Discovery of Classification Rules from Databases

Graeme Richards
University of East Anglia, UK

Beatriz de la Iglesia
University of East Anglia, UK

INTRODUCTION

In descriptive data mining, the objective is to build an *understandable* model that provides insight into the behaviour or characteristics of some data. The data comprise a set of records, each of which assigns values to a set of features or attributes. One of these features is designated the target feature; the value of this feature is described as the class of the record.

For example, given a set of motor insurance records including features describing car and driver details and claims history, we may wish to build a model that will classify drivers as high or low risk with respect to their claims history. This model could then be used when assessing premiums for new customers. We may also wish to understand what characterises low risk drivers so that new marketing campaigns can aim to attract them.

BACKGROUND

Complete classification, such as that usually produced by decision trees, assigns a class to each record in the data. This is often unsuitable for the descriptive data mining task as the models built are often very large and difficult to understand. Also, overall classification accuracy, often used as the guiding criterion to construct the classifier, does not guarantee accurate classification of minority classes (i.e., classes with few representative records, for example high insurance risk).

Partial classification (also known as *nugget discovery*) seeks to find simple and understandable patterns that represent "strong" descriptions of a particular class. It is often convenient to use *rules* to express such patterns (Ali et al., 1999). Rules are of the general form

antecedent* ⇒ *consequent

where the *antecedent* and *consequent* are predicates that are used to define subsets of records from the database and the rule underlines an association between these subsets. In partial classification, the consequent is fixed to be a particular named class. The *strength* of the rule may be expressed by various measures, as described in the following sections.

We are concerned here with the task of partial classification, specifically with the problem of rule discovery. Firstly, we describe the structure of the classification rules used and how rules may be evaluated. We then go on to describe the various techniques that have been developed for the discovery of classification rules. These are:

- Modern Heuristic Methods - The use of optimisation algorithms.
- Multi-Objective Methods – The use of multi-objective evolutionary algorithms.
- All Rules Search – The use of constrained search algorithms.

RULE STRUCTURE

The number of rules that may be constructed is usually very large and often infinite, but imposing constraints on the structure of rules might reduce this. Highly flexible formats allow a rich expression of patterns, which may encapsulate stronger descriptions of a class, but the size of the search space may be very large. Conversely, if the format is too restrictive it will not be possible to express patterns of sufficient interest.

Many rule discovery techniques are restricted to a rule format where the antecedent comprises a conjunction of attribute tests, ATs, and the consequent comprises a single AT representing the class description.

Even with this restriction on rule format, the size of the search space is usually immense for any real-world problem. It is not normally possible to find all rules. Consequently it is necessary to use rule finding techniques that can search effectively within the search space, as described earlier.

EVALUATION OF CLASSIFICATION RULES

Two common measures of rule strength are *confidence* and *coverage,* which are described next.

Given a record, *t*, *antecedent*(*t*) is true if *t* satisfies the predicate, *antecedent.* Similarly, *consequent*(*t*) is true if *t* satisfies the predicate, *consequent.* Then the subsets defined by the *antecedent* or *consequent* are the sets of records for which the relevant predicate is true.

We define three sets of records:

- $A = \{t \in D \mid antecedent(t)\}$, (i.e., the set of records defined by the *antecedent*),
- $B = \{t \in D \mid consequent(t)\}$, (i.e., the set of records defined by the *consequent*),
- $C = \{t \in D \mid antecedent(t) \wedge consequent(t)\}$.

The support for any conjunction of ATs, *M*, *sup*(*M*) is the number of records which satisfy *M*.

Given a rule, *r*, we designate the antecedent of the rule r^a and the consequent r^c.

Then, the support for the antecedent, $sup(r^a)$ = $|A|$ = a

and the support for the consequent, $sup(r^c)$ = $|B|$ = b, (i.e., the cardinality of the target class).

The *support* for *r*, $sup(r)$, is defined as $sup(r^a \wedge r^c)$ = $|C| = c$

The *confidence* of *r*, $conf(r)$, is defined as

$$conf(r) = \frac{sup(r)}{sup(r^a)} = \frac{c}{a}$$

The *coverage* of *r*, $cov(r)$, is defined as

$$cov(r) = \frac{sup(r)}{sup(r^c)} = \frac{c}{b}$$

A strong rule may be defined as one that meets certain confidence and coverage thresholds. Those thresholds are normally set by the user and are based on domain or expert knowledge about the data. Strong rules may be considered interesting if they are found to be novel and useful. That type of criteria, which may be defined subjectively, can only normally be assessed by interpretation of the rule against the domain knowledge, and against the expectations and needs of the data owner, and so forth. In nugget discovery we are therefore interested in presenting a set of strong rules (possibly interesting rules) to the user for further subjective evaluation.

TECHNIQUES FOR THE DISCOVERY OF CLASSIFICATION RULES

Modern Heuristics

Modern heuristic optimisation techniques, namely simulated annealing, genetic algorithms and tabu search, may be used to extract the best classification rules according to a specified measure of interest (de la Iglesia et al., 1996, 2000). In this approach to nugget discovery the problem of finding strong class descriptions becomes an optimisation problem. We represent a conjunctive classification rule as a solution to this problem, and all the classification rules available given a particular rule format constitute the search space. We then evaluate classification rules using some measure of interest so that the search can be guided towards the most interesting rules according to that measure. One such measure is the fitness measure,

$$f(r) = \lambda c - a \text{ where } \lambda \in \Re$$

In this equation *a* and *c* are interpreted as described previously. This measure is capable of partially ordering rules according to confidence and coverage under certain constraints. Under the defined partial ordering, if two rules have the same confidence the rule of higher coverage is preferred, and if two rules have the same coverage the rule of higher confidence is preferred. It follows that if a rule has both higher coverage and confidence than another, then the first rule is preferred. The partial ordering defines a high confidence/coverage boundary from which the heuristic techniques would search for solutions. Variations in the λ parameter allow the algorithms to explore different areas of the upper confidence/coverage boundary, by encouraging the search for rules of high confidence or high coverage.

In the implementation given in de la Iglesia et al. (1996, 2000), a solution or rule is represented as a bit string. Each attribute is assigned a number of bits, with numerical attributes defined by a lower and upper limit and categorical attributes defined by a number of labels. The class label does not need to be represented, as it is fixed. Evaluation is conducted by examining the database to count the support for the antecedent and consequent of