Developments on a Multi-objective Metaheuristic (MOMH) Algorithm for Finding Interesting Sets of Classification Rules

Beatriz de la Iglesia, Alan Reynolds, and Vic J Rayward-Smith

University of East Anglia, Norwich, Norfolk, NR4 7TJ, UK
{bli, ar, vjrs}@cmp.uea.ac.uk

Abstract. In this paper, we experiment with a combination of innovative approaches to rule induction to encourage the production of interesting sets of classification rules. These include multi-objective metaheuristics to induce the rules; measures of rule dissimilarity to encourage the production of dissimilar rules; and rule clustering algorithms to evaluate the results obtained.

Our previous implementation of NSGA-II for rule induction produces a set of cc-optimal rules (coverage-confidence optimal rules). Among the set of rules produced there may be rules that are very similar. We explore the concept of rule similarity and experiment with a number of modifications of the crowding distance to increasing the diversity of the partial classification rules produced by the multi-objective algorithm.

1 Introduction

Data mining is concerned with the extraction of patterns from large databases. One particular task of data mining which is attracting increased research attention is the extraction of classification rules. Partial classification, also known as nugget discovery, involves the production of accurate yet simple rules (nuggets) that describe subsets of interest within a database.

Recently, we have developed a multi-objective metaheuristic algorithm for the extraction of partial classification rules [7]. The problem of nugget discovery was formulated as a multi-objective optimisation problem by using some of the frequently used measures of interest, namely confidence and coverage of a nugget, as objectives to be optimised. NSGA-II was then used to perform the search for Pareto-optimal rules according to the defined objectives.

The approach was evaluated by comparison to another algorithm, ARAC [18] which is guaranteed to find all cc-optimal rules subject to certain constraints. The constraints may affect the number of attribute tests that are allowed in the antecedent of the rule, or the maximum cardinality allowed for any attribute that participates in a test. For small datasets, where constraints do not have to be applied, ARAC can deliver all the cc-optimal (i.e. Pareto-optimal) rules efficiently, so it provides a perfect point of comparison.

C. A. Coello Coello et al. (Eds.): EMO 2005, LNCS 3410, pp. 826–840, 2005.

Results showed the strength of the new multi-objective approach for finding a good approximation to the Pareto front in a number of datasets. For the larger datasets, the multi-objective approach showed real advantage as it could find good sets of solutions in a fraction of the time, with predictable termination times, and without having to apply any restrictions to the number of attributes or their cardinality.

One question raised was whether the set of rules delivered may contain very similar rules or rules that appear to be different but match similar records. There may also be rules that are interesting because they cover different subsets of records, but which are dominated in terms of coverage and confidence and are, therefore, never found.

In this paper we investigate the quality of the rule sets obtained. In particular, we investigate various options for refining the quality of the rule sets obtained in order to deliver an *interesting* set of rules or nuggets. Defining interest in rule induction has been an area of research for some time. Most methods of measuring individual rule interest use a combination of confidence and support for the rule [14]. Considerations about rule novelty or surprise for individual rules are sometimes included [9, 10]. We study the novelty of the rules in relation to other rules within the set; that is, we would like to deliver a set of rules of high quality in terms of confidence and coverage, but also where rules are as diverse as possible with respect to other rules in the same set. This should increase the interest of the rule set, as opposed to the interest of the individual rules. We examine the interest of the rule sets obtained with our previous approach and attempt various modifications to improve our rule sets.

Section 2 covers the basic concepts and terminology used in the paper. Section 3 describes briefly the original multi-objective nugget discovery algorithm. Section 4 describes measures of rule dissimilarity and their applicability to the algorithm and introduces the concept of clustering rules for better interpretation of results. We describe some initial experimentation in section 5 and give our conclusions and ideas for further work in section 6.

2 Concepts and Terminology

2.1 Nugget Discovery

The task of partial classification [1] is also known as nugget discovery; it seeks to find patterns that represent "strong" descriptions of a specified class, even when that class has few representative cases in the data. For example, in insurance data, groups of people that constitute an unacceptably high risk are in a minority. However, if an insurer can identify such groups, with their defining characteristics, they may gain a competitive advantage.

Let Q be a finite set of attributes where each $q \in Q$ has an associated domain, $\mathrm{Dom}(q)$. Then a record specifies values for each attribute in Q. A tabular database, D, is defined to be a finite set of such records. A classification tabular dataset is one in which a class attribute is present.