Mach Learn (2008) 71: 307–323
DOI 10.1007/s10994-008-5046-x

TECHNICAL NOTE

Layered critical values: a powerful direct-adjustment approach to discovering significant patterns

Geoffrey I. Webb

Received: 15 July 2007 / Revised: 6 February 2008 / Accepted: 7 February 2008 / Published online: 8 March 2008
Springer Science+Business Media, LLC 2008

Abstract Standard pattern discovery techniques, such as association rules, suffer an extreme risk of finding very large numbers of spurious patterns for many knowledge discovery tasks. The direct-adjustment approach to controlling this risk applies a statistical test during the discovery process, using a critical value adjusted to take account of the size of the search space. However, a problem with the direct-adjustment strategy is that it may discard numerous true patterns. This paper investigates the assignment of different critical values to different areas of the search space as an approach to alleviating this problem, using a variant of a technique originally developed for other purposes. This approach is shown to be effective at increasing the number of discoveries while still maintaining strict control over the risk of false discoveries.

Keywords Pattern discovery · Significant patterns · Significant rules · Layered critical values · Association rules · Statistical procedures

1 Introduction

The current paper presents a technique that is demonstrated to often substantially increase the power of the direct-adjustment approach to controlling the risk of false discoveries in pattern discovery. Relative to the previous state-of-the-art, the new technique increases the number of significant patterns discovered while still maintaining strict control over the risk of false discoveries.

Pattern discovery finds collections of items that co-occur frequently in data. This class of data mining techniques includes *association rule discovery* (Agrawal et al. 1993), *k-optimal* or *top-k pattern discovery* (Webb 1995; Scheffer and Wrobel 2002; Han et al. 2002; Webb and Zhang 2005), *contrast or emerging pattern discovery* (Bay and Pazzani 2001; Dong and

Editor: Johannes Fuernkranz

G.I. Webb (✉)
Faculty of Information Technology, Monash University, Clayton Campus, Wellington Road, Clayton, Vic, Australia
e-mail: geoff.webb@infotech.monash.edu

Li 1999), *subgroup discovery* (Klösgen 1996), *interesting itemset discovery* (Jaroszewicz and Simovici 2004) and *impact* or *quantitative rule discovery* (Aumann and Lindell 1999; Webb 2001; Zhang et al. 2004).

For many applications, such patterns will only be interesting if they represent non-trivial correlations between all constituent items. We will call such patterns *significant patterns* and all remaining patterns *false discoveries*. Many techniques have been developed that seek to avoid false discoveries (Agrawal et al. 1993; Bastide et al. 2000; Bay and Pazzani 2001; Bayardo et al. 2000; Brin et al. 1997; DuMouchel and Pregibon 2001; Gionis et al. 2006; International Business Machines 1996; Liu et al. 1999; Megiddo and Srikant 1998; Možina et al. 2006; Piatetsky-Shapiro 1991; Scheffer 1995; Webb 2002, 2006, 2007; Zaki 2004; Zhang et al. 2004). Two of these, *direct-adjustment* and *holdout evaluation*, have the desirable properties of allowing 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. Previous research has shown that each of these two approaches has relative strengths and weaknesses (Webb 2006, 2007). Of the two, only the direct-adjustment approach is directly applicable to *k-optimal pattern discovery*.

This paper presents an improvement to the direct-adjustment approach. Section 2 describes the false discovery problem. This is an expanded version of the problem statement in (Webb 2007), included here to make the paper self-contained. Section 3 describes the direct-adjustment and holdout evaluation approaches. Section 4 describes the new *Layered Critical Values* extension to the direct-adjustment approach. Section 5 describes a series of experiments to assess the effect of the new technique on the power of the direct-adjustment approach. We end with concluding remarks.

2 Problem statement

Pattern discovery seeks to identify patterns $\rho \in \mathcal{P}$ that satisfy constraints ϕ with respect to distribution Θ. However, whether ϕ is satisfied is assessed by reference to sample data D drawn from Θ. Although the principles extend directly to further contexts, the current research limits consideration to two types of data, *transactional data* and *attribute-value data*, and one type of pattern, *rules*.

For both data types, D is a multiset of n records and each record $R \in D$ is a set of items $R \subseteq I$. For transactional data, items are atomic terms. For attribute-value data, there exists a set of a attributes $A_1 \dots A_a$, each attribute A_i has a domain of $\#A_i$ values $\mathrm{dom}(A_i)$, each item is an attribute-value pair denoted as $A_i{=}v$, where $v \in \mathrm{dom}(A_i)$, and each record $R \in D$ contains exactly one item for each attribute.

In the current work, rules take the form $X \rightarrow y$, where $X \subseteq I$, $|X| \geq 1$ and $y \in I$. X is called the *antecedent* and y the *consequent* of the rule. For attribute-value data, $X \cup \{y\}$ may contain no more than one item for any one attribute. While some rule discovery systems support multiple elements in the consequent of a rule, such rules can be represented by multiple single-element consequent rules, and limiting the search space to the latter greatly decreases its size. As limiting the size of the search space is important for the current work, we limit rules to single-element consequents.

Association rule discovery finds all rules that satisfy constraints ϕ specified as a minimum *support* (Agrawal et al. 1993), together with other constraints, if desired, such as minimum *confidence* (Agrawal et al. 1993). Support and confidence are defined with respect to a rule $X \rightarrow y$ and dataset D as follows:

$$support(X \rightarrow y) = |\{R \in D : X \cup \{y\} \subseteq R\}|, \tag{1}$$

$$confidence(X \to y) = support(X \to y)/|\{R \in D : X \subseteq R\}|. \tag{2}$$

Confidence can be viewed as a maximum likelihood estimate of the *true confidence*, $P_\Theta(y \mid X)$.

Each assessment of whether a given pattern ρ satisfies constraints ϕ is accompanied by a risk that ρ will satisfy ϕ with respect to the sample data D but not with respect to Θ. Most pattern discovery systems fail to effectively control this risk.

Statistical hypothesis tests are applicable to such a scenario. To apply such a test it is necessary to specify a *null hypothesis*, in our context the hypothesis that the negation of ϕ is true. The test returns a value p, an upper bound on the probability that the sample data, or sample data with a more extreme distribution, would be observed if the null hypothesis were true. The value p is compared to a significance-level α, a user-specified upper limit on the allowed risk of rejecting a null hypothesis if it is false. If $p \leq \alpha$ the null hypothesis is rejected and the alternate hypothesis accepted. In the case of pattern discovery, this means accepting the pattern as valid and hence "discovering" it.

While the generic techniques examined support arbitrary statistical hypothesis tests, the current research considers only tests for *productive rules*. A rule is productive if it has higher confidence than all of its generalizations:

$$productive(X \to y) = \forall Z \subset X,\ confidence(X \to y) > confidence(Z \to y). \tag{3}$$

We use a Fisher exact test for productive rules, as described in Appendix 1 of (Webb 2007). As this seeks to assess whether the patterns are productive with respect to Θ rather than simply with respect to D, the null hypothesis is

$$\forall Z \subset X,\ P(y \mid X) \leq P(y \mid Z). \tag{4}$$

If the discovery process "discovers" a pattern ρ that in actuality satisfies the null hypothesis, ρ is considered to be a *false discovery* or equivalently, a *type*-1 *error*. Any pattern ρ that is not "discovered" and does not satisfy the null hypothesis is called a *type*-2 *error*.

The techniques presented herein allow arbitrary statistical hypothesis tests to be applied in a manner that allows the user to place a strict upper bound on the *experimentwise risk of false discoveries*. In the context of pattern discovery, this is the risk of any pattern from those found in a single session being a false discovery.

3 The direct-adjustment and holdout approaches

A standard statistical solution to the multiple tests problem is to use an adjustment such as the well-known Bonferroni adjustment. These control the experimentwise risk of false discoveries (Holland and Copenhaver 1988) by adjusting the critical value employed with the statistical test to allow for the number of hypotheses tested.

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 strictly controls the experimentwise risk of false discoveries, even if the hypothesis tests are correlated with one another, as is often the case in the context of pattern discovery.

The *direct-adjustment* approach applies the appropriate Bonferroni adjustment directly to any statistical test employed during the search process. This requires an upper bound on the number of hypothesis tests in the search space. Webb (2007) describes how this may be calculated in a context where there is an upper limit on the size of the antecedent.

Rather than applying statistical tests during the pattern discovery process, the *holdout* approach partitions the available data into exploratory and holdout sets; discovers candidate patterns using the exploratory data; and then tests those patterns using the holdout data. It accepts only those patterns that pass relevant statistical tests for significance. It is necessary to correct for multiple tests, but only with respect to the number of patterns found in the exploratory stage, not the full size of the search space considered. As the former is likely to be many magnitudes smaller than the latter, the adjustment will be substantially smaller.

Holdout evaluation can be applied as a simple wrapper to any existing pattern discovery system. In contrast, the direct-adjustment approach may require substantial re-engineering of a system. The holdout approach is less susceptible to decreases in its power as a result of increases in the size of the search space, can utilize more powerful corrections for multiple tests such as the Holm procedure (Holm 1979), and can support procedures to control the false discovery rate (Benjamini and Hochberg 1995) as well as the experimentwise error rate. Further, the actual number of tests that must be applied will often be orders of magnitude lower than under the direct-adjustment approach, providing a considerable computational advantage when employing computationally demanding statistical tests. On the other hand, only the direct-adjustment approach directly supports k-optimal (also known as top-k) pattern discovery and it also utilizes all available data for both pattern detection and pattern evaluation.

4 Layered critical values

It is desirable to increase the power of these techniques, the probability that each will discover valid patterns. Webb (2007) identified two competing forces that affect the power of the direct-adjustment approach as the size of the search space is altered. As the search space is increased there is an upwards pressure on power. When more patterns are considered, more potentially valid patterns are available to be discovered. On the other hand, as the search space increases, the size of the adjustment r increases proportionally, decreasing the critical value α' employed in the statistical test. This requires patterns to be stronger in order to be discovered. As a result there is a downward pressure on power, as weaker patterns that would be discovered in a smaller search space will no longer pass the statistical test at the lower critical value.

In practice, the two counteracting pressures on power result in the number of patterns discovered initially increasing as the size of the allowed antecedent is increased, reaching a peak and then declining as the numbers of patterns discovered at lower antecedent sizes but no longer accepted with a lower critical value exceeds the number of additional valid patterns encountered that can pass the low critical value.

In the existing direct-adjustment techniques, the size of the search space is altered by changing the number of items allowed in the antecedent of a rule. This is illustrated in Table 1, which shows the rules in the search space given that there are four items. At each of antecedent sizes 1 and 2 there are twelve rules and at antecedent size 3 there are four rules. With larger numbers of items, the size is initially relatively small for antecedents of size 1 and rapidly grows to very large numbers. For example, with 50 items there are 2,450 rules with one element in the antecedent, 58,800 with two elements, 921,200 with three elements and over 10 million with four elements.

It is useful to consider *candidate patterns*. These are patterns that pass standard selection criteria, such as minimum support, that may result in patterns being discarded before being subjected to a statistical test. It is notable that, for real-world data and standard pattern selection criteria, the proportion of the search space that consists of candidate patterns

Table 1 Search space at each level given four items

Antecedent size		
1	2	3
$a \rightarrow b$	$a, b \rightarrow c$	$a, b, c \rightarrow d$
$a \rightarrow c$	$a, b \rightarrow d$	$a, b, d \rightarrow c$
$a \rightarrow d$	$a, c \rightarrow b$	$a, c, d \rightarrow b$
$b \rightarrow a$	$a, c \rightarrow d$	$b, c, d \rightarrow a$
$b \rightarrow c$	$b, c \rightarrow a$	
$b \rightarrow d$	$b, c \rightarrow d$	
$c \rightarrow a$	$a, d \rightarrow b$	
$c \rightarrow b$	$a, d \rightarrow c$	
$c \rightarrow d$	$b, d \rightarrow a$	
$d \rightarrow a$	$b, d \rightarrow c$	
$d \rightarrow b$	$c, d \rightarrow a$	
$d \rightarrow c$	$c, d \rightarrow b$	

tends to decrease as the size of the antecedent increases. One reason for this is that support is anti-monotone on antecedent size. Most selection criteria will discard patterns with zero support and the proportion of zero support patterns cannot decrease as the antecedent size increases. Table 2 illustrates this phenomena with respect to the experiments on real-world data in (Webb 2007). In these experiments the minimum even-valued setting of minimum support was found that resulted in fewer than 10,000 productive rules with antecedent sizes of no more than six. Table 2 shows for each dataset—the size of the search space at each antecedent size n; the number of productive rules that satisfy the minimum support constraint for the dataset; the density of candidates within the search space at the level (*cand*/*size*); the adjusted critical value that is employed when the search space allows antecedents of size no more then n; and the change in the number of patterns discovered relative to a search limited to antecedents of size no more than $n - 1$.

For all of these data sets, as the antecedent size increases the density decreases and the critical value increases rapidly. In some sense, the higher levels of the search space are the richest, containing the highest density of candidates. They also have the smallest search spaces and hence receive the least strict critical values. As the search space is increased to include ever less dense levels, the critical value applied to the denser higher levels increases exponentially, depleting the proportion of candidates that can be 'discovered.'

These observations lead to the insight that it would be desirable to protect the discoveries that can be made in the denser smaller search spaces whenever the search space is enlarged to allow larger antecedents.

Bay and Pazzani (2001) developed a direct-adjustment technique that could apply a Bonferroni-like adjustment without knowing the size of the search space in advance. As they noted, the underlying theoretical basis for the Bonferroni correction is that the sum of all critical values be no more than the desired upper bound on the risk of any false discovery, α. It is not necessary that the critical values employed across multiple tests be identical. To create a Bonferroni-like adjustment that could be used without knowing in advance the depth to which the search would extend, they developed a scheme whereby the critical value employed at a level of the search space was no more than

$$\alpha'_L = \alpha/(2^L \times H_L) \tag{5}$$

Table 2 The density of candidate patterns at each level of the search space

Dataset	$X_{\max}$	Size	Cand.	Density	α'	Δdisc
BMS-WebView-1	1	2.46×10^{05}	3126	0.012707	2.03×10^{-07}	+3,010
	2	6.10×10^{07}	4422	7.25×10^{-05}	8.17×10^{-10}	+2,985
	3	1.00×10^{10}	1963	1.96×10^{-07}	4.95×10^{-12}	−555
	4	1.24×10^{12}	254	2.05×10^{-10}	4.00×10^{-14}	−487
	5	1.22×10^{14}	7	5.75×10^{-14}	4.07×10^{-16}	−450
	6	9.98×10^{15}	0	0	4.95×10^{-18}	−430
Covtype	1	1.78×10^{04}	74	0.004148	2.80×10^{-06}	+68
	2	1.15×10^{06}	216	0.000188	4.28×10^{-08}	+177
	3	4.86×10^{07}	745	1.53×10^{-05}	1.00×10^{-09}	+501
	4	1.50×10^{09}	1794	1.2×10^{-06}	3.23×10^{-11}	+944
	5	3.65×10^{10}	3138	8.61×10^{-08}	1.32×10^{-12}	+1,158
	6	7.18×10^{11}	4028	5.61×10^{-09}	6.61×10^{-14}	+1,045
IPUMS LA 99	1	3.08×10^{06}	526	0.000171	1.62×10^{-08}	+440
	2	2.23×10^{09}	1626	7.3×10^{-07}	2.24×10^{-11}	+1,029
	3	9.58×10^{11}	3172	3.31×10^{-09}	5.21×10^{-14}	+1,279
	4	2.76×10^{14}	3058	1.11×10^{-11}	1.81×10^{-16}	+735
	5	5.70×10^{16}	1381	2.42×10^{-14}	8.73×10^{-19}	+39
	6	8.87×10^{18}	235	2.65×10^{-17}	5.60×10^{-21}	−96
KDDCup98	1	1.50×10^{08}	402	2.69×10^{-06}	3.34×10^{-10}	+78
	2	4.39×10^{11}	1483	3.38×10^{-09}	1.14×10^{-13}	+15
	3	7.49×10^{14}	2753	3.68×10^{-12}	6.68×10^{-17}	−10
	4	8.75×10^{17}	2963	3.39×10^{-15}	5.71×10^{-20}	−8
	5	7.65×10^{20}	1783	2.33×10^{-18}	6.53×10^{-23}	−2
	6	5.27×10^{23}	604	1.15×10^{-21}	9.47×10^{-26}	0
Letter recognition	1	4.66×10^{03}	854	0.183262	1.07×10^{-05}	+606
	2	1.30×10^{05}	3149	0.024285	3.72×10^{-07}	+1,433
	3	2.16×10^{06}	3531	0.001636	2.18×10^{-08}	+705
	4	2.45×10^{07}	1909	7.79×10^{-05}	1.87×10^{-09}	−47
	5	2.00×10^{08}	496	2.48×10^{-06}	2.20×10^{-10}	−123
	6	1.24×10^{09}	25	2.01×10^{-08}	3.40×10^{-11}	−126
Mush	1	1.52×10^{04}	778	0.051117	3.29×10^{-06}	+686
	2	8.59×10^{05}	2723	0.003169	5.72×10^{-08}	+1,908
	3	3.03×10^{07}	3578	0.000118	1.60×10^{-09}	+2,250
	4	7.54×10^{08}	2150	2.85×10^{-06}	6.37×10^{-11}	+1,041
	5	1.40×10^{10}	656	4.68×10^{-08}	3.38×10^{-12}	+87
	6	2.01×10^{11}	113	5.62×10^{-10}	2.31×10^{-13}	−127
Retail	1	2.72×10^{08}	5250	1.93×10^{-05}	1.84×10^{-10}	+882
	2	2.23×10^{12}	3693	1.66×10^{-09}	2.24×10^{-14}	−234
	3	1.23×10^{16}	904	7.35×10^{-14}	4.07×10^{-18}	−120
	4	5.05×10^{19}	62	1.23×10^{-18}	9.90×10^{-22}	−73

Table 2 (*Continued*)

Dataset	$X_{\max}$	Size	Cand.	Density	α'	Δdisc
Retail	5	1.66×10^{23}	0	0	3.01×10^{-25}	−42
	6	4.56×10^{26}	0	0	1.10×10^{-28}	−30
Shuttle	1	1.03×10^{03}	380	0.37037	4.87×10^{-05}	+322
	2	1.36×10^{04}	2046	0.150585	3.42×10^{-06}	+1,263
	3	1.04×10^{05}	4206	0.040481	4.22×10^{-07}	+1,291
	4	5.11×10^{05}	2788	0.005456	7.94×10^{-08}	+237
	5	1.65×10^{06}	550	0.000333	2.19×10^{-08}	−94
	6	3.55×10^{06}	23	6.48×10^{-06}	8.58×10^{-09}	−89
Splice Junction	1	5.80×10^{04}	6846	0.118034	8.62×10^{-07}	+578
	2	6.82×10^{06}	2265	0.000332	7.27×10^{-09}	−60
	3	5.25×10^{08}	586	1.12×10^{-06}	9.40×10^{-11}	−136
	4	2.99×10^{10}	46	1.54×10^{-09}	1.64×10^{-12}	−102
	5	1.33×10^{12}	1	7.52×10^{-13}	3.68×10^{-14}	−38
	6	4.86×10^{13}	0	0	1.00×10^{-15}	−38
TICDATA 2000	1	4.68×10^{05}	454	0.00097	1.07×10^{-07}	+78
	2	1.56×10^{08}	1880	1.21×10^{-05}	3.20×10^{-10}	−8
	3	3.40×10^{10}	3328	9.78×10^{-08}	1.46×10^{-12}	−2
	4	5.50×10^{12}	3008	5.47×10^{-10}	9.04×10^{-15}	−16
	5	6.96×10^{14}	1024	1.47×10^{-12}	7.12×10^{-17}	−16
	6	7.25×10^{16}	0	0	6.83×10^{-19}	0

where L is the level of the search space and H_L is the number of hypotheses evaluated at L. Unfortunately, however, as H_L did not include the entire search space of patterns from which those to be evaluated were selected, this approach did not enforce the desired upper bound of α on the risk of any false discovery (Webb 2007).

Nonetheless, we can adopt this general idea, but using the size of the search space at each level instead of the number of hypotheses evaluated, thereby ensuring strict control over the risk of any false discovery. The Bay and Pazzani scheme used a schedule of adjustments structured so that it was not necessary to cap the maximum size of a pattern. Our approaches assume that such a cap exists, as is often the case in real-world pattern discovery applications. Further, because the search space size is usually so much smaller at the lower pattern sizes, we do not want to disproportionately weight the available critical value mass toward the smaller sizes. In consequence we use

$$\alpha'_L = \alpha/(L_{\max} \times S_L) \tag{6}$$

where L is the level of the search space, $L_{\max}$ is the maximum value of L for the current search, and S_L is the size of the search space at L. In rule discovery applications that do not utilize an upper limit on L, the following variant of the Bay and Pazzani scheme might be used in place of (6):

$$\alpha'_L = \alpha/(2^L \times S_L). \tag{7}$$

The underlying basis for these formulae is to ensure that $\alpha \geq \sum_{L=1}^{L_{\max}} \alpha'_L \times S_L$.

Table 3 Layered vs. uniform adjustments for BMS-WebView-1 at different antecedent sizes

Lvl	1	2	3	4	5	6
Space	2.46×10^{5}	6.10×10^{7}	1.00×10^{10}	1.24×10^{12}	1.22×10^{14}	9.98×10^{15}
Uniform	2.03×10^{-7}	8.16×10^{-10}	4.95×10^{-12}	4.01×10^{-14}	4.06×10^{-16}	4.95×10^{-18}
Layered to lvl 1	2.03×10^{-7}					
Layered to lvl 2	1.01×10^{-7}	4.10×10^{-10}				
Layered to lvl 3	6.76×10^{-8}	2.73×10^{-10}	1.66×10^{-12}			
Layered to lvl 4	5.07×10^{-8}	2.05×10^{-10}	1.24×10^{-12}	1.01×10^{-14}		
Layered to lvl 5	4.06×10^{-8}	1.64×10^{-10}	9.95×10^{-13}	8.08×10^{-15}	8.21×10^{-17}	
Layered to lvl 6	3.38×10^{-8}	1.37×10^{-10}	8.29×10^{-13}	6.73×10^{-15}	6.84×10^{-17}	8.36×10^{-19}
U_n/L_6	6.00	5.98	5.96	5.95	5.94	5.93
U_6/L_6	1.47×10^{-10}	3.63×10^{-8}	5.97×10^{-6}	7.36×10^{-4}	7.24×10^{-2}	5.93×10^{0}

Using (6), the critical value employed at a level n of the search space will be no less than $1/L_{\max}$ of the value it would be if the search space were capped at L. Thus, increasing the search space an extra level from n to $n+1$ can be expected to have only modest impact on the number of patterns discovered at levels 1 to n while opening up the possibility of discovering additional patterns at level $n+1$. Table 3 illustrates this effect with respect to the BMS-WebView-1 dataset. The first row lists the levels. The second row shows the size of the search space at the level. This is the number of rules with the specified number of items in the antecedent. The third row lists the uniform critical value that is applied at all levels if search is capped at the specified level. The rows labeled *Layered to lvl* 1–6 show the critical value employed at the level with which the column is labeled when search is capped at the level with which the row is labeled. The row labeled U_n/L_6 shows the uniform critical value if the search is capped at the level for the column divided by the value in the row headed *Layered to lvl* 6. This demonstrates that when the search is increased to allow antecedents containing up to six items, in no case does the critical value employed at a level n decrease more than six-fold relative to the critical value that would be employed under a uniform critical value if the search space were capped at n. In other words, under the layered approach the critical value employed at a level is never greatly reduced from the minimum critical value under the uniform approach that is capable of including rules from that level. The row headed U_6/L_6 shows the critical value employed under the uniform approach if the search space is capped at level 6 divided by the value in the row headed *Layered to lvl* 6. This shows that when the search space is capped at level 6, the critical values employed at all lower levels of the search space are substantially lower if uniform critical values are used, some times by as much as a factor of 10^{10}.

We call this strategy for setting critical values within the direct adjustment approach *Layered Critical Values*. For the reasons outlined above, we expect it in general to find more patterns than the standard direct adjustment approach, which utilizes a uniform critical value for all patterns.

5 Experiments

To evaluate whether Layered Critical Values do indeed increase the statistical power of our techniques, we replicate the key experiments in (Webb 2007), using Layered Critical Values in place of the uniform critical value employed by the initial direct-adjustment technique.

Three such experiments were performed. Two, with synthetic data, evaluated performance in a context where the actual true and false patterns were known. Such synthetic experiments provide a means of examining the false discovery rates of alternative techniques, as the complete set of true patterns is usually not known in the context of real-world data. One experiment, with real-world data, evaluated the relative power of the techniques in real-world contexts. A fourth experiment, for which there is no equivalent in the earlier research, compares the two techniques in the context of k-optimal pattern discovery. For all four experiments we used the Fisher exact test for productive rules, described in Appendix 1 of (Webb 2007).

The original versions of the first three experiments compared a number of techniques including a number of variants of the holdout evaluation approach as well as the direct adjustment approach. For the holdout treatments, half of each dataset was used for exploration and the remaining half for statistical evaluation. A number of alternative approaches to finding patterns at the exploratory stage were compared. We reproduce here only the approach that proved most effective, and only for the experiment on real-world data.

Note that when the size of the antecedent is limited to no more than 1 item, the use of either layered or uniform critical values will be equivalent and hence have identical results.

5.1 Experiment 1

The first experiment explored the effect of increases to the size of the search space that do not increase the number of true patterns available to be found. The original experiment generated random data for ten pairs of binary variables x_0 and y_0 through to x_9 and y_9. Each x_i was generated at random with each value being equiprobable. The probability of $y_i = 1$ was $1.0 - i \times 0.05$ if $x_i = 1$, $i \times 0.05$ otherwise. This gave rise to forty valid (productive) rules of the forms $x_i = 0 \rightarrow y_i = 0$, $x_i = 1 \rightarrow y_i = 1$, $y_i = 0 \rightarrow x_i = 0$ and $y_i = 1 \rightarrow x_i = 1$. The rules for x_0 and y_0 represent very strong correlations and were straightforward to detect. The strength of correlation declines for successive indexes with the rules for x_9 and y_9 representing relatively weak correlations. Thus there is considerable variety in the ease with which the rules may be detected. As all valid rules had only one item in the antecedent, any increase in the maximum allowed size of the antecedent served to increase the search space without increasing the number of valid rules in the search space. The maximum allowed antecedent size was varied through every size from 1 to 5.

The quantity of data was varied by generating datasets of the following sizes: 250, 500, 1,000, 2,000, 4,000, 8,000 and 16,000. These sizes were selected by experimentation as those providing the most interesting variations in performance. 100 random datasets were generated at each of these sizes. Each larger dataset was generated by adding additional data to the immediately smaller dataset.

Table 4 presents results for experiment 1 for direct adjustment using alternatively a uniform critical value and layered critical values. For each treatment it lists the total number of true and false discoveries over all 100 runs, together with the number of runs for which any false discoveries occurred. Note that it is the latter that the experimentwise significance test seeks to control. If we have successfully controlled this risk at the 0.05 level then the average number of runs for which any false discoveries occur should be no more than $0.05 \times 100 = 5$. Cells with value zero have been left blank to enhance readability.

As can be seen, at all but the largest dataset sizes, for which direct adjustment with a uniform critical value is able to find all 40 rules at all search space sizes, layered critical values suffers less than the uniform critical value from the effect of decreasing numbers of discoveries as the size of the search space increases. This improved power is obtained at

Table 4 Results for experiment 1

	Data	Total true discoveries					Total false discoveries					Experiment false discoveries				
		1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Direct	250	3128	2972	2876	2800	2720										
	500	3520	3376	3268	3200	3144	4					1				
	1000	3688	3612	3584	3528	3476										
	2000	3856	3752	3680	3656	3636	12					3				
	4000	3996	3984	3924	3904	3876										
	8000	4000	4000	4000	4000	4000	8					1				
	16000	4000	4000	4000	4000	4000										
Layered	250	3128	3096	3056	3052	3044										
	500	3520	3484	3484	3480	3472	4					1				
	1000	3688	3680	3664	3656	3652										
	2000	3856	3836	3800	3796	3788	12	8	8			3	2	2		
	4000	3996	3996	3992	3992	3992										
	8000	4000	4000	4000	4000	4000	8	8	8	8	4	1	1	1	1	1
	16000	4000	4000	4000	4000	4000										

a cost of a modest increase in the frequency of false discoveries, but in no case do more than 0.05 of runs result in any false discoveries, illustrating how the experimentwise false discovery rate is still strictly controlled. While the magnitude of the benefit to the layered approach is small in this experiment, this is only because the artificial domain includes small numbers of weaker rules. The significance of the result is to show that the layered approach is better able to discover weaker correlations in the data.

5.2 Experiment 2

For the second experiment the values of 15 binary variables a, b, c, d, e and $x_0, x_1, \ldots x_9$ were randomly generated independently of one another, with each value equiprobable, except for e for which the probability of value 1 was 0.80 if all of a, b, c and d were 1 and 0.48 otherwise.

This generates a total of 83 productive rules, those with:

- one or more of $a = 1$, $b = 1$, $c = 1$ and $d = 1$ in the antecedent and $e = 1$ in the consequent
- $e = 1$ and zero or more of $a = 1$, $b = 1$, $c = 1$ and $d = 1$ in the antecedent and one of $a = 1$, $b = 1$, $c = 1$ and $d = 1$ in the consequent,
- exactly one of $a = 0$, $b = 0$, $c = 0$ and $d = 0$ in the antecedent and $e = 0$ in the consequent, and
- $e = 0$ and zero or more of $a = 1$, $b = 1$, $c = 1$ and $d = 1$ in the antecedent and one of $a = 0$, $b = 0$, $c = 0$ and $d = 0$ in the consequent.

Note that this meant that each increase in the size of the search space but the last increased the number of productive rules that could be found. There are 16 productive rules with 1 item in the antecedent, 30 rules with 2 items, 28 rules with 3 items and 9 rules with 4 items.

Table 5 Results for experiment 2

	Data	Total true discoveries					Total false discoveries					Experiment false discoveries				
		1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Direct	1000	4					12					3				
	2000	36	8	11	22	14										
	4000	128	96	184	270	192	8	4				2	1			
	8000	592	673	876	1109	941		1					1			
	16000	1396	2131	3110	3229	2742	4					1				
	32000	1600	3922	6192	6668	6302										
	64000	1600	4588	7360	8206	8158	4					1				
Layered	1000	4					12	4				3	1			
	2000	36	28	28	40	32										
	4000	128	139	237	347	329	8	4	4	4	4	2	1	1	1	1
	8000	592	866	1220	1568	1517										
	16000	1396	2270	3493	4047	3963	4	4	4			1	1	1		
	32000	1600	3822	6400	7236	7172										
	64000	1600	4584	7380	8280	8276	4	4				1	1			

Identical treatments were applied to those for Experiment 1 except that datasets sizes were varied from 1,000 to 64,000, as larger datasets were required in order to find the more subtle patterns.

Table 5 presents results for experiment 2. Like Table 4, it lists for each treatment the total number of true and false discoveries over all 100 runs, together with the number of runs for which any false discoveries occurred.

At the two largest dataset sizes with antecedents of size up to 2, Layered Critical Values finds very slightly fewer rules than uniform critical values. This is because the reduced critical value at level 2 finds slightly fewer of the larger number of rules at this antecedent size, and this is not offset by as many retentions of the fewer rules from antecedent size 1. At most other data set sizes and antecedent sizes Layered Critical Values finds substantially more rules. Again there is a modest increase in the false discovery rate, but in no case does it approach, let alone exceed a total of 5 runs, which corresponds to the significance level 0.05.

5.3 Experiment 3

The third experiment evaluates the performance of Layered Critical Values on real-world data. The respective approaches were applied to eight of the largest attribute-value datasets from the UCI machine learning (Newman and Hettich 2007) and KDD (Hettich and Bay 2007) repositories together with the BMS-WebView-1 (Zheng et al. 2001) and Retail (Brijs et al. 1999) datasets. These datasets are described in Table 6.

The original study (Webb 2007) first found for each dataset the minimum even value for minimum-support that produced fewer than 10,000 productive rules when applied with respect to the dataset as a whole with antecedents of size up to six. These values are listed in the min sup column of Table 6. Each treatment was then applied to each dataset six times, once with each maximum limit on the size of the antecedent $L_{\max}$ from 1 to 6. All runs used

Table 6 Datasets

Dataset	Records	Items	Min sup	Description
BMS-WebView-1	59,602	497	60	E-commerce clickstream data
Covtype	581,012	125	359,866	Geographic forest vegetation data
IPUMS LA 99	88,443	1,883	42,098	Census data
KDDCup98	52,256	4,244	43,668	Mailing list profitability data
Letter recognition	20,000	74	1,304	Image recognition data
Mush	8,124	127	1,018	Biological data
Retail	88,162	16,470	96	Retail market-basket data
Shuttle	58,000	34	878	Space shuttle mission data
Splice Junction	3,177	243	244	Gene sequence data
TICDATA 2000	5,822	709	5,612	Insurance policy holder data

the minimum-support specified, except for the holdout treatments which only use half the data for rule discovery and for which the minimum-support was therefore halved.

Table 7 presents the results of these experiments. Each row presents results for a specified dataset and setting of $L_{\max}$. The meanings of the columns are as follows:

- *Dataset*: The dataset.
- $L_{\max}$: The maximum number of items in the antecedent.
- *Prod*: The number of productive rules 'discovered.' This shows the number of candidates from which the final discoveries are being winnowed under the two direct adjustment approaches. Note that the number of candidates for the holdout approach may differ from this as we apply a smaller minimum support with respect to a smaller dataset.
- *Holdout*: The number of rules found using holdout evaluation. Where holdout discovers more rules than either of the direct-adjustment approaches, this value is underlined.
- *Search space*: The number of rules in the search space. Note that this number is the sum of the number of rules at each level of the search space up to and including $L_{\max}$. The uniform direct-adjustment technique used a critical value of 0.05 divided by this value.
- *Uniform*: The number of rules 'discovered' that passed a significance test with a uniform direct adjustment. Where uniform discovers more rules than layered, this value is set in boldface.
- *Layered*: The number of rules 'discovered' that passed a significance test with a layered direct adjustment. Where layered discovers more rules than direct, this value is set in boldface.

For three of the ten datasets, uniform critical values found slightly more rules than the layered approach at $L_{\max} = 2$, and in one case at $L_{\max} = 3$, for the same reasons as this effect was apparent in experiment 2. For most other combinations of dataset and maximum antecedent size the layered approach found more rules than did a uniform adjustment. In many cases the layered approach found substantially more rules than the uniform approach. At the higher values of $L_{\max}$ the layered approach exhibited a consistent clear advantage. For Splice Junction layered finds more than three times as many rules and for TICDATA 2000 and Retail it finds more than twice as many rules.

For Splice Junction and TICDATA 2000, the number of rules found with a uniform adjustment peaked at antecedent size 1, and increasing the search space resulted in a decrease in the number of rules found. In contrast, layered adjustments found more rules when the

Table 7 Number of rules found in the experiments on real-world data

Dataset	$L_{\max}$	Prod	Holdout	Search space	Uniform	Layered
BMS-WebView-1	1	3126	3316	2.46×10^{05}	3010	3010
	2	7548	7386	6.12×10^{07}	**5995**	5985
	3	9511	7206	1.01×10^{10}	5440	**5932**
	4	9765	7200	1.25×10^{12}	4953	**5907**
	5	9772	7200	1.23×10^{14}	4503	**5886**
	6	9772	7200	1.01×10^{16}	4073	**5867**
Covtype	1	74	68	1.78×10^{04}	68	68
	2	290	247	1.17×10^{06}	245	245
	3	1035	755	4.98×10^{07}	746	**750**
	4	2829	1752	1.55×10^{09}	1690	**1702**
	5	5967	3116	3.80×10^{10}	2848	**2914**
	6	9995	4390	7.56×10^{11}	3893	**4032**
IPUMS LA 99	1	526	452	3.08×10^{06}	440	440
	2	2152	1508	2.23×10^{09}	**1469**	1464
	3	5324	3017	9.60×10^{11}	2748	**2764**
	4	8382	4103	2.77×10^{14}	3483	**3593**
	5	9763	4398	5.73×10^{16}	3522	**3769**
	6	9998	4400	8.93×10^{18}	3426	**3760**
KDDCup98	1	402	88	1.50×10^{08}	78	78
	2	1885	112	4.39×10^{11}	93	**95**
	3	4638	116	7.49×10^{14}	83	**93**
	4	7601	116	8.76×10^{17}	75	**91**
	5	9384	116	7.66×10^{20}	73	**91**
	6	9988	116	5.28×10^{23}	73	**91**
Letter recognition	1	854	574	4.66×10^{03}	606	606
	2	4003	1905	1.34×10^{05}	2039	**2040**
	3	7534	2581	2.29×10^{06}	2744	**2808**
	4	9443	2702	2.68×10^{07}	2697	**2952**
	5	9939	2703	2.27×10^{08}	2574	**2942**
	6	9964	2703	1.47×10^{09}	2448	**2911**
Mush	1	778	690	1.52×10^{04}	686	686
	2	3501	2567	8.75×10^{05}	2594	**2599**
	3	7079	4838	3.12×10^{07}	4844	**4893**
	4	9229	6039	7.85×10^{08}	5885	**6049**
	5	9885	6346	1.48×10^{10}	5972	**6336**
	6	9998	6412	2.16×10^{11}	5845	**6377**
Retail	1	5250	1036	2.72×10^{08}	882	882
	2	8943	1099	2.23×10^{12}	648	**877**
	3	9847	1099	1.23×10^{16}	528	**856**
	4	9909	1099	5.05×10^{19}	455	**846**
	5	9909	1099	1.66×10^{23}	413	**840**
	6	9909	1099	4.56×10^{26}	383	**838**

Table 7 (*Continued*)

Dataset	$L_{\max}$	Prod	Holdout	Search space	Uniform	Layered
Shuttle	1	380	316	1.03×10^{03}	322	322
	2	2426	1446	1.46×10^{04}	**1585**	1565
	3	6632	2507	1.19×10^{05}	**2876**	2868
	4	9420	2768	6.30×10^{05}	3113	**3200**
	5	9970	2785	2.28×10^{06}	3019	**3189**
	6	9993	2785	5.83×10^{06}	2930	**3169**
Splice Junction	1	6846	308	5.80×10^{04}	578	578
	2	9111	430	6.88×10^{06}	518	**678**
	3	9697	485	5.32×10^{08}	382	**689**
	4	9743	488	3.04×10^{10}	280	**684**
	5	9744	488	1.36×10^{12}	242	**667**
	6	9744	488	5.00×10^{13}	204	**654**
TICDATA 2000	1	454	86	4.68×10^{05}	78	78
	2	2334	78	1.56×10^{08}	70	**86**
	3	5662	78	3.42×10^{10}	68	**86**
	4	8670	78	5.53×10^{12}	52	**86**
	5	9694	78	7.02×10^{14}	36	**86**
	6	9694	78	7.32×10^{16}	36	**86**

search space was increased, retaining the majority of the discoveries at antecedent size 1 while added further discoveries at larger antecedent sizes.

Holdout evaluation found more patterns than direct-adjustment with a uniform critical value for 46 treatments while uniform found more for just 13. The introduction of layered critical values greatly increases the number of treatments for which direct-adjustment finds the most patterns. Holdout evaluation still finds the most patterns more often, doing so for 33 treatments, but the number for which direct-adjustment finds the most doubles to 26. The introduction of layered critical values makes the direct adjustment strategy much more competitive with the holdout strategy.

5.4 Experiment 4

K-optimal (also known as *top-k*) pattern discovery finds the k patterns that optimize a metric of interest within any user-specified constraints (Webb 1995; Scheffer and Wrobel 2002; Han et al. 2002; Webb and Zhang 2005). This approach is attractive when the user can specify an upper-limit on the number of patterns that it might be useful to consider. In particular, the top-k constraint often removes the need for a minimum support constraint. However, the holdout evaluation strategy does not integrate well into this approach as it is not possible to determine how many patterns will be accepted, thus undermining the rationale for the approach. In contrast, the direct-adjustment strategy can be implemented as a constraint within the k-optimal approach, the result being that the k significant patterns are found that optimize the measure of interest. In this context it is desirable to maximize the power of the statistical test, so that high-value patterns will not be excluded from the results.

In this experiment, the Magnum Opus software was run on each of the ten datasets used in Experiment 3, seeking the 1000 rules that maximize support within the constraints that the

Table 8 Results for experiment 4

Dataset	Strategy	Num. rules	Min	Max	Mean
BMS-WebView-1	layered	1000	0.002	0.020	0.003
	uniform	1000	0.002	0.020	0.003
Covtype	layered	1000	0.882	0.949	0.906
	uniform	1000	0.879	0.949	0.905
IPUMS LA 99	layered	1000	0.565	0.981	0.613
	uniform	1000	0.562	0.981	0.610
KDDCup98	layered	1000	0.717	0.997	0.769
	uniform	1000	0.700	0.997	0.754
Letter recognition	layered	1000	0.111	0.487	0.168
	uniform	1000	0.104	0.487	0.159
Mush	layered	1000	0.251	0.973	0.334
	uniform	1000	0.248	0.973	0.326
Retail	layered	1000	0.001	0.331	0.006
	uniform	1000	<0.001	0.331	0.004
Shuttle	layered	1000	0.076	0.656	0.137
	uniform	1000	0.074	0.656	0.135
Splice Junction	layered	1000	0.034	0.401	0.102
	uniform	247	0.032	0.401	0.142
TICDATA 2000	layered	1000	0.142	0.999	0.329
	uniform	1000	0.102	0.998	0.238

rule must be statistically significant and the antecedent contain no more than six elements. Table 8 presents the number of rules found by each approach, and the minimum, maximum and mean values of support for those rules. Note that fewer than k (1000) rules will only be found if there are not sufficient rules that pass the specified constraints. This was only the case with respect to the uniform approach on the Splice Junction data for which uniform could find only 247 rules while layered could still find the requested 1000.

For many of the datasets there is little difference between the minimum, maximum and mean values, layered never being worse off and usually enjoying a slight advantage. The reason that the advantage is often small is because strong patterns will have high support and so for a dataset with many strong patterns the significant patterns with high support are likely to have very low p-values. For datasets with few strong patterns a more obvious difference in performance is more likely. This is apparent with the Splice Junction data for which the uniform approach fails to find many rules found by layered. It is also apparent with the Retail data. While both approaches find 1000 rules, the average value of those found by uniform is only 66% of that found by layered. Likewise, for TICDATA 2000 there is a substantial difference in the minimum and mean support of the rules found by the two approaches.

6 Conclusions

False discoveries are a serious problem for pattern discovery. In many cases the majority of patterns discovered by standard techniques can be spurious artifacts of the sample data.

The direct adjustment and holdout evaluation approaches bring the power of statistical hypothesis testing to bear upon this problem. Both approaches allow any applicable hypothesis test to be used to screen the discovered patterns while strictly bounding the risk of any false discoveries at a user specified rate. Each approach has relative strengths and weaknesses. The holdout approach can be applied as a simple wrapper to any existing pattern discovery system, enabling statistically sound pattern discovery to be easily retrofitted to any existing software. It is also suffers less from decreases in the numbers of discoveries as a result of increases in the size of the search space. It can employ the more powerful Holm procedure in place of the Bonferroni adjustment. In addition to controlling the experimentwise error rate, it can also control the false discovery rate. However, only the direct-adjustment approach integrates directly with the k-optimal (also known as k-top) pattern discovery paradigm. It also uses all available data for both detection and evaluation of patterns.

While it is important to strictly control the risk of false discoveries, it is also important to maximize the number of discoveries that can be made while such control is enforced. This research has demonstrated substantial improvement can be obtained in the numbers of patterns found within the direct adjustment strategy by using the layered critical values approach, notwithstanding that it was developed for the different purpose of enabling a Bonferroni-like adjustment without knowing in advance the maximum depth of the search. Indeed, while previous research had suggested that holdout evaluation was generally more powerful than direct adjustment (Webb 2007), the layered critical values approach lifts the performance of direct adjustment to be in many cases competitive with holdout evaluation, with the additional advantage that it is directly applicable to k-optimal pattern discovery techniques.

Acknowledgements This research has been supported by Australian Research Council Grant DP0772238. We are grateful to Janice Boughton for assistance in the collation of experimental results and to our editor Johannes Fürnkranz and anonymous reviewers for numerous helpful suggestions and comments.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining associations between sets of items in massive databases. In *Proceedings of the 1993 ACM-SIGMOD international conference on management of data* (pp. 207–216). Washington, DC, May 1993.

Aumann, Y., & Lindell, Y. (1999). A statistical theory for quantitative association rules. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining (KDD-99)* (pp. 261–270).

Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., & Lakhal, L. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In *First international conference on computational logic—CL 2000* (pp. 972–986). Berlin: Springer.

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

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

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

Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (1999). Using association rules for product assortment decisions: A case study. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 254–260). New York: ACM.

Brin, S., Motwani, R., & Silverstein, C. (1997). Beyond market baskets: Generalizing association rules to correlations. In J. Peckham (Ed.), *SIGMOD 1997, Proceedings ACM SIGMOD international conference on management of data*, May 1997 (pp. 265–276). New York: ACM.

Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining (KDD-99)* (pp. 15–18). New York: ACM.

DuMouchel, W., & Pregibon, D. (2001). Empirical Bayes screening for multi-item associations. In *KDD-2001: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, August 2001 (pp. 76–76). New York: ACM.

Gionis, A., Mannila, H., Mielikainen, T., & Tsaparas, P. (2006). Assessing data mining results via swap randomization. In *12th international conference on knowledge discovery and data mining (KDD)* (pp. 167–176).

Han, J., Wang, J., Lu, Y., & Tzvetkov, P. (2002). Mining top-K frequent closed patterns without minimum support. In *International conference on data mining* (pp. 211–218).

Hettich, S., & Bay, S. D. (2007). *The UCI KDD archive*. Department of Information and Computer Science, University of California, Irvine, CA. http://kdd.ics.uci.edu.

Holland, B. S., & Copenhaver, M. D. (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin*, *104*(1), 145–149.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

International Business Machines (1996). *IBM intelligent miner user's guide, version 1, release 1*.

Jaroszewicz, S., & Simovici, D. A. (2004). Interestingness of frequent itemsets using Bayesian networks as background knowledge. In R. Kohavi, J. Gehrke, & J. Ghosh (Eds.), *KDD-2004: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, August 2004 (pp. 178–186). New York: ACM.

Klösgen, W. (1996). Explora: A multipattern and multistrategy discovery assistant. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 249–271). Menlo Park: AAAI.

Liu, B., Hsu, W., & Ma, Y. (1999). Pruning and summarizing the discovered associations. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining (KDD-99)*, August 1999 (pp. 125–134). New York: AAAI.

Megiddo, N., & Srikant, R. (1998). Discovering predictive association rules. In *Proceedings of the fourth international conference on knowledge discovery and data mining (KDD-98)* (pp. 27–78). Menlo Park: AAAI.

Možina, M., Demšar, J., Žabkar, J., & Bratko, I. (2006). Why is rule learning optimistic and how to correct it. In *Machine learning: ECML 2006* (pp. 330–340). Berlin: Springer.

Newman, D. J., Hettich, S., Blake, C., & Merz, C. J. (2007). *UCI repository of machine learning databases* (*Machine-readable data repository*). Department of Information and Computer Science, University of California, Irvine, CA.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro & J. Frawley (Eds.), *Knowledge discovery in databases* (pp. 229–248). Menlo Park: AAAI/MIT Press.

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

Scheffer, T., & Wrobel, S. (2002). Finding the most interesting patterns in a database quickly by using sequential sampling. *Journal of Machine Learning Research*, *3*, 833–862.

Webb, G. I. (1995). OPUS: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research*, *3*, 431–465.

Webb, G. I. (2001). Discovering associations with numeric variables. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining (KDD-2001)* (pp. 383–388). New York: ACM.

Webb, G. I. (2002). *Magnum opus, version 1.3. Software*. Melbourne: G.I. Webb & Associates.

Webb, G. I. (2006). Discovering significant rules. In L. Ungar, M. Craven, D. Gunopulos, & T. Eliassi-Rad (Eds.), *Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining, KDD-2006* (pp. 434–443). New York: ACM.

Webb, G. I. (2007). Discovering significant patterns. *Machine Learning*, *68*(1), 1–33.

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

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

Zhang, H., Padmanabhan, B., & Tuzhilin, A. (2004). On the discovery of significant statistical quantitative rules. In *Proceedings of the tenth international conference on knowledge discovery and data mining (KDD-2004)*, August 2004 (pp. 374–383). New York: ACM.

Zheng, Z., Kohavi, R., & Mason, L. (2001). Real world performance of association rule algorithms. In *Proceedings of the seventh international conference on knowledge discovery and data mining (KDD-2001)*, August 2001 (pp. 401–406). New York: ACM.