Discovering Predictive Association Rules

Nimrod Megiddo and **Ramakrishnan Srikant**

IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120, U.S.A.
{megiddo,srikant}@almaden.ibm.com

Abstract

Association rule algorithms can produce a very large number of output patterns. This has raised questions of whether the set of discovered rules "overfit" the data because all the patterns that satisfy some constraints are generated (the Bonferroni effect). In other words, the question is whether some of the rules are "false discoveries" that are not statistically significant. We present a novel approach for estimating the number of "false discoveries" at any cutoff level. Empirical evaluation shows that on typical datasets the fraction of rules that may be false discoveries is very small. A bonus of this work is that the statistical significance measures we compute are a good basis for ordering the rules for presentation to users, since they correspond to the statistical "surprise" of the rule. We also show how to compute confidence intervals for the support and confidence of an association rule, enabling the rule to be used predictively on future data.

1. Introduction

The problem of mining association rules was introduced in (Agrawal, Imielinski, & Swami 1993). The input consists of a set of transactions, where each transaction is a set of literals (called items). An example of an association rule is: "30% of transactions that contain beer and potato chips also contain diapers; 2% of all transactions contain all of these items". Here 30% is called the *confidence* of the rule, and 2% the *support* of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints. Applications include discovering affinities for market basket analysis and cross-marketing, catalog design, loss-leader analysis and fraud detection. See (Nearhos, Rothman, & Viveros 1996) for a case study of a successful application in health insurance, and (Ali, Manganaris, & Srikant 1997) for applications in medical research and telecommunications diagnosis.

Association Rules Overview. Let $\mathcal{I} = \{l_1, l_2, \ldots, l_m\}$ be a set of literals, called items. Let $\mathcal{D}$ be a set of transactions, where each transaction T is a set of items such that $T \subseteq \mathcal{I}$. We say that a transaction T *contains* a set A of some items in $\mathcal{I}$, if $A \subseteq T$. An *association rule* is an implication of the form $A \Rightarrow B$, where $A \subset \mathcal{I}$, $B \subset \mathcal{I}$, and $A \cap B = \emptyset$. The rule $A \Rightarrow B$ holds in the transaction set $\mathcal{D}$ with *confidence*[1] c if c% of transactions in $\mathcal{D}$ that contain A also contain B. The rule $A \Rightarrow B$ has *support* s in the transaction set $\mathcal{D}$ if s% of transactions in $\mathcal{D}$ contain $A \cup B$.[2] Given a set of transactions $\mathcal{D}$, the computational task of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence respectively.

The task of mining association rules is decomposed into two steps:

- Find all combinations of items that have transaction support above minimum support. Call those combinations *frequent itemsets*.
- Use the frequent itemsets to generate the desired rules. The general idea is that if, say, $ABCD$ and AB are frequent itemsets, then we can determine if the rule $AB \Rightarrow CD$ holds by computing the ratio $r =$ support$(ABCD)$/support(AB). The rule holds only if $r \geq$ minimum confidence. Note that the rule will have minimum support because $ABCD$ is frequent.

The first step is responsible for most of the computation time, and has been the focus of considerable work on developing fast algorithms, e.g. (Agrawal *et al.* 1996) (Brin *et al.* 1997).

[1] In other words, the confidence is the conditional probability $p(B|A)$.

[2] The support is the probability of the *intersection* of the events.

Related Work. Brin et al. (Brin, Motwani, & Silverstein 1997) used the chi-squared test to look for correlated associations, but did not take into account the number of hypotheses were being tested. Piatetsky-Shapiro (Piatetsky-Shapiro 1991) had a similar idea when he argued that a rule $X \Rightarrow Y$ is not interesting if support$(X \Rightarrow Y) \approx$ support$(X) \times$ support(Y), but again did not consider the number of hypotheses.

2. Statistical Significance of Association Rules

In this section, we discuss issues of statistical significance of a set of association rules. First, in Section 2.1, we discuss the statistical significance of a single rule. However, when we analyze many rules simultaneously, the significance test has to take into account the number of hypotheses being tested. This is the so-called multiple comparisons problem (Hochberg & Tamhane 1987). Typically the test statistics corresponding to the hypotheses being tested are not independent. It is important to observe that the number of hypotheses implicitly being tested may be much greater than the number of output rules; we give an upper bound for this number in Section 2.2. This bound may, in general, be too conservative. We offer a practical way of dealing with this problem in Section 2.3; the idea is to use resampling to determine a good acceptance threshold. Finally, in Section 2.4, we describe the computation of confidence intervals for the support and confidence of a rule.

2.1. Statistical Significance of a Single Association

We view a dataset consisting of n transactions as the realizations of n independent identically distributed random boolean vectors, sampled from the "real world" distribution. Let π^S denote the "real world" probability that a transaction contains a given itemset S. Thus, the number of transactions N^S in the sample (i.e., dataset) that contain S is a binomial random variable with success probability $\pi = \pi^S$ and n trials.

Hypothesis testing and minimum support. The minimum support requirement can be cast in the hypotheses testing framework as follows. For example, suppose our minimum support requirement is 10%. For each itemset S, let H_0^S be the *null hypothesis* that $\pi^S = 0.1$, and let us test it against the *alternative* hypothesis H_S^1 that $\pi^S > 0.1$. Let p^S be the fraction of transactions in the dataset that contain S. The test is to compare p^S with a threshold value p_0 and reject H_0^S if and only if $p^S \geq p_0$. There are two kinds of possible errors (Cryer & Miller 1994):

	Truth	
Decision	H_0^S true	H_1^S true
Do not reject H_0^S	Correct decision	Type II error
Reject H_0^S	Type I error	Correct decision

The selection of p_0 is determined by a bound on the desired probability of Type I error, which is called the *significance level.*

P-values. In general, the *p-value* of a test result is the probability of getting an outcome at least as extreme as the outcome actually observed; the p-value is computed under the assumption that the null hypothesis is true (Cryer & Miller 1994). In our example, the p-value corresponding to an observed fraction p^S is equal to the probability, under the assumption that $\pi^S = 0.1$, that the fraction of transactions that contain S is greater than or equal to p^S.

In order to compute the p-value, we use either the normal approximation, the Poisson approximation, or the exact binomial distribution, depending on the actual values of n, the minimum support requirement, and the observed support. For example, suppose $n = 10,000$, the minimum support is $\pi = 0.1$ and the observed support is $p = 0.109$. We use the normal approximation. The mean is $\pi = 0.1$, and the standard deviation is $\sqrt{(\pi(1-\pi)/n)} = \sqrt{(0.09/10000)} = 0.003$. Since p is three standard deviations greater than π, the p-value is 0.0013.

Testing Independence. Consider an association rule $S \Rightarrow T$, where S and T are sets of items. As a null hypothesis we assume that S and T occur in transactions independently. Thus, under the null hypothesis,[3] $\pi^{S \wedge T} = \pi^S \times \pi^T$. As an alternative hypothesis, we can use the inequality $\pi^{S \wedge T} > \pi^S \times \pi^T$, which means that the conditional probability of T given S is greater than the probability of T.

If the values π^S and π^T are assumed to be known with sufficient accuracy, we can use the value $\pi^S \times \pi^T$ to compute a p-value for $S \Rightarrow T$. This p-value corresponds to the probability, under the assumption that S and T are independent, that the empirical frequency of the set $S \cup T$ will be greater than $p^{S \wedge T}$. Since we don't know the actual values of π^S and π^T, we use p^S and

[3] We denote the event that $S \cup T$ is included in the transaction by $S \wedge T$ since this is indeed the intersection of the events corresponding to S and T.

p^T (the fractions of transactions that contain S and T, respectively) as estimates for π^S and π^T. The lower the p-value, the more likely it is that S and T are *not* independent.

2.2. Statistical Significance of a Set of Associations

Suppose we are testing k null hypotheses $H_0^1, \ldots, H_0^k$, and denote by q^i the probability of rejecting H_0^i when it is true. The probability of rejecting at least one of the null hypotheses when they are all true is at most $q^1 + \cdots + q^k$. Thus, if we wish the latter to be smaller than, say, 0.05, it suffices to determine thresholds for the individual tests so that $q_i < 0.05/k$. This bound may be very small if the number of hypotheses implicitly being tested is very large. Indeed, since under the null hypothesis the empirical p-value is distributed uniformly, when we test k true null hypotheses whose test statistics are independent random variables, the expected value of the smallest p-value is $1/(k+1)$, so in order to achieve a small probability of rejecting any true null hypothesis we would have to choose thresholds even smaller than that. Note that when we test independence of pairs in a set of, say, 10,000 items, then the value of k would be greater than 10^7.

Obviously, if we wish to achieve a good probability of rejecting most "false discoveries", we have to increase the probability of rejecting some true ones as well. In other words, when we attempt to discover more true rules, we also increase the risk of false discoveries (i.e., rejecting true null hypotheses). However, we would like to have an idea how many false discoveries there may be for any given threshold. We explain below how to compute an upper bound on the number of hypotheses that are implicitly tested; this number can be used to estimate the number of false discoveries for any given threshold.

Upper Bound on the Number of Hypotheses. The number of hypotheses that we are implicitly testing in the associations algorithm is typically much larger than just the number of frequent itemsets or the number of rules that are generated. To understand why, consider the set of frequent pairs. Let the null hypothesis H_0^{ij} be that the items i and j are independent. To find the set of frequent pairs, the associations algorithm counts the cross-product of all the frequent items. Suppose there are 100 frequent items. Then there will be roughly 5000 pairs of items whose support is counted. If the algorithm throws away 4000 of these *at random* and tests H_0^{ij} only for the remaining 1000 pairs, then only 1000 hypotheses have been tested. On the other hand, if the algorithm picks the 1000 pairs with the smallest p-values, then 5000 hypotheses have been tested. If the algorithm first considers the 1000 pairs with the highest support, and only then looks at p-values, then the actual number of hypotheses being tested is not readily available. In this case, we can use the number 5000 as an upper bound on the number of hypotheses.

We can extend this upper bound to include itemsets and rules with more than two items. Consider an itemset with three frequent items that does not include any frequent pairs. We do not need to include such an itemset while counting the number of hypotheses because this itemset clearly cannot have minimum support (in the dataset) and hence it's properties are never examined by the algorithm. Hence for itemsets with three items, the number of hypotheses is less than the product of the number of frequent pairs times the number of frequent items. In fact, we can further bound the number of hypotheses to just those itemsets all of whose subsets are frequent. This number is exactly the number of candidates counted by current algorithms. By summing this over all the passes, we get

Number of Hypotheses
$\leq$ Number of Candidate Itemsets
$\leq$ Number of Frequent Itemsets $\times$ Number of Frequent Items.

where the set of candidate itemsets includes any itemset all of whose subsets are frequent. We emphasize that this is only an upper bound and may be much higher than necessary. Another serious problem is that even when all the null hypotheses are true, their test statistics are clearly not independent, thus prohibiting a direct calculation of appropriate thresholds. Below we present a practical solution to this problem.

2.3. Determining thresholds by resampling

Given the observed singleton frequencies p_i of the items, we generate a few synthetic data sets of transactions under a model where the occurrences of all the items are independent. Thus, the transactions are generated independently; for each transaction j and for item i we pick a number x_{ij} from a uniform distribution over $[0, 1]$ and include i in j if and only if $x_{ij} < p_i$. Typically, we would generate 9 data sets, each consisting of $10,000$ transactions. These numbers depend on the number of frequent items and minimum support rather than the number of transactions. We run the association rules algorithms on these data sets. Let v_{ij} denote the ith smallest p-value in dataset j. Let V_i denote the

Dataset	Supermarket	Dept. Store	Mail Order
Number of Customers	6200	Unknown	214,000
Number of Transactions	1.5 million	570,000	3 million
Items per Transaction	9.6	4.4	2.6
Min. Support (for exp.)	2%	1%	0.02%
Min. Conf (for exp.)	25%	25%	25%
# Frequent Items	201	283	2849
# Frequent Itemsets[4]	2541	943	10,173
# Candidates	30,000	42,000	4,090,000
# Rules	4828	1020	2479

Table 1: Dataset Characteristics

mean of the values $v_{i1}, v_{i2}, \ldots$ The value V_i estimates the expectation of the ith smallest p-value when all the null hypotheses are true. So, we expect at most i false discoveries when we place the threshold at V_i. These estimates become useful when we wish to assess the quality of the set of rules we mine from the real data set. For example, if in the real data set we consider reporting all the rules with p-values smaller than some threshold t, and if $V_i < t \leq V_{i+1}$, then we expect no more than i of these rules to be false discoveries, since even in a purely synthetic data base where all the null hypotheses are true, no more than an expected number of i turn out to have such small p-values. As the value of t increases, more rules would be reported, but a larger number of them is expected to be false.

We tried this approach on three real-life datasets, whose characteristics are shown in Table 1. We present results for a specific minimum support and confidence for each rule; we got similar results for other values of support and confidence. Table 2 gives for each dataset the results of the simulation. We present results with three different random seeds to give an idea of the variation in p-values. For the supermarket and department store data, we also ran with three different data sizes: 1000, 10,000 and 100,000 transactions.[5] Notice that the average p-values are quite similar for the three data sizes.

We estimated the smallest p-value for each dataset based on the conservative upper bound on the number of hypotheses that we derived in the previous section. There was more than a factor of 100 difference between the expected lowest p-value and the actual least p-value on all three datasets.

For the Supermarket data, only two rules (out of 4828 rules) had p-values higher than 10^{-9}: their p-values were .0037 and .0051. For the Department Store data, only nine rules (out of 1020 rules) had p-values higher than 10^{-100}, and all their p-values were greater than 0.09. For the Mail Order data, none of the rules (out of 2479 rules) had p-values greater than 10^{-40}. Hence the number of "false discoveries" was extremely small.

The reason for the extremely low number of false discoveries is that the support and confidence threshold already do an excellent job of pruning out most rules that are not statistically significant. For instance, consider a rule where the support of the consequent is 5%. For this rule to meet the minimum confidence constraint, the support (confidence) of this rule must be at least 5 times the expected support (confidence) assuming that the antecedent and consequent are independent. Hence, unless the minimum support was extremely low, this rule would have a very low p-value.

2.4. Confidence Intervals

Denote by $B(k; n, s)$ the probability that a binomial random variable with success probability s and n trials will have a value greater than k. The p-value of a rule with observed frequency p, with respect to a desired support level of s is equal to $B(np; n, s)$. Let π denote the true frequency. The probability of the event $\pi - x \leq p \leq \pi + y$ is the same as the confidence level of an interval of the form $[p - y, p + x]$. The symmetry of the normal approximation allows calculating confidence intervals based on the observed value p. If we construct for each rule a confidence interval of level 95%, then for each rule there is an apriori probability of 95% that the true frequency lies within the interval. This means that the expected proportion of the rules where the true frequency lies within the respective interval is 95%. With regard to constructing a confidence interval for the confidence of a rule, we can argue the following. In general, consider events $E_1 \subset E_2$. If $[a, b]$ and $[c, d]$ are confidence intervals of level $1 - \epsilon$ for $\pi(E_1)$ and $\pi(E2)$, respectively, and if $c > 0$, then $[a/d, b/c]$ is a confidence

[5] For the mail order data, the minimum support was too low to get meaningful results with the first two data sizes. With 10,000 transactions, minimum support corresponds to just 2 transactions.

Simulated Dataset	Number of Transactions	Expected Lowest p-value	Lowest p-value	Next Lowest p-value
Supermarket	1,000	3e-5	.0026, .0048, .0072	.0038, .0074, .0089
	10,000		.0030, .0044, .0064	.0049, .0110, .0140
	100,000		.0011, .0022, .0086	.0049, .0055, .0096
Dept. Store	1,000	2e-5	3e-5, .0025, .0025	.0010, .0027, .0029
	10,000		.0013, .0025, .0032	.0032, .0040, .0090
	100,000		.0002, .0021, .0045	.0006, .0022, .0090
Mail Order	100,000	2e-7	2e-5, 6e-5, .0002	7e-5, 8e-5, .0003

Table 2: Simulation Results

interval for $p(E_1|E_2)$ with confidence level of at least $1-2\epsilon$.

These confidence intervals allow users to use associations rules predictively by giving them an idea of how much variance they can expect in the support and confidence of a rule in the future.

3. Conclusions

We looked at the issue of whether association rule algorithms produce many "false discoveries". It is straightforward to compute the statistical significance of a single rule. However, when looking at a set of rules, the significance test has to take into account the number of hypotheses being tested. We showed that the number of hypotheses implicitly being tested can be much greater than the number of output rules, and derived an upper bound for the number of hypotheses. Unfortunately deriving an acceptance threshold for the statistical significance test from this bound may be too conservative. We presented a novel approach of using resampling to determine the acceptance threshold for the significance test. The threshold value derived using this approach was typically more than 100 times greater than the threshold value derived from the upper bound.

We then used this threshold to evaluate the number of "false discoveries" on three real-life dataset. We found that less than 0.1% of the rules were false discoveries: the reason for this surprisingly low number is that the minimum support and confidence constraints already do an excellent job of pruning away the statistically insignificant rules. A bonus of this work is that the statistical significance measures we compute are a good basis for ordering the rules for presentation to users, since they correspond to the statistical "surprise" of the rule.

Finally, we derived confidence intervals for the support and confidence of an association rule, enabling users to use the rule predictively over future data.

References

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, A. I. 1996. Fast Discovery of Association Rules. In Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press. chapter 12, 307–328.

Agrawal, R.; Imielinski, T.; and Swami, A. 1993. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Conference on Management of Data*, 207–216.

Ali, K.; Manganaris, S.; and Srikant, R. 1997. Partial Classification using Association Rules. In *Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining*.

Brin, S.; Motwani, R.; Ullman, J. D.; and Tsur, S. 1997. Dynamic itemset counting and implication rules for market basket data. In *Proc. of the ACM SIGMOD Conference on Management of Data*.

Brin, S.; Motwani, R.; and Silverstein, C. 1997. Beyond market baskets: Generalizing association rules to correlations. In *Proc. of the ACM SIGMOD Conference on Management of Data*.

Cryer, J., and Miller, R. 1994. *Statistics for Business*. Belmont, California: Duxbury Press.

Hochberg, Y., and Tamhane, A. 1987. *Multiple Comparison Procedures*. New York: Wiley.

Nearhos, J.; Rothman, M.; and Viveros, M. 1996. Applying data mining techniques to a health insurance information system. In *Proc. of the 22nd Int'l Conference on Very Large Databases*.

Piatetsky-Shapiro, G. 1991. Discovery, analysis, and presentation of strong rules. In Piatetsky-Shapiro, G., and Frawley, W. J., eds., *Knowledge Discovery in Databases*. Menlo Park, CA: AAAI/MIT Press. 229–248.