Pre-publication draft of a paper which appeared iin Proceedings of the 2003 Australasian Datamaing Workshop AUSDM03 pp 1-9. Publisher: University of Technology, Sydney

Preliminary investigations into statistically valid exploratory rule discovery

Geoffrey I. Webb

School of Computer Science and Software Engineering, Monash University,
Melbourne, Vic 3800, Australia
Email: webb@csse.monash.edu.au
Telephone: +61 3 9905 3296
Fax: +61 3 9905 5146

Abstract. Exploratory rule discovery, as exemplified by association rule discovery, is has proven very popular. In this paper I investigate issues surrounding the statistical validity of rules found using this approach and methods that might be employed to deliver statistically sound exploratory rule discovery.

Keywords: exploratory rule discovery; association rule discovery; k-most-interesting rule discovery; statistically sound rule discovery

1 Introduction

Association rule discovery has proven very popular. However, it is plagued by the problem that it often delivers unmanageably large numbers of rules. As the current work reveals, not only are the rules numerous, but in at least some cases the vast majority are spurious or unproductive specialisations of more general rules. This paper discusses the issues of spurious and unproductive rules and presents preliminary approaches to address them. Experimental results confirm the practical realisation of the concerns and suggests that the preliminary techniques presented are effective.

2 Exploratory rule discovery

I use the term *exploratory rule discovery* to encompass data mining techniques that seek multiple rather than single models, with the objective of allowing the end-user to select between those models. It is distinguished from *predictive data mining* that seeks a single model that can be used for making predictions.

Exploratory data mining is often applicable when there are factors that can affect the usefulness of a model but it is difficult to quantify those factors in a manner that may be used by an automated data mining system. By delivering multiple alternative models to the end-user they are empowered to evaluate the available models and to select those that best suit their business or other objectives.

Three prominent frameworks for exploratory rule discovery on which I here focus are *association rule discovery* [1], *k-most-interesting rule discovery* [6] and *contrast* or *emerging pattern discovery* [4, 2] as it is variously known. These techniques all discover qualitative rules, rules that represent relationships between nominal-valued variables.

Each such rule $A \rightarrow C$ represents the presence of an interesting relationship between the antecedent A and the consequent C, where A is a conjunction of nominal-valued terms and C is a single nominal valued term[1]. The rules are usually presented together with statistics that describe the relationship between A and C.

2.1 Association rule discovery

Association rule discovery [1] is the most widely deployed exploratory rule discovery approach. It grew out of market-basket analysis, the analysis of transaction data for combinations of products that are purchased in a single transaction. Association rule discovery uses the so called *support-confidence framework*. It finds all rules that satisfy a user-specified minimum support constraint together with whatever further constraints the user may specify. Essentially, the approach generates all rules that satisfy the minimum support constraint but discards at the final stage any rules that fail the further constraints.

Support is the proportion of records in the training data that satisfy both the antecedent and consequent of the rule.

Initial approaches used a further constraint on minimum *confidence*. To avoid potential confusion with the statistical concept of *confidence* I will hereafter refer to this metric as *strength*.

$$strength = support/coverage$$

where *coverage* is the proportion of records that satisfy the antecedent.

More recent approaches typically use a constraint on minimum lift in preference to a constraint on strength:

$$lift = strength/prior$$

where *prior* is the proportion of records that satisfy the consequent.

A limitation of association rule discovery is that the main control over the number of rules that are discovered is the value that is specified for minimum support. However, it is usually difficult to anticipate which values of minimum support will result in manageable numbers of rules. Too large a value will result in no rules. Too small a value will result in literally millions of rules. In practice there may be a very narrow range of values of support below which there

[1] While association rules are often described in terms of allowing C to be an arbitrary conjunction of terms, in most implementations C is restricted to a single term. In the current work I follow this practice as it greatly reduces the complexity of the rule discovery task while satisfying many rule discovery needs.

are extremely few rules discovered and above which there are too many rules discovered [9].

2.2 K-most-interesting rule discovery

K-most-interesting rule discovery [6] addresses that problem by empowering the user to specify both a metric of interestingness and a constraint on the maximum number of rules to by discovered. In place of a minimum support constraint, k-most-interesting rule discovery uses these two pieces of information to prune the search space. They return the k rules that optimise the interestingness metric within whatever other constraints the user might specify.

2.3 Contrast discovery

Contrast discovery [2] (initially developed under the name *emerging pattern discovery* [4]) identify conditions whose frequency differs between groups of interest. It has been shown that this is equivalent to rule discovery restricted to a consequent that signifies group membership [8].

3 Spurious rules

A problem for all three of these forms of exploratory rule discovery is that they suffer a high risk of discovering *spurious rules*. These are rules that appear interesting on the sample data but which would not be interesting if the true population probabilities were used to assess their level interestingness in place of the observed sample frequencies.

For example, suppose that there is a rule with coverage of one record and a lift of 2.0. This provides very little evidence that the lift that would be obtained by substituting population probabilities for sample frequencies would have high lift as a rule with one record coverage must have either a support of zero or one record and hence, irrespective of the population lift, the observed lift must either be 0.0 or $1.0/prior$. Put in another framework, when the coverage is low the statistical confidence is low that the observed relative frequencies are strongly indicative of the population probabilities.

The support-confidence framework of association rule discovery attempts to counter this problem by enforcing a minimum support constraint in the expectation that considering only rules with high support will lead to the observed frequencies being strongly representative of the population frequencies.

4 The multiple comparisons problem

However, this ignores the problem of multiple comparisons [5]. If many observations are made then one can have high confidence that some events that are unlikely in the context of a single observation are likely to occur in some of the many observations that are made. For example, suppose a hypothesis test is applied to evaluate whether a rule is spurious with a significance level of 0.05.

Consider a spurious rule $A \rightarrow C$ for which A and C are independent. The probability that this spurious rule will be accepted as not spurious 5%. If this process were applied to 1,000,000 rules in a context where all rules were spurious (for example, the data were generated stochastically using uniform probabilities) we could reasonably expect that 5% or 50,000 would be accepted as non-spurious despite all being spurious. In practice the rule spaces explored by rule discovery systems are many magnitudes greater than 1,000,000, and hence we should expect many spurious rules to be generated even if we apply a significance test before accepting each one.

5 Filters for spurious rules

One response to this problem is to apply a correction for multiple comparisons, such as the Bonferroni adjustment that divides the critical value α by the number of rules evaluated. This is the approach adopted in the contrast discovery context by STUCCO [2]. A problem with this approach is that the search process may require the evaluation of very large numbers of rules and hence α may be driven to extremely low values. The lower the value of α the higher the probability of type-2 error, that is, of rejecting rules that are not spurious.

What is required is an approach that minimises the risk of type-1 error, that is, of accepting spurious rules, without in the process discarding the most interesting non-spurious rules.

6 Unproductive rules

A further problem for rule discovery is that of unproductive rules. A rule $A \rightarrow C$ is unproductive if it has a generalisation $B \rightarrow C$ such that $strength(A \rightarrow C) \leq strength(B \rightarrow C)$. An unproductive rule will arise when a variable that is unrelated to either B or C is added to B. As the strength is unaltered, the lift of the unproductive rule will equal that of the generalisation. In practice data sets often involve many variables that do not impact upon the rules of interest and hence very large numbers of unproductive rules are generated.

The problem of unproductive rules interacts with the problem of spurious rules. It is straightforward to add a filter to the rule discovery process that discards any rule for which the observed strength is not greater than the observed strength of all its generalisations. However, random variations in the data sample will lead to almost half the unproductive rules appearing to be productive (albeit in many cases only very slightly). A statistical test of significance may be applied, but we again face the multiple comparisons problem.

7 A new approach

Hypothesis testing is designed for controlling the risk of type-1 error in the context of evaluating a prior hypothesis against previously unsighted data. It is inadequate to the task of both generating hypotheses and evaluating them from the same set of data.

An approach that has been used in other data mining contexts is to use a holdout set for hypothesis testing. Models are inferred from a training set and then evaluated against a holdout set. My proposal is to utilise this framework in an exploratory rule discovery context. The available data will be divided into an exploratory data set from which rules will be discovered. This will be treated as a hypothesis generation process. The rules discovered are treated as hypotheses that are then evaluated against the holdout set. As the holdout data is partitioned from the exploratory data, the huge number of rules considered during rule discovery does not affect the subsequent evaluation. A simple multiple comparisons adjustment need only divide the selected alpha value by the number of rules delivered by the rule discovery phase. Thus the α value need not be set prohibitively low, minimising the problem of type-2 error.

I propose the use of k-most-interesting rule discovery for the rule discovery phase rather than association rule discovery, because it is desirable to find a constrained number of rules during the rule discovery phase. If too many rules are discovered the necessary multiple comparisons adjustment will result in a raised risk of type-2 error. If too few rules are discovered then there is a raised risk of failing to discover sufficient interesting rules to satisfy the user. Standard association rule discovery provides only very imprecise control over the number of rules discovered. Tightening or weakening each of the constraints will respectively decrease or increase the number of rules discovered, but typically it is not possible to predict by exactly how much a particular alteration to the constraints will affect the number of rules discovered. In contrast, k-most interesting rule discovery always returns k rules, except in the unusual circumstance that the other constraints applied are satisfied by fewer than k rules.

7.1 Holdout evaluation tests

For each rule $A \rightarrow C$ we wish to assess whether the observed $strength(A \rightarrow C)$ is significantly higher than would be expected if there were no relationship between the antecedent and consequent and also whether it is significantly higher than the strength of all its generalisations[2] I use a binomial test to assess whether an observed strength is signficantly higher than a comparator strength. The large number of subsets of an antecedent containing many conditions would make testing against all generalisations infeasible. In consequence I test $strength(A \rightarrow C)$ against the sample frequency of C (which equals $strength(\rightarrow C)$) and against the strength of all its immediate generalisations (rules formed by removing a single condition from A). While it is theoretically possible for a rule to have higher strength than all of its immediate generalisations but lower strength than a further generalisation, to do so requires a very specific type of interaction between four or more variables of a form that might make the resulting rule interesting in its own right despite being unproductive with respect to one of its generalisations.

[2] Actually, as $\emptyset \rightarrow C$ is a generalisation of $A \rightarrow C$ the latter condition subsumes the first.

8 Evaluation

The Magnum Opus [7] k-most interesting rule discovery system was extended to support the form of holdout evaluation described above.

I first sought to evaluate what proportion of rules discovered by a traditional association rule approach to rule discovery might be either spurious or unproductive. To this end I investigated rule discovery performance on two large data sets from the UCI repository [3], covtype (581012 records) and census-income (199,523 records).

Each data set was randomly divided into two equal sized subsets, the exploratory data used to discover rules and the holdout data used for holdout evaluation.

I started by seeking to find values of minimum support and minimum lift that resulted in constrained numbers (less than 100,000) of rules. After a number of trials I found for the covtype data that a minimum support of 0.25 and minimum lift of 2.75 resulted in 1997 rules of which 1936 (96.9%) were rejected by holdout evaluation. For the census-income data minimum support of 0.4 and minimum lift of 2.0 resulted in 7502 rules of which 7462 (99.4%) were rejected as spurious or unproductive when assessed against the holdout data. These figures provide a dramatic illustration of the degree to which traditional association rule discovery results may be dominated by rules that are effectively noise.

To separate the issues of unproductiveness from spuriousness, I applied a filter to discard unproductive rules during the rule discovery phase. That is, during rule discovery a rule was discarded if it was unproductive as assessed using the observed strength on the exploratory data without application of a significance test.

With the same support and lift constraints, for covtype 433 rules were found of which 377 (87.1%) were rejected by holdout evaluation. Whereas only 40 rules passed the holdout evaluation when the support confidence framework was employed, when filtering of unproductive rules is added this is raised to 63 rules, as the number of multiple comparisons is reduced and hence the adjusted α value used in holdout evaluation is raised.

When a binomial test was applied during rule discovery to evaluate whether a rule was significantly productive (on the exploratory data), using $\alpha = 0.05$, the number of rules found was further reduced to 73 of which 45 (61.6%) were rejected by holdout evaluation. Note that the number of rules the have passed the holdout test (18) has decreased. This illustrates the problem of filtering so as to adequately balance the risks of type-1 and type-2 error. The filter applied during rule discovery has discarded 45 rules that were found with a weaker filter and then accepted after holdout evaluation.

Applying yet stronger filters, for example by adjusting for multiple comparisons the α used in the statistical test applied during rule discovery, can be expected to improve the proportion of rules that pass holdout evaluation, but to decrease the absolute number of rules that pass.

For census-income when unproductive rules were discarded during the rule discovery phase, 48 rules were discovered of which 8 were rejected by holdout evaluation. This resulted in the same 40 rules passing holdout evaluation as the rule discovery that did not discard unproductive rules during the rule discovery phase. Tightening the filter applied during rule discovery by adding a significance test resulted in the discovery of 45 rules of which 5 were discarded by holdout evaluation, leaving the same 40 rules.

As a final test, I applied k-most-interesting rule discovery in place of the support-confidence framework. As a measure of interestingness I used *leverage*,

$$leverage(A \rightarrow C) = support/coverage(A) \times coverage(C).$$

This represents the difference between the observed join frequency and the joint frequency that would be expected if the antecedent and consequent were independent. I sought the 100 rules that maximised this value without any other constraints other than that all rules had to be significantly productive at the 0.05 level, that is that they had to pass a binomial test at the 0.05 level indicating that they had higher strength than any immediate generalisation. Note that this process did not require the time consuming and error prone business of identifying a suitable minimum support constraint.

For covtype all 100 rules passed the holdout evaluation. All rules found had extremely high support, the lowest being 0.436. The lowest lift was 2.19. It is interesting that the search for the 100 most interesting rules found quite a different trade-off between support and lift than I found during my manual attempt to find a set of constraints that provided sufficiently few rules for consideration, resulting in rules with higher support but lower lift. It is also notable that all rules so found passed holdout evaluation, as the search explicitly sought rules that were most exceptional on the exploratory data and hence the most valuable to evaluate on the holdout data.

For census-income, of the 100 rules found 13 were discarded by holdout evaluation. All rules found had high support, the lowest being 0.413. The lowest lift of a rule was 1.91. This illustrates the difficulty of finding appropriate constraints to apply within the traditional association rule framework, as it lay just outside the minimum lift that I had found after some experimentation in the attempt to return only a constrained number of rules.

For each data set, the k-most-interesting approach to rule discovery delivered higher numbers of statistically sound rules without need for manual determination of appropriate support and other constraints.

9 Conclusion

I have presented an approach to addressing the problems of spurious and unproductive rules in exploratory rule discovery. Two examples have demonstrated that over 99% of rules discovered using the support-lift framework can be spurious or unproductive. I have shown that the use of k-most-interesting rule discovery with holdout evaluation can overcome this problem, delivering for the first time statistically sound exploratory discovery of potentially interesting rules from data.

References

[1] Agrawal, R., Imielinski, T., and Swami, A. Mining associations between sets of items in massive databases. In *Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data* (Washington, DC, May 1993), pp. 207–216.

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

[3] Blake, C., and Merz, C. J. UCI repository of machine learning databases. [Machine-readable data repository]. University of California, Department of Information and Computer Science, Irvine, CA., 2001.

[4] Dong, G., and Li, J. Efficient mining of emerging patterns: Discovering trends and differences. In *ACM SIGKDD 1999 International Conference on Knowledge Discovery and Data Mining* (1999), ACM, pp. 15–18.

[5] Jensen, D. D., and Cohen, P. R. Multiple comparisons in induction algorithms. *Machine Learning 38*, 3 (2000), 309–338.

[6] Webb, G. I. Efficient search for association rules. In *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, 2000), The Association for Computing Machinery, pp. 99–107.

[7] Webb, G. I. Magnum Opus version 1.3. Computer software, Distributed by Rulequest Research, http://www.rulequest.com, 2001.

[8] Webb, G. I., Butler, S., and Douglas, N. On detecting differences between groups. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)* (2003), pp. 256–265.

[9] Zheng, Z., Kohavi, R., and Mason, L. Real world performance of association rule algorithms. In *KDD-2001: Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining* (New York, NY, August 2001), ACM, pp. 401–406.