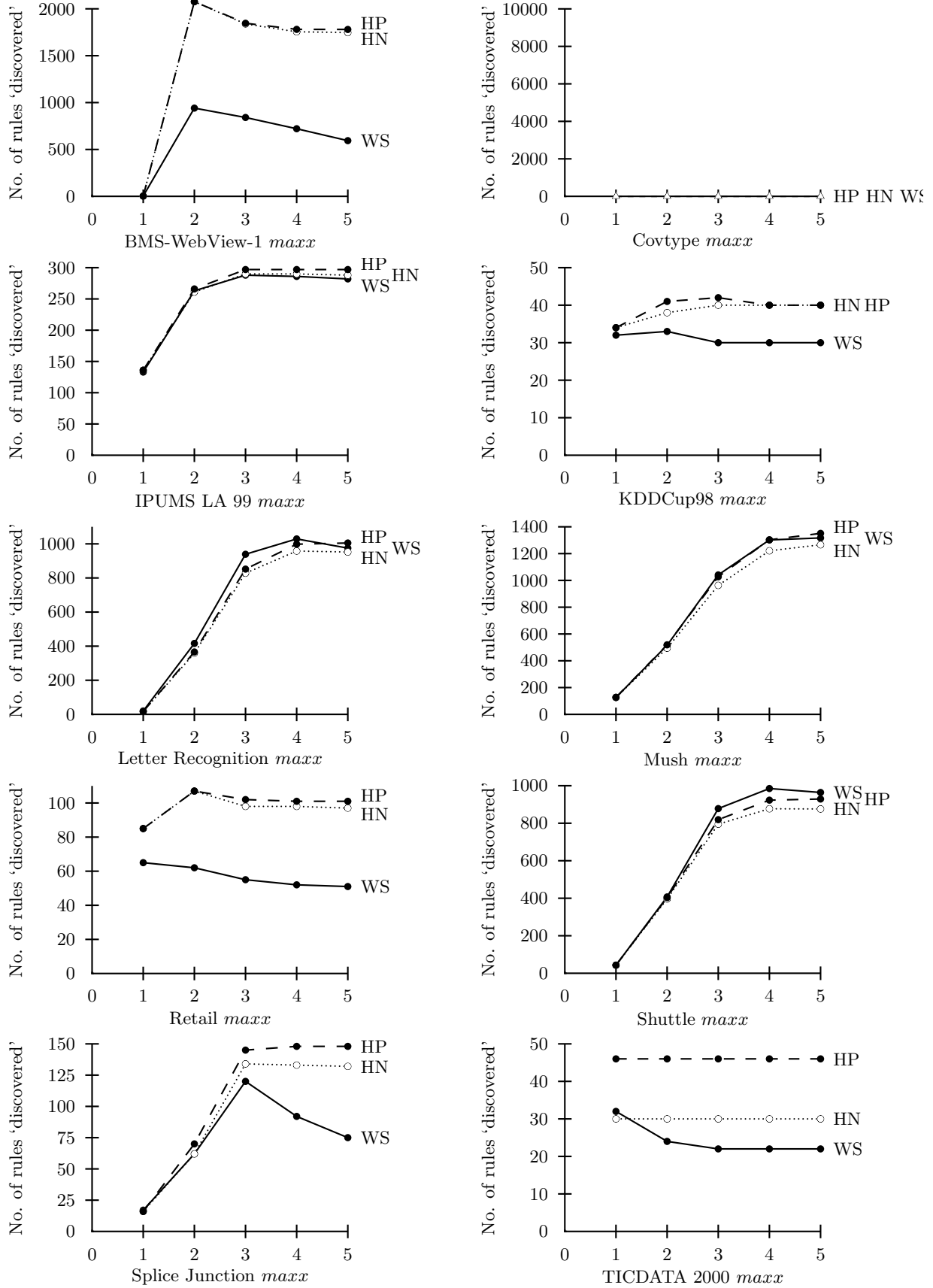
Figure 2: Numbers of rules 'discovered'

redundant rules demonstrates the value of excluding from holdout evaluation rules that are unlikely to pass. On the other hand, however, the small number of rules that were found using non-redundant rather than productive rules demonstrates the danger of performing too strong a filter on the rules to be subjected to holdout evaluation. The development of effective techniques for performing such filtering provides a promising direction for future investigation.

In the current within-search approach, the adjustment to the significance level takes account of the size of the entire rule space under consideration. It takes no account of the fact that the number of candidates must vary as user specified constraints such as minimum-support are varied. It seems credible that the adjustment should get smaller as minimum-support gets higher, as fewer significance tests will be applied. However, it would be incorrect to simply adjust for the number of rules that pass a minimum support constraint. To understand why this is so, consider the example of random data in Experiment 1, above, which consists of 10,000 transactions each containing values for 100 binary variables with each value being equiprobable. For one such random dataset, 12 non-redundant rules were found with support greater than 2600. If a significance level of 0.05 were adjusted to allow for 12 tests, the resulting adjusted significance level would be 0.00417. Two of the twelve rules would pass a Fisher exact test at this significance level, both obtaining p-values of 0.00112. The reason that it is not appropriate to adjust the significance test only for the rules that pass a minimum-support constraint is that sampling effects will exaggerate the support of some rules and the number of such exaggerated support counts experienced relates to the total number of rules in the global rule space, not just to the number that pass the minimum support constraint. It would be valuable if sound techniques could be developed for modifying adjustments in line with relaxation or strengthening of constraints.

It is possible to vary the size of the rule space for the within search strategy by varying constraints such as $maxx$, or by deleting items from consideration. It is clear that doing so will alter the number of discoveries that are made, and that in some circumstances increasing the size of the rule space will decrease the number of discoveries. It would be valuable to develop techniques for selecting constraints that will maximize the expected number of discoveries. One approach might be to explore a variety of constraints and then select the one that delivers the most discoveries, but care would need to be taken that the selection criteria were in no way selecting the settings that were most likely to have resulted in false discoveries, because doing so could defeat the current strict control over the risk of false discoveries. It is possible that selecting the settings that deliver the most rules might favor settings that produce false discoveries, as any false discoveries produced will increase the number of discoveries.

The current research has adopted the association rules support-confidence framework. The ability to assess significance during search holds open the prospect of a new form of rule discovery, in which all statistically significant rules are discovered without the need to specify any arbitrary parameter other than the risk of false discoveries (the global significance level). As search through the massive search spaces involved in rule discovery relies upon efficient pruning from consideration of the majority of the search space, the development of such techniques will depend critically upon whether effective techniques can be developed for identifying sections of the search space that cannot contain rules that could pass a significance test. This is a further intriguing direction for future research.

A somewhat related line of research to that pursued herein involves the discovery of condensed representations [19, 8, 11] for datamining. Like the current research, this is concerned with reducing the number of discoveries. However, while we seek here to limit discoveries to those that represent non-trivial interactions between variables, condensed representations seek to limit discoveries to a limited set from which all other discoveries can be inferred. It would be useful to develop condensed representations for significant rules, thereby coupling the representational power of condensed representations with the ability to avoid false discoveries.

8. CONCLUSIONS

This research has demonstrated the effectiveness of two alternative techniques for controlling the risk of false discoveries during association rule discovery. Within-search significance tests adjust the significance level to allow for the size of the search space from which rules have been drawn. Despite the resulting use of extremely low significance levels (in our experiments as low as 3×10^{-25}), in some cases large numbers of rules can be discovered while providing strict control over the risk of false discoveries. Holdout evaluation can also provide strict control over the risk of false discoveries by first discovering candidate rules from an exploratory dataset and then testing those rules on different holdout data.

Our experimental evaluation has demonstrated that neither of these approaches clearly dominates the other. While in our experiments the holdout approach made substantially more discoveries overall, it is clear that the within-search approach holds an advantage when the the size of the search-space is small or when the number of candidate rules is large.

Both techniques provide considerable flexibility. Most importantly, while we have examined in this paper only techniques for discarding unproductive rules, the techniques are directly applicable to any statistical hypothesis test, allowing users to identify and discard false discoveries using whatever definition of false discovery is applicable to their specific application.

The problem of false discoveries appears to be a serious one, as evidenced by the large numbers of non-redundant rules that are not even productive on the sample data, let alone in the domain from which the data have been sampled. The Covtype dataset is real-world data for which all of the 197,183,686 highest support rules are unproductive. Experiment 1 demonstrated that it is possible to 'discover' large numbers of association rules from totally random data. These examples starkly illustrate that there is a serious risk of discovering large numbers of spurious rules if we do not perform appropriate statistical evaluation during association rule discovery that takes account of the size of the search space that is explored. We, the members of the data mining community, are doing a serious disservice to ourselves, as well as to the communities we seek to serve, if we present sets of 'discoveries' to our clients of which the majority are spurious.

10. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining associations between sets of items in massive databases. In *Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data*, pages 207–216, Washington, DC, May 1993.

[2] A. Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153, February 1992.

[3] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In *First Int. Conf. Computational Logic - CL 2000*, pages 972–986, Berlin, 2000. Springer-Verlag.

[4] S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

[5] R. J. Bayardo, Jr., R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4(2/3):217–240, 2000.

[6] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A new and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300, 1995.

[7] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[8] J.-F. Boulicaut and A. Bykowski. Frequent closures as a concise representation for binary data mining. In *Proc. Pacific-Asia Conf. Knowledge Discovery and Datamining (PAKDD-2000)*, pages 62–73, Kyoto, Japan, 2000.

[9] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using association rules for product assortment decisions: A case study. In *Knowledge Discovery and Data Mining*, pages 254–260, 1999.

[10] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In J. Peckham, editor, *SIGMOD 1997, Proc. ACM SIGMOD Int. Conf. Management of Data*, pages 265–276, New York, NY, May 1997. ACM Press.

[11] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. In *Proc. 6th European Conf. Principles and Practice of Knowledge Discovery in Databases, PKDD 2002*, pages 74–85, Berlin, 2002. Springer.

[12] W. DuMouchel and D. Pregibon. Empirical Bayes screening for multi-item associations. In *KDD-2001: Proc. Seventh ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pages 76–76, New York, NY, August 2001. ACM Press.

[13] S. Hettich and S. D. Bay. The UCI KDD archive. [http://kdd.ics.uci.edu] Irvine, CA: University of California, Department of Information and Computer Science., 2006.

[14] B. Holland and C. M. D. Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin*, 104(1):145–149, 1988.

[15] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[16] International Business Machines. IBM intelligent miner user's guide, version 1, release 1, 1996.

[17] C. Jermaine. Finding the most interesting correlations in a database: How hard can it be? *Information Systems*, 30(1):21–46, 2005.

[18] B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proc. Fifth ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD-99)*, pages 125–134, New York, August 1999. AAAI.

[19] H. Mannila and H. Toivonen. Multiple uses of frequent sets and condensed representations. In *Proc. Int. Conf. Knowledge Discovery and Data Mining KDD'96*, pages 189–194, Portland, USA, 1996.

[20] N. Megiddo and R. Srikant. Discovering predictive association rules. In *Proc. Fourth Int. Conf. Knowledge Discovery and Data Mining (KDD-98)*, pages 27–78, Menlo Park, US, 1998. AAAI Press.

[21] D. J. Newman, S. Hettich, C. Blake, and C. J. Merz. UCI repository of machine learning databases. [Machine-readable data repository]. University of California, Department of Information and Computer Science, Irvine, CA., 2006.

[22] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, Menlo Park, CA., 1991.

[23] T. Scheffer. Finding association rules that trade support optimally against confidence. *Intelligent Data Analysis*, 9(4):381 – 395, 1995.

[24] J. P. Shaffer. Multiple hypothesis testing. *Ann. Rev. Psychology*, 46:561–584, 1995.

[25] G. I. Webb. Magnum Opus Version 1.3. Software, G. I. Webb & Associates, Melbourne, Aust., 2002.

[26] G. I. Webb. Magnum Opus Version 3.1. Software, G. I. Webb & Associates, Melbourne, Aust., 2006.

[27] G. I. Webb and S. Zhang. K-optimal rule discovery. *Data Mining and Knowledge Discovery*, 10(1):39–79, 2005.

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

[29] H. Zhang, B. Padmanabhan, and A. Tuzhilin. On the discovery of significant statistical quantitative rules. In *Proc. Tenth Int. Conf. Knowledge Discovery and Data Mining (KDD-2004)*, pages 374–383, New York, NY, August 2004. ACM Press.

[30] Z. Zheng, R. Kohavi, and L. Mason. Real world performance of association rule algorithms. In *Proc. Seventh Int. Conf. Knowledge Discovery and Data Mining (KDD-2001)*, pages 401–406, New York, NY, August 2001. ACM.