Theoretical Investigations into Interestingness and Classification of Association Rules in Data Mining

Abstract

Association rules (ARs) bring out hidden relationships among items on the basis of their co-occurrence in a database of transactions. Initially employed in the market-basket context, ARs inform the user about items likely to be purchased in a single transaction, by a customer. This can give an insight into customer purchasing behaviour. Association rule discovery has been plagued by generation of too many insignificant, irrelevant and obvious rules. Many of these rules do not provide any new knowledge to the user (typically a retail-store manager). Thus it is important to mine and present the most interesting, relevant and significant rules to a manager. Approaches that have tried to mitigate this problem include visualization, organization and summarization of rules, grouping of rules and rule pruning.

Interestingness measures try to quantify the amount of interest a rule generates. Interestingness can be classified into objective and subjective categories. Most studies in literature use 'unexpectedness' as the main criterion for the operationalization and quantification of subjective interestingness. However, current studies on subjective interestingness do not consider the underlying dimensions that can result in unexpectedness.

Items can get 'related' in diverse ways. These relationships determine item combinations that may be perceived as 'interesting'. ARs that contain unrelated or weakly related item-pairs may be interesting. Thus, 'item-relatedness' can be used for identifying interesting association rules. In the first part of the study, we use a taxonomy-tree representation for capturing relatedness between item-pairs and subsequent quantification of interestingness.

Any taxonomy includes only 'is-a' type of relationships between items. Consequently, relationships like complementarities may not get captured. In the second part of the study, we include other relationships using occurrence of items in different market baskets (i.e. frequently occurring item-sets). We analyze relationships between items on the basis of

purchase transactions alone. We evolve a framework to represent these relationships using purchased items as the only parameters. This analysis leads us to the development of non-knowledge-based interestingness measures.

Interestingness measures deal with rules in isolation. There is no framework for evaluating groups of rules. A user would find it helpful to get a holistic view, if 'similar' rules are presented together as a group. This brings us to 'clustering' of association rules. In the last part we develop two similarity measures that can be used as bases for clustering of association rules. We provide intuitive justification for these measures. We evolve a distance scheme that utilizes the two similarity measures. We prove the metric properties of this scheme. We demonstrate its applicability by clustering a set of association rules and comparing the clusters obtained by the proposed measures with those obtained by using an appropriate measure from relevant literature.