A Comparison of Standard and Interval Association Rules*

Choh Man Teng
cmteng@ai.uwf.edu
Institute for Human and Machine Cognition
University of West Florida
40 South Alcaniz Street, Pensacola FL 32501, USA

Abstract

The standard formulation of association rules is suitable for describing patterns found in a given data set. A number of difficulties arise when the standard rules are used to infer about novel instances not included in the original data. In previous work we proposed an alternative formulation called *interval association rules* which is more appropriate for the task of inference, and developed algorithms and pruning strategies for generating interval rules. In this paper we present some theoretical and experimental analyses demonstrating the differences between the two formulations, and show how each of the two approaches can be beneficial under different circumstances.

Standard Association Rules

One of the active research areas in data mining and knowledge discovery deals with the construction and management of association rules. We will call the formulation typified in (Agrawal, Imielinski, & Swami 1993) the *standard* formulation. A *standard association rule* is a rule of the form

$$X \Rightarrow Y,$$

which says that if X is true of an instance in a database Δ, so is Y true of the same instance, with a certain level of significance as measured by two indicators, *support* and *coverage*:

[support] proportion of XYs in Δ;

[coverage] proportion of Ys among Xs in Δ.

(Note that "*coverage*" is typically called "*confidence*" in the standard association rule literature. However, we will be using "confidence" to denote the level of certainty associated with an interval derived from a statistical procedure. To avoid confusion, we will refer to the above measure of rule accuracy as the *coverage* of the rule, and restrict the use of the word "confidence" to terms such as "the confidence interval" as are traditionally used in statistics.)

The goal of standard association rule mining is to output all rules whose support and coverage are respectively above some given support and coverage thresholds. These rules encapsulate the relational associations between selected attributes in the database, for instance,

coke $\Rightarrow$ potato chips: 0.02 support; 0.70 coverage (^{*1})

denotes that in our database 70% of the people who buy coke also buy potato chips, and these buyers constitute 2% of the database. This rule signifies a positive (directional) relationship between buyers of coke and potato chips.

We would like to go from an observation obtained from a data sample, such as (^{*1}), to an inference rule about the population at large, such as

"buying coke is a good predictor for buying potato chips."

A number of difficulties arise if we are to take association rules, as typically formulated, to be rules of inference. These difficulties stem from a fundamental difference in how these rules are conceived. We will examine this distinction and some of its implications in the following.

The Difference between Description and Inference

There are many reasons to abstract rules of association from a data set. For example, we may wish to extract an intensional description of the data set, or we may want to use the insights provided by the rules obtained from the data set as a guide to similar occurrences in the world at large. We argue that standard association rules are descriptive, tailored for the first task, while interval association rules are inferential, intended for the second task.

For the first task mentioned above, the standard formulation of association rules is appropriate for the kind of information we seek. Standard association rules present a description of the patterns found among the attributes, as manifested by the instances that belong to the *existing data set*. This is useful when we need a succinct summary of the data set in lieu of a listing of all the instances. However, these rules may not be directly applicable to describe patterns in the instances that are not part of the given data set. This latter usage is instead part of the second task.

For this second task we would like the rules we derive from the given data set to be indicative of the patterns that can be found in a much larger domain. The target domain may be the list of *all* potential transactions, for example, including both those performed by current customers who did

*This work was supported by NASA NCC2-1239 and ONR N00014-01-1-0926.

not shop on the day we collected the data, as well as those that will be performed by customers who have not shopped here yet but one day (in the not too distant future) will. The target population thus is not only much larger than the given data set, but also is typically of infinite cardinality. Examining the whole population is out of the question in many cases of interest, both because of the size of the population, and also because many members of the population, such as those that exist in the future, cannot be available for examination at any cost.

This is not to suggest that the descriptive task is trivial. In general discovering a pattern that characterizes unstructured data is as difficult a task as making predictions about unseen instances. We are merely pointing out that the two problems are distinct, with different issues that need to be addressed. For instance, consider the notion of "interestingness" or "unexpectedness". A rule that says

$$\text{being in kindergarten} \Rightarrow \text{less than 10 years old}$$

is likely to be highly accurate, but not very surprising. The value of this rule thus would be low in a descriptive context, from a knowledge discovery perspective, while its utility as a predictive rule in an inferential context may very well be much higher.

Making Inferences with Association Rules

The distinction between description and inference is a point worth emphasizing as often the implicit goal of an association rule mining session is inferential rather than descriptive. We look for rules that are expected to hold in a population that typically extends into the inaccessible (for the moment) future, and in any case is far greater than the sample data set we gathered the rules from.

A number of considerations make it unattractive to adopt standard association rules as inference rules mechanically. The former rules are abstracted from a given sample data set, while the latter rules are to be applicable to a larger population.

From Sample to Population

First, we need to take into account variations inherent in the process of sampling. Although the larger the sample size, the higher the proportion of samples that resemble the population, any given sample of any given size is unlikely to have exactly the same characteristics as the population. Giving an exact point value for the rule support and coverage parameters can easily convey an illusion of certainty of the findings.

Note that for the purpose of statistical inference, the sample relevant to a rule $X \Rightarrow Y$ is not the whole given data set Δ, but only that portion of Δ containing X. Thus, even from a single given data set, different rules may require the consideration of different samples (portions of Δ). In addition, the central limit theorem is based on the *absolute number* of instances in a sample, not the proportion it constitutes of the parent population (which is typically infinite).

This cannot be easily modelled in the standard association rule framework. Standard rules of the same coverage are considered to be of equal standing unless we also take note of their respective sample sizes. The support of a rule $X \Rightarrow Y$ is construed as the *proportion* of XYs in Δ. This proportion is irrelevant here. What we need is the *absolute number* of instances of X in order to establish the degree of certainty, or statistical confidence, concerning the inference from the rule coverage in a sample to the rule coverage in the parent population.

Evaluation

Mining standard association rules is a clearly defined task. The objective there is to generate all rules of the form $X \Rightarrow Y$ which are above some given support and coverage thresholds. The problem of evaluation and validation is thus reduced to one of correctness and efficiency. Correctness in this case is unambiguous. Any algorithm is required to return *the* set of rules meeting the given criteria. Since there is no difference between the set of rules returned by one algorithm and the next, much of the research effort in this area has been understandably focused on efficiency issues, aiming to overcome the challenges imposed by the tremendous size of the data sets involved and the potential number of rules that can be generated. (Mannila, Toivonen, & Verkamo 1994; Savasere, Omiecinski, & Navathe 1995; Agrawal *et al.* 1996; Zaki *et al.* 1997, for example)

In those cases where variations to the standard framework are investigated, the refinements are mostly restricted to imposing additional constraints on top of the support and coverage criteria to pick out the more interesting and relevant rules from the huge pool of acceptable rules. Alternative measures to determine the fitness of a rule include, for instance, correlation, gain, Gini, Laplace, χ^2, lift, and conviction. These metrics provide grounds to pre- or post-prune the standard association rules in order to arrive at a smaller set of rules. (Silberschatz & Tuzhilin 1996; Brin, Motwani, & Silverstein 1997; Bayardo & Agrawal 1999; Liu, Hsu, & Ma 1999, for example).

The several measures that have been used in the standard association rule literature are not entirely satisfactory as indicators of the quality of an inference rule. *Correctness* is not as well defined in the case of inference. *Efficiency* and the *quantity* of rules are important, but they should be supplementary to a measure of the substance of the rules. *Interestingness*, as we have already noted, is relevant for description but not as much of a concern for inference. In this paper we employ a measure from first principles, namely, comparing rules known to exist (probabilistically) in the parent population to rules obtained from a data set sampled from this population.

Interval Association Rules

The task of deriving predictive rules can be construed as a statistical inference problem. The parent population (all potential transactions) is typically very large if not infinite, and the data set we have at hand (transactions recorded on a given day) constitutes a sample drawn from this population. The problem then can be cast as the problem of projecting the associations found in the sample to justifiably probable associations in the parent population.

In (Teng & Hewett 2002) we advanced the interval association rule framework as an approach to deriving associations that is grounded in the theory of statistical inference. Instead of point-based rules that have to satisfy some minimum coverage and support in the given data set, association coverage is given in terms of an interval, encompassing a range of values in which we can claim the true rule coverage in the parent population falls with a certain level of confidence.

What sets our approach apart is that instead of using statistical measures as a descriptive summary of the characteristics in the sample data set, or as a way to select a subset of more relevant rules from the exhaustive set of standard association rules, or as a handle to deal with numeric data, we relate explicitly the rule coverage in the given sample data to the rule coverage in the population

Perhaps the work that is closest in spirit to our approach is that of (Suzuki 1998), where a production rule (a rule with a single-atom consequent) is considered reliable if its generality and accuracy measures are above certain constant thresholds based on a statistical confidence interval. Our approach can be regarded as a more general formulation with respect to association rules. Confidence intervals are formally attached to the rules, allowing for more informative ways of rule selection and pruning.

Specification

Let us briefly summarize the formulation of interval association rules. Details of the framework as well as algorithms and pruning strategies pertaining to the computational aspects of interval rules can be found in (Teng & Hewett 2002).

Let $A = \{a_1, \ldots, a_m\}$ be a set of m binary attributes. Let Δ be a data set of transactions, each transaction being a subset of A. For a set of attributes $X \subseteq A$, let $\#(X)$ denote the number of transactions containing X in the data set, that is, $\#(X) = |S|$, where $S = \{\delta \in \Delta : X \subseteq \delta\}$. Similarly, for $Y \subseteq A$ and $z \in A$, let $\#(XY)$ denote $\#(X \cup Y)$ and $\#(Xz)$ denote $\#(X \cup \{z\})$.

An *interval association rule* is a rule of the form

$$X \Rightarrow Y \qquad [l, u] : 1 - \alpha, \qquad (^*2)$$

where $X, Y \subseteq A$, $X \cap Y = \emptyset$, and l, u, and α are all real numbers in the interval $[0, 1]$. The goal of our interval rule mining exercise is to assemble an appropriate set of interval association rules, such that for each rule of the form (^{*2}), the proportion of transactions containing Ys in those containing Xs in the parent population is in the interval $[l, u]$ with confidence $1 - \alpha$.

Using the normal approximation to the binomial distribution, the interval association rule in (^{*2}) can be rewritten as

$$X \Rightarrow Y \qquad [p - e,\ p + e] : 1 - \alpha, \qquad (^*3)$$

where (when $\#(X) > 0$)

$$p = \frac{\#(XY)}{\#(X)}; \qquad e = z_\alpha \sqrt{\frac{p\,(1-p)}{\#(X)}}.$$

In the above z_α is the (two-tailed) z-score determined from the confidence parameter α. (For example, $z_{0.05}$ is 1.96.) Note that the value $\#(X)$ used in the calculation of p and e above is *not* the size of the whole data set $|\Delta|$. Rather, $\#(X)$ is the number of occurrences of the antecedent of the rule in question.

For simplicity we will omit the confidence parameter $1-\alpha$ from the rule specification in the following discussion.

When to Adopt Which Approach?

One might ask why we bother with the confidence interval at all, even if it is backed by some interesting statistical theory. This is especially true considering that the sample rule coverage p is always included in the interval. What do we gain by adopting the interval structure?

We argued that the standard approach is descriptive, while the interval approach is inferential. In practice, however, the standard approach has been widely used in many situations, for both description and inference, with little appreciable difficulty. Let us see why this is the case through an example, and then we will see through another example why in some other situations the standard approach is inadequate.

The Case for the Standard Approach

Consider the following rule.

$r : x \Rightarrow y$, where $\Pr(x) = 0.05$ and $\Pr(y \mid x) = 0.6$.

That is, the true coverage of rule r in the population is 60%. Now consider samples of size 10,000 drawn from this population. We expect that in at least 92% of these samples (taking into account the variation in the numbers of both x and y) the coverage of r is in the interval $[0.5551, 0.6449]$. Depending on the utility and the sensitivity of the application, the width of this interval may not worry the user too much. (Is it 56% or 61% that consumers who buy beach balls also buy frisbees?) Thus, the sample rule coverage may be considered a practical approximation of the true population rule coverage. This is especially true when the sample size in question is large, which is bolstered in the standard association rule framework by the combined effect of huge data sets and a reasonably high support threshold.

The Case for the Interval Approach

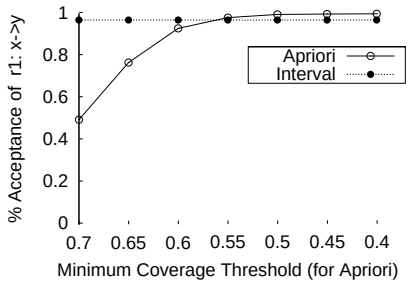
While the standard approach suffices in many situations, there are cases where the additional inferential power of interval association rules is desirable. For example, instead of beach balls and frisbees, we are considering yachts and real estate. Association rules involving commodities that are relatively rarely purchased but of high stakes would be of great utility. In addition, by taking into account the relevant sample sizes, interval association rules are better able to discriminate between rules that are justified and rules that are inconclusive. Let us illustrate with some experiments.

Experiments

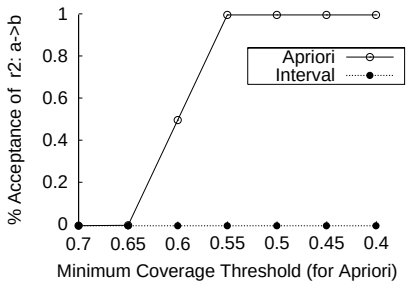
Consider two rules

$r_1 : x \Rightarrow y$, where $\Pr(x) = 0.005$ and $\Pr(y \mid x) = 0.7$;
$r_2 : a \Rightarrow b$, where $\Pr(a) = 0.1$ and $\Pr(b \mid a) = 0.6$.

Suppose the user specified a minimum coverage threshold p^* of 0.7. That is, we would like to accept rules whose coverage in the *population* at large is at least p^*, and reject all



(a) $r_1 : x \Rightarrow y$



(b) $r_2 : a \Rightarrow b$

Figure 1: Percentage of acceptance of rules r_1 and r_2 over 1000 runs, varying the minimum coverage threshold used by Apriori (this parameter does not affect interval rules). We would like to accept r_1 but reject r_2.

others. According to this threshold, we would like to be able to accept r_1 but reject r_2. In other words, the ideal acceptance rates of the two rules are 100% and 0% respectively.

This scenario was evaluated experimentally. We considered sample data sets Δ of size 10,000, drawn randomly from a population with the above distribution constraints for the attributes x, y, a, and b. The interval approach was compared to the standard approach using an algorithm such as Apriori (Agrawal *et al.* 1996).

Given a hypothetical rule coverage p^*, the 95% confidence interval with respect to a sample size $\#(x)$ is $p^* \pm 1.96\sqrt{p^*(1-p^*)/\#(x)}$. In the interval framework we required that rule r_1 be accepted if its actual sample coverage (the ratio between $\#(xy)$ and $\#(x)$ in the sample Δ) was greater than $p^* - 1.96\sqrt{p^*(1-p^*)/\#(x)}$, and rule r_2 be accepted if its actual sample coverage was greater than $p^* - 1.96\sqrt{p^*(1-p^*)/\#(a)}$. This gave us a 97.5% confidence[1] that the sample Δ has been drawn from a population in which the true coverage of an accepted rule (either r_1 or r_2) is at least p^* (0.7 in our experiments).

In the standard framework, without degrading the qualitative performance, the minimum support threshold for Apriori was held deliberately low at 0.01%. We successively lowered the minimum coverage threshold for Apriori from 0.7 down to 0.4. The results over 1000 runs each are shown in Figures 1 and 2.

[1]97.5% corresponds to the area under the standard normal curve in the *one-tailed* interval $(-z, +\infty)$.

Figure 1 shows the percentage of runs in which each of the two rules was accepted. On the interval approach, rule r_1 was accepted 97.1% of the time, while rule r_2 was never accepted. For Apriori, with the minimum coverage threshold set at 0.7, the acceptance rates of the two rules were 49.8% and 0% respectively.

We investigated the effect of lowering the minimum coverage threshold. Figure 1 shows that the lower the threshold, the more often Apriori accepted *both* r_1 and r_2. With a threshold of 0.6, the acceptance rate of r_1 has risen to 93.1%, but at the same time r_2 was also accepted in 50.1% of the runs. Lowering the threshold further, both rules were accepted most of the time, and in some cases r_2 was accepted even (slightly) more often than r_1.

These results are further broken down into four cases in Figure 2, based on the combination of rules that were accepted in each run: (a) both rules were rejected; (b) r_1 was rejected and r_2 accepted; (c) r_1 was accepted and r_2 rejected; and (d) both rules were accepted. The case of particular interest is shown in Figure 2(c), where the two rules received their desirable respective treatments.

Interval rules achieved this desirable scenario 97.1% of the time (in the remaining 2.9% both rules were rejected). For Apriori, as we lowered the minimum coverage threshold, the percentage of the desired outcome rose from 50.2% to 76.8%, but then dropped eventually to 0%. This slack was taken up in Figure 2(d), which shows a sharp rise in the percentage of runs in which both rules were accepted. In other words, as we lowered the minimum coverage threshold for Apriori, r_1 was accepted more often, but at the expense of also accepting the undesirable r_2. Apriori lacks the mechanism to discriminate between the circumstances of the two rules.

Conclusion

We have presented some theoretical and experimental analyses comparing the standard and interval approaches to association rule mining. The standard formulation is geared toward description, while the interval formulation is geared toward inference. Under certain circumstances, the two formulations behave similarly. However, there are cases in which the additional inferential power of the interval framework is beneficial.

The interval formulation can make finer distinctions between inequivalent scenarios that are treated indifferently in the standard formulation, where the minimum coverage threshold (let us put aside the minimum support criterion) dictates that for all rules with the same coverage level, we either accept them all or reject them all. The standard approach does not discriminate between the situation where the sample size is small, in which case the sample rule coverage can be expected to have a large spread, and the situation where the sample size is large, in which case the rule coverage of samples drawn from an identical population would be more closely clustered around the population mean.

Although we can of course approximate such differentiation of rules in the standard framework by devising a goodness measure based on a heuristic combination of the support and coverage values, the interval formulation provides

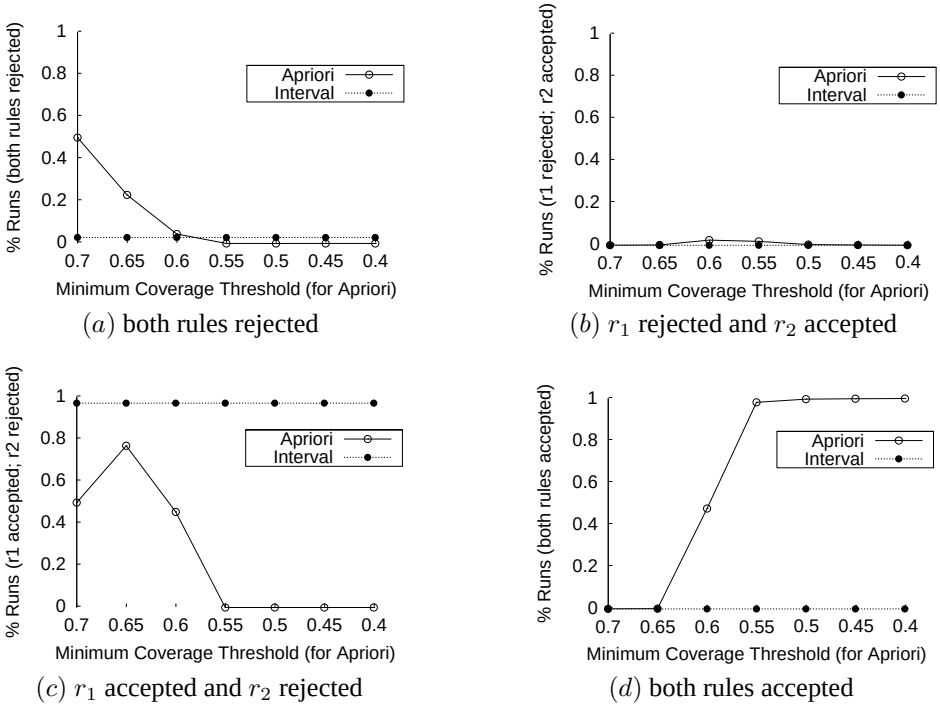


Figure 2: Percentage of occurrences of the four cases over 1000 runs: (a) both rules rejected; (b) r_1 rejected and r_2 accepted; (c) r_1 accepted and r_2 rejected; (d) both rules accepted. Note that case (c) is the desired outcome.

a more principled basis for making normative choices based on a formal statistical theory. The crux of the problem lies in distinguishing between a rule that is justifiably acceptable and one whose supporting evidence is inconclusive. The interval formulation achieves this differentiation by taking into account the inherent uncertainty associated with the task of inferring the characteristics of a population from the characteristics of a sample.

References

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, A. 1996. Fast discovery of association rules. In Fayad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press. 307–328.

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

Bayardo, R., and Agrawal, R. 1999. Mining the most interesting rules. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 145–154.

Brin, S.; Motwani, R.; and Silverstein, C. 1997. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the ACM SIGMOD Conference on the Management of Data*, 265–276.

Liu, B.; Hsu, W.; and Ma, Y. 1999. Pruning and summarizing the discovered associations. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 125–134.

Mannila, H.; Toivonen, H.; and Verkamo, A. 1994. Efficient algorithms for discovering association rules. In *KDD-94: AAAI Workshop on Knowledge Discovery in Databases*, 181–192.

Savasere, A.; Omiecinski, E.; and Navathe, S. 1995. An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21st Conference on Very Large Databases*, 432–444.

Silberschatz, A., and Tuzhilin, A. 1996. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering* 8(6):970–974.

Suzuki, E. 1998. Simultaneous reliability evaluation of generality and accuracy for rule discovery in databases. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*, 339–343.

Teng, C. M., and Hewett, R. 2002. Associations, statistics, and rules of inference. In *Proceedings of the International Conference on Artificial Intelligence and Soft Computing*, 102–107.

Zaki, M. J.; Parthasarathy, S.; Ogihara, M.; and Li, W. 1997. New algorithms for fast discovery of association rules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 283–296.