A New Method for Rule Finding Via Bootstrapped Confidence Intervals

Norman Matloff*

Abstract

Association rule discovery in large data sets is vulnerable to producing excessive false positives, due to the *multiple inference effect*. This paper first sets this issue in precise mathematical terms and presents some analytical results. These show that a common concern regarding effects of filtering is not as problematic as had been previously thought. The analytical results also shed new light on a recently proposed method for dealing with the problem. The paper then proposes a new method based on simultaneous confidence intervals, computed via a novel use of the statistical bootstrap tool. The proposal here differs markedly from previous bootstrap/resampling approaches, not only in function but also in basic goal, which is to enable much more active participation by domain experts.

1 Introduction

There is a rich literature on methods for association rule discovery [4], one of the major issues being control of the number of *false positives*, i.e. discovered "rules" which are only statistical artifacts. As seen in the long list of citations in [12], many solutions have been proposed, but a fully satisfactory approach has yet to be devised. The present paper will present a different approach that has certain desirable features that should make it a valuable tool in the field. The issue is related to the statistical *overfitting* problem, which itself has a very rich literature, for instance [10] [7].

In Section 2, the problems will be given a precise formulation, with a deeper look at the process. It will be shown that though previous work has raised concerns regarding the two-stage nature of standard rule-finding approaches, for typical data mining applications there is actually very little to worry about. This will have a direct positive consequence for the bootstrap method proposed here, as it will reduce computational needs. Theoretical results will also be presented, showing that Bonferonni-based approaches may have insurmountable problems, thus motivating the alternative methodology proposed here.

The focus of the present work is on controlling the number of false positives through resampling operations. There has been much previous work that takes a resampling approach [6] [14] but Section 3 will present a new method, based on a novel type of bootstrap-based confidence intervals. This new methodology is aimed at maximizing the active involvement of domain experts in the rule discovery process. Numerical investigations of the proposed method will be presented.

In order to more clearly illustrate the issues, the present paper will focus on the market basket setting. However, the analysis and methods will be seen to easily generalize to non-binary attributes.

2 A Closer Look at the Rule Finding Problem

In order to prepare for presenting the new proposed methodology, this section will describe the rule-finding problem in precise mathematical terms, and will present some theoretical results that shed light on the issues.

2.1 Problem Statement Let r denote the number of attributes. As noted, in this market basket setting, the j^{th} attribute is 1-0 valued, indicating the presence or absence, respectively, of a purchase of the j^{th} item. Suppose we have data on n transactions, and let X_{ij} denote the value of the j^{th} attribute in the i^{th} transaction. Finally, define $T_i = (X_{i1}, ..., X_{ir})$ to be the i^{th} transaction as a whole, i = 1,...,n. Statistical approaches to the rule finding problem model the n transactions as a sample from some population [8].

A number of measures have been proposed for the effectiveness of a rule [2]. For simplicity, this paper will focus on the "classical" measures, *support* and *confidence*. Again, the results here will be easily extendable to other measures.

Let $S = (s_1, ..., s_r)$ denote an itemset, i.e. a collection of distinct numbers from {1,...,r}. For instance, $S = (2, 5, 6)$ would represent items 2, 5 and 6. For any transaction T, write $S \subset T$ to indicate that T contains the items in S, and possibly others. Let freq(S) denote the number of T_i, i=1,...,n such that $S \subset T_i$.

Consider disjoint itemsets U and V, and a potential rule $U \Rightarrow V$. Let $U \cap V$ denote the situation in which a transaction contains all items of both U and V. (The sets being intersected are sets of transactions, not sets of items.) The support and confidence of the rule are defined to be $supp(U \cap V) = freq(U \cap V)/n$ and $conf(V|U) = supp(U \cap V)/supp(U)$.

*Department of Computer Science, University of California, Davis

2.2 The Multiple Inference Problem This then brings us to the heart of the multiple inference problem. The random quantities supp(U) and conf(V|U) are then estimators of P(U) and P(V|U), the population proportions of transactions satisfying the specified situations. As such, the estimators are subject to statistical sampling error. Thus a rule that looks strong, i.e. with a high value of conf(V|U), may actually be a sampling artifact with the true value of P(V|U) being considerably smaller. Worse, the more possible rules we consider, the more chance there is that at least one of the estimates will be much larger than the true population value.

Rule discovery procedures typically use the classical statistical hypothesis testing approach (including the equivalent use of p-values). The tests typically have null hypotheses such as $H_0 : P(U) = 0$, $H_0 : P(U) \leq minsupp$, H_0: U and V are independent, $H_0 : P(V|U) \leq minconf$, and so on. Here *minsupp* and *minconf* are desired minimum levels of support and confidence.

A statistical significance level will be chosen, say the classical 0.05, and a rule will be used if the test statistic exceeds the corresponding threshhold. Again, even though the Type I error probability for each test is 0.05, that probability for all the tests collectively—the *experimentwise or familywise error rate*—is much higher than 0.05. This *multiple inference problem*, also called the *simultaneous inference problem*, has been extensively covered in the statistical literature [5].

2.3 Another Type of Multiple Inference Problem A related concern has been a possible compromise of a nominal Type I error level due to filtering [14] [12] [13]. Suppose one follows the typical two-stage process: In the first stage, we find a set of candidate rules with estimated support exceeding *minsupp*, and then in the second stage we searche among that candidate set for rules with high confidence values. To be statistically sound, any inference in the second stage must be based on the conditional distribution of the estimated confidence values, given the result of the first stage—a random event, with different samples possibly producing different candidate sets. Yet typical practice assumes the unconditional distribution, so that the nominal value set for α, the probability of a Type I error, is technically incorrect.

One could solve this problem by simultaneously filtering on both support and confidence. However, this greatly increases the number of tests to be performed, and thus seriously exacerbates the problem of loss of power due associated with multiple inference procedures. This would also create a huge computational problem for the bootstrap method presented later in this paper. However, fortunately the issue of conditional inference is not a serious one for large samples. This can be seen analytically:

LEMMA 2.1. *Consider a vector-valued sequence of random variables $\{W_n\}$, and let f() be a scalar-valued function of the W_n such that*

$$\lim_{n\to\infty} P[f(W_n) \leq t] \tag{2.1}$$

exists. Consider a sequence of events $\{Q_n\}$ such that $\lim_{n\to\infty} P(Q_n) = 1$. Then

$$\lim_{n\to\infty} P[f(W_n) \leq t|Q_n] = \lim_{n\to\infty} P[f(W_n) \leq t] \tag{2.2}$$

Proof. Define R_n to be 1 or 0, according to whether $f(W_n) \leq t$, and define S_n similarly for the occurrence of Q_n. Then

$$\begin{aligned} & |P[f(W_n) \leq t] - P[f(W_n) \leq t \text{ and } Q_n)| \\ = \; & |E(R_n) - E(R_n S_n)| \\ = \; & |E(R_n(1 - S_n)| \\ \leq \; & |E(1 - S_n)| \\ = \; & 1 - P(Q_n) \\ \to \; & 0 \; as \; n \to \infty \end{aligned} \tag{2.3}$$

The result then follows from the relation $P[f(W_n) \leq t|Q_n) = P[f(W_n) \leq t \text{ and } Q_n]/P(Q_n)$

Here the function f would be the one that computes a hypothesis test statistic, and the events Q_n would represent the random candidate set chosen during the support-oriented stage of the rule discovering process. Then $P[f(W_n) \leq t|Q_n]$ is 1.0 minus our probability of a Type I error if the null hypothesis is true. We hope this to be α (asymptotically). The result says that for large samples, which we typically have in data mining contexts, we can safely compute our inferences on confidence values as if the distributions are unconditional. We thus need not worry that the true value of α is different than the value at which we set it.

2.4 Bonferroni-Based Approaches The simplest approach to the multiple inference problem is to use the Bonferroni Inequality. If one does b tests and desires an experimentwise error rate of at most α, one can achieve this by performing each test with a Type I error level of α/b. In order to motivate this paper's proposal of "yet another" solution to the multiple inference problem, it is important to first discuss the fundamental problems arising from Bonferroni-based approaches.

In data mining contexts, b can be quite large, often in the tens of thousands or more. This can severely

compromise statistical *power*, i.e. the ability to reject a hypothesis which is false. For instance, the "Z cutoff value" for a one-sided 0.05 level test based on a normal distribution with b = 1 is 1.65, but for b = 100 it is already doubled, to 3.29. This makes rejection of a hypothesis much more difficult, and thus has reduced power to discover true rules. In other words, though we are controlling the probability of discovering false rules, we are limiting our ability to discover true ones.

This can be a very subtle phenomenon. For example, consider the *holdout* procedure proposed by Webb [12] [13]. The approach is quite ingenious, typically achieving a major reduction in b, via holdout: The sample is split into two subsamples. Tests are performed in the first subsample without regard to multiple inference issues at all, i.e. each test is perfomed at the full value of α. Suppose k rules are found in that stage. These rules are then subjected to verification on the second subsample, using multiple inference methodology, say some form of Bonferroni with b = k. Since k will typically be much smaller than the number of potential rules, this achieves a much smaller value of b than would have been the case for straightforward application of Bonferroni, referred to as the *direct* method in [12] [13].

The empirical findings in those papers were mixed. Although the holdout method discovered more rules when applied to a real data set, one of course cannot tell whether the rules were true ones. In the simulation experiments, the results for the holdout and direct methods were generally within sampling error of each other, though the holdout method seemed to bring an improvement of as much as 20% in some settings.

Unfortunately, due to the huge computational work involved in these simulations, the empirical studies in [12] [13] were limited. In the two experiments performed, the space of potential rules consisted of only 40 rules in the first experiment and 83 rules in the second. It is thus desirable to perform some theoretical analysis, so as to investigate the behavior of the holdout and direct approaches for much larger rule spaces.

To this end, let consider the following abstraction of the problem. Say we have m samples of size h, with the i^{th} being from a $N(\mu_i, 1)$ distribution. Let G_i denote the i^{th} sample mean, and suppose we are testing $H_0 : \mu_i = 0$, i = 1,...m. This can be thought of, say, as an abstraction of the problem of testing for nonzero support for m different itemsets.

Suppose that unknown to us, the true value of μ_i is d for i = 1,...,k and that μ_i is 0 for i = k+1,...,m. Let's take $\alpha = 0.05$. Our test statistic for μ_i is

$$Z_i = \frac{G_i}{1/h^{0.5}} \tag{2.4}$$

Under a standard Bonferroni approach to rule finding, we reject H_0 if $Z_i > g_m = \Phi^{-1}(1 - 0.05/m)$, where Φ is the cdf for N(0,1). For instance, as noted, for m = 100 we have $g_m = 3.29$.

For i = 1,...,k, the power of the test is

$$P(Z_i > g_m) = 1 - \Phi\left(g_m - dh^{0.5}\right) \tag{2.5}$$

So, the expected number of correct rule discoveries is

$$k\left[1 - \Phi\left(g_m - dh^{0.5}\right)\right] \tag{2.6}$$

What would happen with the Webb method? Say we apportion our two subsamples to h/2 observations each. For i = 1,...,k, we can obtain the power by applying (2.5) and replacing m by 1 and h by h/2. So the expected number of correct rules found in the first subsample will be

$$\nu_1 = k\left[1 - \Phi\left(g_1 - d(0.5h)^{0.5}\right)\right] \tag{2.7}$$

For i = k+1,...,m the power is 0.05. Thus the expected number of false discoveries found in the first subsample will be

$$\nu_2 = (m - k)0.05 \tag{2.8}$$

As a rough approximation, let's then assume that the number of rules that reach the second stage is the constant $\nu = \nu_1 + \nu_2$, again breaking down into ν_1 true rules and ν_2 false ones.

Then in the second stage the analyst will apply the Bonferroni method with b = ν. The power for the true and false cases is computed as above, so that the expected number of true discoveries will be

$$\nu_1\left[1 - \Phi\left(g_\nu - d(0.5h)^{0.5}\right)\right] \tag{2.9}$$

We evaluated (2.9) for various parameter values (not presented here, producing very mixed results. For instance, for m = 1000000, h = 10000, k = 100 and d = 0.1, the direct method has an expected number of correct rule discoveries of 95.95, almost double the 59.72 value for the holdout method. In some cases, the direct method outperformed the holdout method for a factor of nearly 4. On the other hand, with m = 1000, h = 50, k = 20 and d = 0.2, the holdout method was superior, with a mean number of correct discoveries of 0.35, compared to the direct method's 0.12. Thus the results were even more mixed than in Webb's papers. At present, then, there appears to be no good away to deal with the problem that Bonferroni-based approaches simply cover too much ground. This leads to the the alternative methodology proposed below.

2.5 Bootstrap/Resampling Methods The *bootstrap* enables an analyst to perform a rich variety of inference without making assumptions of a parameteric distribution [3]. One especially valuable aspect of it is that it handles dependencies among multiple test statistics very well, again without making assumptions on the structure of those dependencies.

The operation of the bootstrap is remarkably simple in concept for such a powerful tool. One resamples (with replacement) from the sample data, calculating the given test statistics on each new set of resampled data, and then rejects the hypothesis if the value of the test statistic on the original data is in the upper-α tail of the bootstrapped test statistics.

A good example (possibly the first) of the use of the bootstrap for multiple inference in association rule finding is in [6] and the authors' previous work they cite there.

A variation on the bootstrap is *permutation testing*. In general, this involves actually exchanging the values of one attribute for those of another, in a manner in which distributions would be unchanged under the null hypothesis. In the association rule finding context, [14] exchanges data within a group of potential antecedents, under the null hypothesis of no association.

3 A New Rule Finding Method Based on Bootstrapped Confidence Intervals

3.1 Motivation A fundamental point here will be the use of statistical confidence intervals rather than hypothesis tests. (For brevity and to avoid confusion with the association rule term *confidence*, let us write SCI for "statistical confidence interval.") The overriding goal is to better empower the analyst, as follows.

Consider a common example from elementary statistics courses, in which a new drug for treating hypertension is to be compared with a placebo. Let μ_1 and μ_2 denote the population mean effectiveness for the two drugs. The classical analysis would be to test the hypothesis

$$H_0 : \mu_1 = \mu_2 \tag{3.10}$$

For convenience in the discussion here, let's suppose the alternate hypothesis is $H_0 : \mu_1 \neq \mu_2$.

Another approach would be to form an SCI for the difference $\mu_1 - \mu_2$. As is well known, one can use the SCI as a mechanism for hypothesis testing; one decides that the drug is effective if and only if the SCI does not contain 0. However, the SCI provides the analyst with much more information than this, in that it gives a range for the value of the difference between the two means. Say for example the SCI excludes 0 but is near 0. Here the analyst may decide that the drug should not be used after all, as its limited effectiveness may be outweighed by considerations such as cost or side effects.

In other words, one obtains much more information from SCIs than from hypothesis testing. See [11] and the references in [9] for extended discussions.

Thus the main purpose of this paper is to provide the analyst, for each potential rule $U \Rightarrow V$, information of the form, "The support of this rule is at least c," or "The confidence of this rule is at least d," at a given global α level. In this manner, the analyst can browse freely through the potential rules, applying his/her domain expertise. In light of that expertise, the analyst may wish not to adopt a rule even if it satisfies support and confidence threshholds *minsupp* and *minconf*, and possibly adopt some other rules that come close to meeting these criteria but do not formally exceed these values. Or, it may be, for instance, that the *minsupp* criterion is not quite met, but the *minconf* level is greatly exceeded. This situation may be of high interest to the analyst.

The point is to de-automate the rule-finding process, again providing the flexibility needed to enable the analyst's domain expertise to be exploited. Given the large number of potential rules, some kind of automatic filtering would still be used, but in the end, the analyst would still be in control of the process.

As mentioned, all of this must be done with proper multiple inference controls. This will be done with a novel use of bootstrapping.

3.2 Basics of the Proposed Methodology Let $\theta_i,\ i = 1, ..., g$ be a collection of population values, with estimators $\widehat{\theta_i}$ calculated from sample data $Y_1, ..., Y_m$.

Our goal is to form aproximate one-sided SCIs, i.e. of the form (c, ∞), which hold simultaneously at level $1 - \alpha$. The statistical literature does include work on use of the bootstrap for generating simultaneous SCIs. However, a different approach will be taken here. Define

$$M = \max_i \frac{\widehat{\theta_i}}{\theta_i} \tag{3.11}$$

M is not observable, since the θ_i are unknown. But it still has a distribution, and let q denote the $1 - \alpha$ quantile of M, i.e. the value such that $P(M \leq q) = 1 - \alpha$.

If it were known, the value of q would then provide us with the desired simultaneous confidence intervals, as follows:

$$\begin{aligned} 1 - \alpha &= P(M \leq q) \\ &= P(\widehat{\theta_i}/\theta_i \leq q,\ i = 1, ..., g) \\ &= P(\widehat{\theta_i}/q \leq \theta_i,\ i = 1, ..., g) \end{aligned} \tag{3.12}$$

In other words, the intervals $(\widehat{\theta}_i/q, \infty)$ would be simultaneous SCIs for the θ_i.

The value of q is unknown, but we can estimate it from our data, using the bootstrap. We generate v new samples, each of size m, by sampling with replacement from $Y_1, ..., Y_m$, and then calculate the new values of the $\widehat{\theta}_i$ on each of these new samples. This gives us estimators $\widetilde{\theta}_{ij}$, i = 1,...,g, j = 1,...,v. In analogy to (3.11), we compute

$$\widetilde{M}_j = \max_i \frac{\widetilde{\theta}_{ij}}{\widehat{\theta}_i} \tag{3.13}$$

for j = 1,...,v.

Intuitively, the values $\widetilde{M}_j$ give us an approximation to the distribution of M, and an extensive theory has been developed that shows that the bootstrap indeed works this way. We thus take our estimate of q, denoted $\widehat{q}$, to be the k^{th}-smallest value among $\widetilde{M}_1, ...\widetilde{M}_v$, where $k = \lfloor (1-\alpha)v \rfloor$.

In summary, then, our simultaneous SCIs for the θ_i will be

$$(\frac{\widehat{\theta}_i}{\widehat{q}}, \infty),\ i = 1, ..., g \tag{3.14}$$

3.3 SCIs for Support and Confidence Values Returning to the notation of Section 2.1, let W be a collection of subsets of {1,...,r}, i.e. a collection of sets of attributes. Suppose we wish to find simultaneous SCIs for the population support values for all sets in W.

We apply the methodology in Section 3.2 as follows. The number m will be n; θ_i will be the population value of the support for the i^{th} set A_i in W, i.e. the population mean of the product of all the attributes with indices in A_i; and $\widehat{\theta}_i$ is the corresponding sample mean. The SCIs constructed for all the sets in W using (3.14) will hold *simultaneously* at an approximate $1-\alpha$ level.

The case of SCIs for confidence values follows the same principle. Here we take our collection W to consist of (antecedent, consequent) sets, and the $\widehat{\theta}_i$ are the sample confidence values, with θ_i being the corresponding population quantities.

3.4 How Large Will $\widehat{q}$ Be? With multiple inference procedures in general, the concern is that the methods may lack power, which in the SCI context means wide intervals. In (3.14), this corresponds to values of $\widehat{q}$ that are much larger than 1. (With only one interval, the value is 1.) The overriding question, then is whether $\widehat{q}$ would stay reasonably small in practice, near 1 or 2, as opposed to increasing to a value so large as to be of little use, say 10. The present section addresses this issue.

Clearly, as the size of W increases for fixed n, q will tend to grow, as we are taking a maximum in (3.11) over larger collections. Since our θ_i are proportions, the maximum possible value of M in (3.11) is

$$\frac{1}{\min_{i \epsilon W} \theta_i} \tag{3.15}$$

Putting these two points together, we see that the deleterious effects of browsing through large collections W, in terms of q, will be be acceptable in practice as long as $\min_{i \epsilon W}$ stays bounded away from 0. Since $\widehat{q}$ is an estimator of q, $\widehat{q}$ may not be too large even with large collections W, as long as the sample size is large enough for $\widehat{q}$ to be a fairly accurate estimator of q.

For fixed W, $\widehat{q}$ will tend to become smaller as n increases. This is because $Var(\widehat{q})$ decreases, so that values of $\widehat{q}$ that are far above q become less likely.

Potential rules with very small support can also be cause for concern in a slightly different sense. Consider estimating the population support value for an itemset U for which freq(U) = 1. The dataset investigated below, for instance, has at least two such variables. Since the bootstrap (in its basic version) samples with replacement, that single transaction that contains U can be chosen multiple times in a single bootstrap sample. This would then lead to a large value of $\widehat{q}$.

In order to get a quantitative idea of the magnitudes of these effects, the method proposed here was applied to the Accidents data set [1]. In order to assess the effects of varying sample size n on $\widehat{q}$, the experiment considered only the first nr of n transactions, for different values of nr. Then, to gauge the effects of increasingly larger collections W, the experiment considered only the first nv attributes, for varying values of nv.

Figure 1 shows the results, for support values of singleton sets. Though the curves are not monotonic (nor should they be, as this is real data with heterogeneous attributes), clearly the trend is downward in nr, as expected. Also, the larger attribute collections tended to result in larger values of $\widehat{q}$. Nevertheless, the values were mostly well under 2.0, a gratifying result.

Then, with (3.15) in mind, we filtered out all attributes having a singleton support level of under 0.05. (Recall from Section 2.3 that we can do this without jeopardizing the statistical validity of our results.) The results are shown in Figure 2. This did indeed reduce the values of $\widehat{q}$, as surmised. Additional experiments, not presented here, had similar results.

4 Discussion and Conclusions

The new method proposed here gives domain experts the flexibility to fully exploit their expertise. It is readily interpretable, and produces (asymptotically)

exact inference, as opposed to Bonferroni methods, which tend to have low power.

The new method produced good results on the Accident dataset investigated here. Our intuition that it will work well even with large numbers of attributes was confirmed in the context here, though settings with much larger numbers still should be examined. Based on the findings here, the proposed method should also work well on datasets that are sparser than the Accident set, as long as the attributes with very small support levels are excluded.

This paper has also clarified the role of conditional inference, and shed further light on Bonferroni methods, providing further evidence that alternative methods such as the one proposed here are needed. method, as they different goals.

R code for the proposed method is available from the author. As with all bootstrap methods, it is very computationally intensive. For very large collections W, it is recommended that a form of parallel R be used, such as Rmpi or the **pappl()** function.

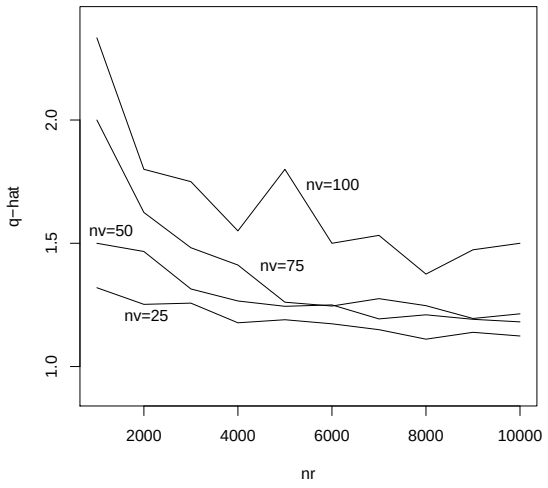

Figure 1: Accidents Data, Full

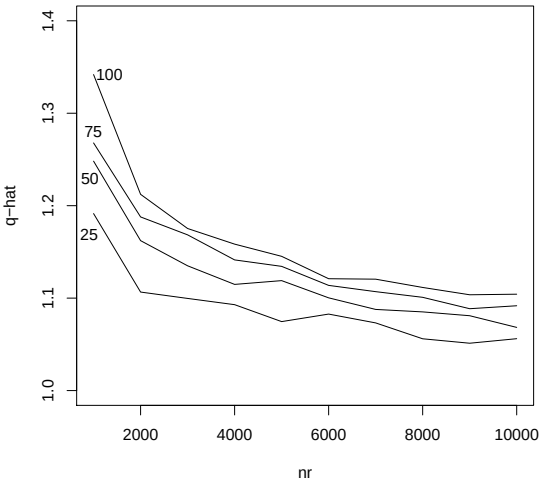

Figure 2: Accidents Data, Small Sets Excluded

References

[1] B. Goethals, M.J. Zaki, *FIMI'03: Workshop on Frequent Itemset Mining Implementations*, 2003.

[2] F. Guillet and H. Hamilton, *Quality Measures in Data Mining*, Springer, 2007.

[3] A. Davison and D. Hinkley, Bootstrap Methods and Their Application, Cambridge University Press, 1997.

[4] F. Höppner, *Association Rules*, in O. Maimon and L. Rokach (eds.), The Data Mining and Knowledge Discovery Handbook, Springer, Berlin (2005), 353-376.

[5] J. Hsu, *Multiple Comparisons: Theory and Methods*, Chapman and Hall, 1996.

[6] S. Lallich, O. Teytaud and E. Prudhomme, *Association Rule Interestingness: Measure and Statistical Validation*, in Quality Measures in Data Mining, F. Guillet and H. Hamilton (eds.), Springer, 2007.

[7] N. Matloff, *A Careful Look at the Use of Statistical Methodology in Data Mining*, in T.Y. Lin, Wesley Chu and L. Matzlack (eds.), Foundations of Data Mining and Granular Computing, Springer-Verlag Lecture Notes in Computer Science, 2005.

[8] N. Megiddo and R. Srikant, *Discovering Predictive Association Rules*, KDD 1998, 274-278.

[9] D. Parkhurst, Commentaries on Significance Testing, `http://www.indiana.edu/~stigtsts`.

[10] S. Portnoy, *Asymptotic Behavior of Likelihood Methods for Exponential Families When the Number of Parameters Tends to Infinity*, Ann. Statist., 16, 356-366, 1988.

[11] C. Wang, Sense and Nonsense of Statistical Inference: Controversy, Misuse and Subtlety, Marcel Dekker, 1993.

[12] G. Webb, *Discovering Significant Rules*, KDD, 2006.

[13] G. Webb, *Discovering Significant Patterns*, Mach. Learn., 2007, 68, 1-33.

[14] H. Zhang, B. Padmanabhan, A. Tuzhilin, *On the Discovery of Significant Statistical Quantitative Rules*, KDD 2004.