Chapter 1

Introduction

Association Rule Mining (ARM) is an important data-mining task. It refers to discovery of associations between different sets of items in very large datasets [Agrawal and Srikant, 1994, Brin et al., 1997, Lin and Kedem, 2002, Bayardo, 1998, Hipp et al., 2000, Han et al., 2004]. The discovered rules help in strategic decision making in both commercial and scientific domains. A classical application of ARM is market basket analysis, an application of data mining in retail sales where associations between different items are discovered to analyze customer’s buying habits in order to develop profitable business strategies [Agrawal and Srikant, 1994].

Association Rule mining comprises of two steps i) Generate all frequent itemsets (FI) ii) Generate strong rules using frequent itemsets discovered in the first step [Agrawal and Srikant, 1994, Brin et al., 1997, Lin and Kedem, 2002, Bayardo, 1998, Hipp et al., 2000, Han et al., 2004]. The process of generating the frequent itemsets is computationally expensive and has attracted attention of several researchers. Set of frequent itemsets is often very large [Zaki, 2000].

The problem of generating frequent itemsets is somewhat mitigated by discovery of set of frequent closed itemsets (FCI), which serves as loss less representation of the set of frequent itemsets and is often much smaller in size than that of FI. Thus mining FCIs instead of FIs in association rule discovery saves computation and memory efforts. Earlier works related to discovery of FCIs derived set of FIs and generated association rules from this set [Pasquier et al., 1999c,
Pasquier et al., 1999b, Pasquier et al., 1999a]. Subsequently, non-redundant rules were generated directly from the FCIs [Stumme et al., 2000, Zaki and Hsiao, 2002, Carpineto and Romano, 2005]. In [Zaki, 2000, Zaki, 2004], it was proved that storing closed itemsets in a lattice facilitates generation of non-redundant rules. M.Zaki was probably the first in the data mining community to exploit Formal Concept Analysis (FCA), a field of applied mathematics for discovery of frequent closed itemsets [Zaki and Hsiao, 2002]. concept lattice, a core structure of FCA, stores concepts where concept is defined in terms of extent (set of transactions) and intent (set of attributes). Stumme et al. proved that intent of a concept represents closed itemset [Stumme et al., 2000]. Thus, set of frequent closed itemsets stored in the form of concept lattice facilitate generation of non-redundant association rules. Substantial reduction in the set of generated rules is the main motivation for exploring concept lattice in association rule mining.

High memory and computational requirement of the concept lattice based algorithms prohibits their use in large and incremental datasets. Further, decaying and pruning away older data to capture recent changes in the dataset has not been addressed in these algorithms making them unsuitable for incremental datasets.

Another issue in ARM is the choice of appropriate interestingness measure (IM) for pruning the set of generated rules. Although the set of non-redundant association rules is relatively small as compared to the set generated using frequent itemsets, still the number is too large for useful interpretation. Traditionally, support and confidence have been used as two measures to prune the set of generated rules. However, these two measures may not be appropriate for the application in hand. For example, in an application requiring negative association rules, it may be better to use correlation coefficient as an interestingness measure. Other rule evaluation measures like recall, certainty factor, variance, laplace, and information gain have been proposed in literature which may be useful in different applications of ARM [Geng and Hamilton, 2006, Heravi and Zaïane, 2010]. In fact, as several interestingness measures may be useful in the same application, association rule mining can be considered a multiobjective problem where the rule evaluation
measures constitute the set of different objectives [H. Ishibuchi and Nojima, 2006, Chan et al., 2010, Dehuri et al., 2008, Ghosh and Nath, 2004].

In summary, two issues related to association rule mining algorithms are the production of small set of non-redundant association rules in both static and incremental datasets and generation of interesting rules where interestingness is defined by user-specified objective measures. In our thesis, we attempt to address these two issues. We proposed a faster method of generation of closed itemsets in static datasets. Further, we proposed a memory efficient lattice based data structure and algorithms for discovery of closed itemsets in various window models of data streams. Use of the proposed data structure for computation of various interestingness measures is demonstrated and an efficient algorithm for discovery of multiobjective association rules is proposed. Further, we introduced the notion of ClosedCARs for generation of associative classifiers. An algorithm to generate accurate and comprehensible classifiers is proposed.

1.1 Contributions of the Thesis

The research work proposed here aims to generate non-redundant and multiobjective association rules in both static and incremental datasets using Closed Itemsets. Contributions of the thesis are detailed below:

1. We propose an efficient algorithm, mineCIL, for generation of closed itemsets in static datasets and stores them in a lattice while generating. The algorithm reduces the repeated computation of same closed itemsets and hence is faster than the competing algorithms.

2. We propose an efficient lattice based data structure, CILattice, which stores closed itemsets along with their support. We propose an algorithm for updation of CILattice. The proposed algorithm, CLICI, scans the dataset exactly once and hence is appropriate for incremental datasets. The algorithm can be efficiently used in data stream environment. The proposed data
structure, \textit{CILattice}, provides the flexibility of using the algorithm in sliding as well as damped window model.

3. An algorithm, \textit{MARM\_DS}, for discovery of non-redundant association rules from the \textit{CILattice} data structure is proposed. The algorithm has been adapted to work in the data stream environment. Use of \textit{CILattice} data structure for computation of various interestingness measures is demonstrated and multiobjective non-redundant association rules are discovered using the proposed algorithm.

4. We introduce the notion of class association rules based on closed itemsets (\textit{ClosedCARs}) and develop an algorithm for generation of \textit{ClosedCARs} from a given dataset. An algorithm for removal of rule conflicts is developed. Based on non-conflicting \textit{ClosedCARs}, we propose to use Pittsburgh approach to build an understandable and accurate classifier. Algorithm for prediction of class label of unseen tuple using lattice of \textit{ClosedCARs} is presented. Lattice based classification and multiobjective genetic algorithm techniques are integrated to improve the comprehensibility of the associative classifiers, where comprehensibility is measured by number of rules in the classifier and number of conditions in the antecedent of each rule.

5. All algorithms are backed by extensive experimentation on real life and synthetic datasets.

1.2 Dissertation Outline

Chapter 2 discusses prominent issues in association rule mining. Different approaches for mining of frequent itemsets are discussed. Notion of closed itemsets is described and their advantages are discussed. Various methods for mining of closed itemsets are described. Procedure for mining of non-redundant associating rules using minimal generators is explained. Chapter 3 describes the algorithm, \textit{mineCIL}, for discovery of closed itemsets lattice in static datasets. Chapter 4 de-
cribes a novel compact lattice based data structure, \textit{CILattice}, for storing closed itemsets and describes the algorithm, \textit{CLICI}, for discovery of closed itemsets in various window models of data streams. Chapter 5 describes a single traversal method for discovery of minimal generators and the non-redundant association rules in data streams. Evaluation of various interestingness measures using the proposed data structure is demonstrated and the algorithm for discovery of multi-objective non-redundant association rules is described. Chapter 6 introduces the notion of \textit{ClosedCARs} and describes the procedure for generation of multiobjective associative classifiers.