CCAIIA: Clustering Categorical Attributes into Interesting Association Rules

Brett Gray and M E Orlowska

School of Information Technology, University of Queensland, QLD 4072 Australia,
brett@psych.uq.edu.au, maria@dstc.edu.au

Abstract. We investigate the problem of mining interesting association rules over a pair of categorical attributes at any level of data granularity. We do this by integrating the rule discovery process with a form of clustering. This allows associations between groups of items to be formed where the groping of items is based on maximising the "interestingness" of the associations discovered. Previous work on mining generalised associations assumes either a distance metric on the attribute values or a taxonomy over the items mined. These methods use the metric/taxonomy to limit the space of possible associations that can be found. We develop a measure of the interestingness of association rules based on support and the dependency between the item sets and use this measure to guide the search. We apply the method to a data set and observe the extraction of "interesting" associations. This method could allow interesting and unexpected associations to be discovered as the search space is not being limited by user defined hierarchies.

1 Introduction

Data mining and knowledge discovery in databases has emerged as a new area of research, attracting significant attention from the database and machine learning communities (See [4] for an overview). An interesting sub field is the problem of discovering association rules.

1.1 Association Rules

Initially introduced in [1], the original motivation was to analyse supermarket transactions and observe how often items are purchased together. Given a set of transactions, each representing a set of purchased items, an association rule is a rule of the form $X \Rightarrow Y$ where X and Y are both sets of items. The rule has support s%, and confidence c%: s% of all transactions contain all items from $X \cup Y$ and c% of all transactions containing all items from X also contain all items from Y. The problem as introduced in [1] is to mine all association rules that achieve a minimum support *minsup*, and confidence *minconf* with *minsup* and *minconf* being user specified parameters. This original formulation of the problem has undergone significant research to develop more efficient algorithms for extracting rules [2], [14].

This supermarket transaction formulation can be viewed as discovering associations over a large relational table of boolean attributes. Each tuple corresponds to a transaction and each boolean attribute corresponds to a supermarket item, adopting a value 1 if the item is present in the transaction or 0 otherwise. Further research has extended the problem to include quantitative attributes [13], [7] which have some natural notion of a distance measure (eg. age, income) and categorical attributes [13] containing a range of values with no obvious distance measure (eg. zip code, customer id).

A problem that has been identified with mining association rules is that of granularity [3], [12]. Consider the supermarket transaction case described above. While the association {Milk} $\Rightarrow$ {Bread} may be an association that achieves the minimum support and confidence, if the data is represented at a different level of granularity, this association will not be found. The database may contain items such as brand X skim milk, brand Y full cream milk, brand Z multigrain bread, etc. At this finer level of granularity there may not be any association (eg. {brand X skim milk} $\Rightarrow$ {brand Z multigrain bread}) that achieves minimum support or confidence and the {Milk} $\Rightarrow$ {Bread} association will remain undiscovered.

The problem of granularity has been addressed on quantitative attributes by allowing associations between ranges of values where an attributes value must lie within a range to conform to the association [13]. Work has also been done on mining associations where the left hand side is restricted to two quantitative attributes and the right hand side to a particular value of a boolean attribute [7]. The left hand side then adopts a rectangle or an admissible region (connected, x-monotone region). For a tuple to conform to the association the pair of quantitative attributes must be within the defined region and the boolean attribute must adopt the specified value.

The granularity problem has been addressed on sets of boolean attributes (eg. supermarket transaction data) by assuming a taxonomy (is-a hierarchy) over the attributes [12], [9]. Associations can then be mined between items and nodes of the taxonomy. For example, a taxonomy could be constructed which grouped all forms of milk to a milk node and all forms of bread to a bread node. The {Milk} $\Rightarrow$ {Bread} association could then be mined as a generalised association between the milk and bread nodes of the taxonomy. It is easy to see how this method could be applied to categorical attributes with a taxonomy over the values of the attribute.

The use of a taxonomy to mine generalised association rules has the benefit of providing a method for incorporating additional domain information into the mining process, in the form of the taxonomy. However it is also limiting in that it not only requires the construction of the taxonomy but it also limits the forms of associations that can be discovered. Consider an example of mining generalised associations between a pair of categorical attributes, Customer Id and Item Code. Each tuple represents the identification of a customer and an item that customer has purchased at a particular store. We wish to mine generalised association rules in order to analyse customer profiles. Particular groups of customers may