Understanding the Crucial Differences Between Classification and Discovery of Association Rules – A Position Paper

Alex A. Freitas

Pontificia Universidade Catolica - Parana
Dept. of Computer Science
Rua Imaculada Conceicao, 1155
Curitiba – PR, 80215-901. Brazil
http://www.ppgia.pucpr.br/~alex
alex@ppgia.pucpr.br

ABSTRACT

The goal of this position paper is to contribute to a clear understanding of the profound differences between the association-rule discovery and the classification tasks. We argue that the classification task can be considered an ill-defined, non-deterministic task, which is unavoidable given the fact that it involves prediction; while the standard association task can be considered a well-defined, deterministic, relatively simple task, which does *not* involve prediction in the same sense as the classification task does.

Keywords

Classification, association rules, induction, prediction.

1. INTRODUCTION

Classification and association-rule discovery are probably the two tasks most addressed in the data mining literature. Hence, it is crucial that some fundamental differences between these two tasks be clearly understood.

Unfortunately, this does not seem to be the case. We have observed that there is some confusion about important characteristics of these two tasks in the data mining community. Actually, a confusion between these two tasks (or at least a confusion of terminology) is sometimes even present in papers published at major international conferences. We mention here only two examples of this kind of confusion.

The first example comes from Domshlak et al. (1998). Although this is overall a high-quality paper (like the others published at such a prestigious conference), we have to disagree with a statement made by the authors: "Association Rules are straightforwardly collected from the decision tree." (p. 187.) Clearly, the rules extracted from a decision tree are classification rules, rather than association ones.

The second example comes from Bayardo (1997). Again, this is overall a high-quality paper, but we believe it makes the mistake of calling the association rules discovered by their method "classification rules". The task being addressed in that paper is still association-rule discovery and the rules discovered by their method are still association rules, in the sense that they lack characteristics inherent to classification rules, as will be explained later (particularly in sections 2.2 through 2.5).

The goal of this position paper is to contribute to a clear understanding of the profound differences between the association-rule discovery and the classification tasks. More precisely, this position paper argues that there are crucial differences between the association-rule discovery and the classification task, and that these differences involve the key notion of prediction. We argue that the classification task can be considered an ill-defined, non-deterministic task, which is unavoidable given the fact that it involves prediction; while the standard association task can be considered a well-defined, deterministic, relatively simple task, which does *not* involve prediction in the same sense as the classification task does.

This paper is organized as follows. In section 2 we discuss several differences between the classification and the association-rule discovery tasks. These differences include the following issues: syntactical differences and attribute (a)symmetry (section 2.1); ill-defined, non-deterministic vs. well-defined, deterministic tasks (section 2.2); overfitting and underfitting (section 2.3); and inductive bias (section 2.4). We also argue, in section 2.5, that generally previous work on integrating classification and association-rule discovery still can be categorized as solving either the classification or the association task, but not both. Finally, section 3 discusses some implications of the ideas discussed in previous sections and concludes the paper.

2. DIFFERENCES BETWEEN THE CLASSIFICATION AND THE ASSOCIATION-RULE DISCOVERY TASKS

Throughout this section we refer to the standard framework of association rule discovery (hereafter referred to as the standard association framework, for short). By standard association framework we mean the well-known support-confidence framework introduced by Agrawal et al. (1993), in which the algorithm discovers all association rules having support and confidence greater than user-specified thresholds.

It should be noted that recently there have been several proposals for extending this standard framework, and some of these proposals blur the distinction between association and classification. We do *not* claim that our arguments generalize to all such extended association frameworks (although we believe they still hold for some of these extended frameworks). We focus on the standard association framework mainly because it is still the most used in the literature. In addition, a comprehensive

discussion of those extended frameworks would require an amount of space far greater than the available in this short position paper. In any case, we will say a few words about some extended association frameworks later.

2.1 Syntactic Differences and Attribute (A)symmetry

Probably the most obvious difference between classification and association rules is on a syntactical level. Classification rules have only one attribute in their consequent (THEN part), whereas association rules can have more than one attribute in their consequent.

In addition, the classification and association tasks can also be distinguished according to the (a)symmetry of the attributes being mined. One can say that classification is asymmetric with respect to attributes, since in this task we aim at predicting the value (class) of a special, user-defined goal attribute based on the values of all the other (predictor) attributes. By contrast, one can say that the association task is symmetric with respect to attributes, since no attribute is given special treatment in this task - i.e. any attribute can occur either in the rule antecedent or in the rule consequent.

Granted, if we consider only these two difference criteria, we can blur the distinction between the two kinds of rules by discovering only a subset of association rules, namely the ones having just a value of the goal attribute (a class) in their consequent – as in done e.g. by Liu et al. (1998). However, these two simple difference criteria are only the beginning of the story. We now move on to discuss more profound, "semantic" differences, which have to do with the core of the nature and purpose of these tasks.

2.2 Ill-Defined, No n-Deterministic vs. Well-Defined, Deterministic Tasks

Classification is an *ill-defined, non-deterministic* task, in the sense that in general, using only the training data, one *cannot* be sure that a discovered classification rule will have a high predictive accuracy on the test set, which contains examples *unseen* during training. (There are, however, theoretical bounds on test set error for some classifiers, such as support vector machines – see e.g. Burges (1998), under certain conditions.)

Another way of putting this is to consider that in classification we are essentially using data about "the past" (the training set) to induce rules about "the future", i.e. rules that predict the value that a goal attribute will take on for an example to be observed. Clearly, predicting the future is a non-deterministic problem.

Yet another way of understanding the non-determinism of classification is to recall that classification can be regarded as a form of induction, and that induction (unlike deduction) is not truth-preserving. To see why induction is ill-defined and non-deterministic, consider for instance the inductive task of predicting which is the next number in the following series: 1, 4, 9, 16, ?. (We suggest the reader actually spends a couple of minutes trying to predict the next number in the series, before moving on.)

The reader will probably have guessed 25, after inducing that the generator polynomium is n^2. However, the correct answer is 20, because the generator polynomium, borrowed from Bramer (1996), is: $(-5n^4 + 50n^3 - 151n^2 + 250n - 120) / 24$. There are, of course, many other polynomia which could be the correct answer, since, strictly speaking, there is an infinite number of curves passing through a finite, small number of points. In other words, there is a virtually infinite number of hypotheses consistent with a training set, but the vast majority of them will make a wrong prediction on the test set. Clearly, we humans have a bias favoring the simpler hypothesis, but this is no guarantee that the simpler hypothesis will make the correct prediction – see e.g. Domingos (1998) for an excellent discussion about this point.

In passing we note that several AI-related tasks are also ill-defined and non-deterministic. This characteristic is inherent to pattern recognition problems, such as vision – see e.g. Pinker (1997).

In contrast to classification-rule discovery, association-rule discovery is a well-defined, deterministic task. By definition, any association algorithm must discover precisely the same rule set, i.e. the set of *all* rules having support and confidence greater than a user-specified threshold, without exception. Hence, all association algorithms have the same effectiveness – i.e. they discover the same rule set. The differences in the proposed algorithms concern mainly their relative efficiency – i.e. some algorithms are faster than others.

There are well-defined, deterministic algorithms for finding association rules, so there is no need to use non-deterministic search methods – such as neural networks, genetic algorithms, etc. – in this task. (Recall that we are considering the standard association framework. Of course there are plenty of opportunity for non-deterministic search methods in more complex versions of the association task.)

2.3 Overfitting and Underfitting

In essence, overfitting occurs when the induced model (e.g. a rule set) reflects idiosyncrasies of the particular data being mined that are not reliable generalizations for the purpose of predictions involving new data, whereas underfitting is the dual problem.

An important distinction between the classification and the association task is that overfitting/underfitting avoidance is a crucial concern in the former, but not in the latter.

Actually, since the possibility of overfitting/underfitting is one of the reasons why the classification task is so hard – see e.g. Schaffer (1993) – most rule induction algorithms performing classification have quite elaborated procedures to (try to) avoid overfitting/underfitting – see e.g. Breslow & Aha (1997), Jansen & Schmill (1997), Oates & Jensen (1998), Jensen & Cohen (2000).

These elaborated procedures are necessary because, of course, if the discovered rules are overfitting/underfitting the training data this will lead to a degradation of predictive performance on the *unseen* examples of the test set, which is what we really care about in prediction (even though in data mining we also care about rule comprehensibility and interestingness).

By contrast, overfitting/underfitting issues are largely ignored in the specification of an algorithm for discovering association rules. Actually, one can say that overfitting/underfitting issues are *not* a problem for data mining algorithms in the discovery of association rules. In this task the algorithm simply finds all rules with support and confidence greater than user-specified thresholds, regardless of whether or not the rules would be overfitting/underfitting the data.

Indeed, in the standard association framework, we do not even evaluate the discovered association rules on an unseen test set. Hence, in principle we cannot even detect that overfitting/underfitting has occurred.

Perhaps, making the role of devil's layer, one could say that the standard association framework has at least a crude mechanism to avoid overfitting, namely the specification of a minimum support for the discovered rules. We do not find this argument convincing. The "mechanism" is just a comparison of a rule's support with a *user-specified* threshold, which involves much less autonomy and much less sophistication than the overfitting/underfitting-avoidance procedures usually found in classification algorithms. In addition, the minimum support threshold is specified for all rules, regardless of the items occurring in the rule, which is clearly undesirable in many cases, as argued by Liu et al. (1999a). Liu et al. propose that we modify the standard association framework in such a way that the user specifies a minimum support for each item, so that the minimum support for a given rule is a function of the items occurring in the rule. We believe this is a step in the right direction, but there is still a long way to go to make association-rule discovery algorithms more flexible.

2.4 Inductive Bias

Let us briefly recall the important concept of inductive bias, which is well-known by the classification community but relatively less well-known by the association-rule discovery community – for a good reason, as will seen below.

As pointed out by Michalski (1983), given a set of observed facts (data instances), the number of hypotheses – e.g. classification rules - that imply these facts is potentially infinite. Hence, a classification algorithm *must* have an *inductive bias*. An inductive bias can be defined as any (explicit or implicit) basis for favoring one hypothesis over another, other than strict consistency with the data being mined – see Mitchell (1980, 1997). Note that without an inductive bias a classification algorithm would be unable to prefer one hypothesis over other consistent ones. In machine learning terminology, a classification algorithm without an inductive bias would be capable of performing only the simplest kind of learning, namely rote learning.

We emphasize here a well-known fact about classification. Any bias has a *domain-dependent* effectiveness. Since every classification algorithm has a bias, the performance of a classification algorithm strongly depends on the application domain. In other words, claims such as "classification algorithm A is better than data mining algorithm B" should only be made for a given (or a few) application domain(s). This has been shown both theoretically – see Schaffer (1994), Rao et al. (1995), Domingos (1998) - and empirically – see Michie et al. (1994), King et al. (1995).

Now, what is the inductive bias of an association-rule discovery algorithm? None - at least in the standard association framework, which is the focus of this paper. After all, an association algorithm simply returns *all* the rules having support and confidence greater than user-specified thresholds. Among all these rules, the algorithm has no criterion (no bias) to select one rule over another. Once more, similarly to the point discussed at the end of the previous section, perhaps one could make the role of devil's layer and argue that the minimum support and minimum confidence specified by the user define the "inductive bias" of the algorithm. Again, we do not find this argument very convincing, since these thresholds are defined by the user. Perhaps one could say that these thresholds represent the bias of the *user*, rather than the bias of the algorithm.

2.5 Integrating Classification and Association Rule Discovery

Granted, there has been some work on integrating classification and association rule discovery. However, we argue that in general these projects can be better described as performing one of the two tasks (classification or association), but not both. We briefly discuss below two projects on this kind integration.

Liu et al. (1998) propose a CBA (Classification Based on Association) algorithm. Their work adapts the framework of association-rule discovery to the classification task. For instance, their algorithm discovers a subset of association rules, namely the association rules having only the class attribute in their consequent, and uses a classification-rule pruning method to reduce the number of generated rules. The modifications make sense because the rules will eventually be used as classification rules. The discovered rules are used to predict examples in the test set, and the pruning method helps avoiding overfitting. Hence, the task being solved is classification. The fact that the classification algorithm uses the results produced by an association algorithm does not modify the fact that the problem being solved is classification.

As another example, Bayardo (1997) proposes several pruning strategies to control the combinatorial explosion involved in mining association rules from "classification data sets" – i.e. data sets with a well-defined class attribute and usually used for evaluating classification (rather than association) algorithms. However, as we mentioned in the introduction, in our opinion this work essentially addresses the association-rule discovery task, despite its use of "classification data sets" and despite its claim of discovering "classification rules". The discovered *association* rules are evaluated concerning the coverage of the data set, but their performance on an unseen test set is not measured. Classification goes beyond coverage of the data being mined, it involves *prediction*, as discussed above, and this issue is not addressed by Bayardo (1997).

3. DISCUSSION

We have argued that classification and association-rule discovery are fundamentally different data mining tasks. The former involves prediction and induction, whereas the second involves neither prediction nor induction.

Furthermore, we argue that, if we are to seriously consider using association rules for prediction purposes, we would have to modify the association-rule framework in at least two ways. First, we would need to extend association-rule discovery algorithms with some procedure to avoid overfitting/underfitting. We believe that the idea of automatically varying the minimum support, as proposed by Liu et al. (1999a), is a step in the right direction, but we still need to go further to make association algorithms more flexible. Second, we would need to evaluate discovered association rules on an *unseen* test set. This is a basic requirement for evaluation of any kind of prediction rule.

Now, suppose we develop an association-rule discovery algorithm based on the two above extensions to the standard association

framework. What are we left with? Note that we are not performing a classification task, since we can still discover an association rule predicting several attributes, whereas a classification rule has a single goal attribute in its consequent. However, it is not so easy to distinguish this extended association-rule discovery algorithm from a "generalized rule induction" algorithm, such as ITRULE – see Smyth & Goodman (1991). ITRULE performs a data mining task that can be called dependence modeling. In this task there are several goal attributes to be predicted, so that different rules can predict different goal attributes, but discovered rules are evaluated on an unseen test set.

We could, of course, extend an association rule algorithm with overfitting/ underfitting-avoidance procedures, evaluating discovered rules on an unseen test set and discover only association rules having a single goal attribute in its consequent. But in this case we would be essentially left with a classification task.

To summarize, if we want to discover *prediction* rules, we will end up either with classification or with dependence modeling – or other machine learning-related task – but not with the standard association task.

At this point we need to say a few words explaining our motivation for writing this paper. We have nothing against association rules. The large number of projects focusing on association rules in the literature is a good evidence of the importance of this data mining task. Actually, we believe that the introduction of the standard association rule framework by Agrawal et al. (1993) was one of the few truly-new proposals for a new data mining task in the last few years. Most of the data mining tasks, such as classification and clustering, have been extensively studied for quite a long time – even though in the past there was not so much emphasis on the issue of scalability.

Our main concern is that, whenever the standard association rule framework is used, its limitations concerning the predictive power of the discovered rules should be well-understood. We note in passing that the classification task, although leading to rules with more predictive power, also has limitations of its own (like any data mining task). In particular, despite what some people believe, classification rules have no causal semantics. They represent correlations in the data, and correlation is not necessarily causation.

Finally, we should emphasize that the discussion presented in this paper refers only to the standard association framework. Clearly, the association task can become as ill-defined and non-deterministic as the classification task when we consider issues such as pruning/summarizing discovered association rules – see Liu et al. (1999b), selecting interesting rules among all discovered association rules – see Dong & Li (1998), Guillaume et al. (1998), Klemettinen et al. (1999), Klemettinen et al. (1994), etc.

4. REFERENCES

Agrawal, R.; Imielinski, T. and Swami, A. (1993) Mining association rules between sets of items in large databases. *Proc. 1993 Int. Conf. on Management of Data (SIGMOD-93),* 207-216. ACM.

Bayardo, R.J. (1997) Brute-force mining of high-confidence classification rules. *Proc. 3rd Int. Conf. on Knowledge Discovery & Data Mining (KDD-97),* 123-126. AAAI Press.

Bramer, M. (1996) Induction of classification rules from examples: a critical review. *Proc. Data Mining'96 Unicom Seminar,* 139-166. London: Unicom, 1996.

Breslow, L.A. and Aha, D.W. (1997) Simplifying decision trees: a survey. *The Knowledge Engineering Review* 12(1)*,* 1997*,* 1-40.

Burges, C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2), 121-167.

Domingos, P. (1998) Occam's two razors: the sharp and the blunt. *Proc. 4th Int. Conf. on Knowledge Discovery & Data Mining (KDD-98),* 37-43. AAAI Press.

Domshlak, C.; Gershkovich, D.; Gudes, E.; Liusternik, N.; Meisels, A.; Rosen, T.; Shimony, S.E. (1998) FlexiMine – A flexible platform for KDD research and application construction. *Proc. 4th Int. Conf. on Knowledge Discovery & Data Mining (KDD-98),* 184-188. AAAI Press.

Dong, G. and Li, J. (1998) Interestingness of discovered association rules in terms of neighborhood-based unexpectedness. *Research and Development in Knowledge Discovery and Data Mining. (Proc. 2nd Pacific-Asia Conf. – PAKDD-98). LNAI 1394,* 72-86. Springer-Verlag.

Guillaume, S.; Guillet, F. and Philippe, J. (1998) Improving the discovery of association rules with intensity of implication. *Principles of Data Mining and Knowledge Discovery (Proc. 3th European Conf. - PKDD-98), LNAI 1510*, 318-327. Springer-Verlag.

Jensen, D. and Cohen, P. (2000) Multiple comparison in induction algorithms. *Machine Learning* 38(3), 1-30.

Jensen, D. and Schmill, M. (1997) Adjusting for multiple comparisons in decision tree pruning. *Proc. 3rd Int. Conf. on Knowledge Discovery & Data Mining (KDD-97),* 195-198. AAAI Press.

Klemettinen, M.; Mannila, H.; Ronkainen, P.; Toivonen, H. and Verkamo, A.I. (1994) Finding interesting rules from large sets of discovered association rules. *Proc. 3rd Int. Conf. Information and Knowledge Management (CIKM-94)*, 401-407. ACM.

Klemettinen, M.; Mannila, H. and Verkamo, A.I. (1999) Association rule selection in a data mining environment. *Principles of Data Mining and Knowledge Discovery (Proc. 4th European Conf. - PKDD-99), LNAI 1704*, 372-377. Springer-Verlag.

King, R.D.; Feng, C. and Sutherland, A. (1995) STATLOG: comparison of classification algorithms on large real-world problems. *Applied Artificial Inteligence* 9(3). May/June 1995, 289-333.

Liu, B.; Hsu, W. and Ma, Y. (1998) Integrating classification and association rule mining. *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD-98),* 80-86. AAAI Press.

Liu, B.; Hsu, W. and Ma, Y. (1999a) Mining association rules with multiple minimum supports. *Proc. 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD-99),* 337-341. ACM.

Liu, B.; Hsu, W. and Ma, Y. (1999b) Pruning and summarizing the discovered associations. *Proc. 5th ACM SIGKDD Int.*

Conf. on Knowledge Discovery and Data Mining (KDD-99), 125-134. ACM.

Michalski, R.W. (1983) A theory and methodology of inductive learning. *Artificial Intelligence* 20, 1983, 111-161.

Michie, D.; Spiegelhalter, D.J. and Taylor, C.C. (1994) *Machine Learning, Neural and Statistical Classification.* New York: Ellis Horwood.

Mitchell, T.M. (1980) The need for biases in learning generalizations. *Rutgers Technical Report,* 1980. Also published in: J.W. Shavlik and T.G. Dietterich (Eds.) *Readings in Machine Learning,* 184-191. Morgan Kaufmann, 1990.

Mitchell, T.M. (1997) *Machine Learning*. McGraw-Hill.

Oates, T. and Jensen, D. (1998) Large datasets lead to overly complex models: an explanation and a solution. *Proc. 4th Int. Conf. on Knowledge Discovery & Data Mining (KDD-98),* 294-298. AAAI Press.

Pinker, S. (1997) *How the Mind Works.* New York: W.W. Norton & Company.

Rao, R.B.; Gordon, D. and Spears, W. (1995) For every generalization action, is there really an equal and opposite reaction? Analysis of the conservation law for generalization performance. *Proc. 12th Int. Conf. on Machine Learning,* 471-479. Morgan Kaufmann.

Schaffer, C. (1993) Overfitting avoidance as bias. *Machine Learning* 10*,* 1993, 153-178.

Schaffer, C. (1994) A conservation law for generalization performance. *Proc. 11th Int. Conf. on Machine Learning,* 259-265. Morgan Kaufmann.

Smyth, P. and Goodman, R.M. (1991) Rule induction using information theory. In: G. Piatetsky-Shapiro and W.J. Frawley. (Eds.) *Knowledge Discovery in Databases,* 159-176. AAAI Press.

About the author:

Alex A. Freitas received his B.Sc. and M.Sc. degrees in Computer Science from FATEC-SP (College of Technology of Sao Paulo) and UFSCar (Federal University of Sao Carlos), both in Brazil, in 1989 and 1993, respectively. He received his Ph.D. degree in Computer Science, doing research in the area of data mining, from the University of Essex, England, 1997. He is currently an associate professor at "Pontificia Universidade Catolica do Parana" (Pontifical Catholic University of Parana), in Curitiba, Brazil. His main research interests are data mining, knowledge discovery and evolutionary algorithms.