Chapter 91
Exploration of Soft Computing Approaches in Itemset Mining

Jyothi Pillai
Bhilai Institute of Technology, India

O. P. Vyas
Indian Institute of Information Technology, India

ABSTRACT

Data Mining is largely known to extract knowledge from large databases in an attempt to discover existing trends and newer patterns. While data mining refers to information extraction, soft computing is more inclined to information processing. Using Soft Computing, the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth for achieving tractability, robustness, and low-cost solutions can be revealed. For effective knowledge discovery from large databases, both Soft Computing and Data Mining can be merged. Soft computing techniques are Fuzzy Logic (FL), Neural Network (NN), Genetic Algorithm (GA), Rough Set (RS), etc. For handling different types of uncertainty in huge data, FL and RS are highly suitable. NNs are a nonparametric, robust technique and provide good learning and generalization capabilities in data-rich environments. GAs provide efficient search algorithms for selecting a model, from mixed-media data, based on some priority criterion. In one of its realms, Association Rule Mining (ARM) and Itemset mining have been a focus of research in data mining for a decade, including finding most frequent item sets and corresponding association rules and extracting rare itemsets including temporal and fuzzy concepts in discovered patterns. The objective of this chapter is to explore the usage of Soft Computing approaches in itemset utility mining, both frequent and rare itemsets. In addition, a literature review of applications of soft computing techniques in temporal mining is described.

1. INTRODUCTION

The process of knowledge discovery from data bases (KDD) is a real life problem solving paradigm and is defined as the non-trivial process of identifying valid, novel, potentially useful and understandable patterns from large data bases, where the data is frequently ambiguous, incomplete, noisy, redundant and changes with time. One of the fundamental steps in the KDD process is Data mining which is concerned with the algorithmic means by which patterns or structures are

DOI: 10.4018/978-1-4666-9562-7.ch091

enumerated from large data warehouses. Unlike conventional (hard) computing, the aim of soft computing is to uncover the tolerance for vagueness, uncertainty, partial truth, and approximation to achieve tractability, robustness and solutions with low cost for the upcoming field of conceptual intelligence. Soft computing methodologies consisting of FL, NN, GA, and RS, individually or in integrated manner, is turning out to be strong candidates for performing data mining tasks efficiently (Mitra et al, 2002). The resultant technique of the interactive combinations of data mining and soft computing techniques provides a human-interpretable, low cost solution, as compared to conventional techniques.

ARM is one of the most important areas of data mining research and is used to discover frequent occurrences of items that often appear together (itemsets) and their corresponding association rules. In traditional ARM algorithms, the presence of each item in a transaction is only considered without giving importance to its quantity or weight impact on the transaction. However, real world decision problems do require maximizing the utility of the transacted item making quantity and weight as significant parameters.

To account for this apparent shortcoming, Utility Mining emerges as a specialization to ARM. The main objective of Utility Mining is to identify the itemsets with highest utilities, by considering profit, quantity, cost or other user preferences. These itemsets have utility above a user-specified threshold. In many real-life applications, high-utility itemsets consist of rare items (Pillai et al, 2010). Rare items are items that occur less frequently in a transaction dataset. Frequent itemsets may only contribute a small portion of the overall profit of a business whereas rare itemsets may contribute a large portion of the profit. For example, customers purchase microwave ovens or plasma televisions rarely as compared to bread, washing powder, soap etc but the former may yield more profit for the supermarket than the latter. Rare itemsets provide useful information in different decision-making domains,

This chapter focuses on different issues of data mining and soft computing as well as analysis of these techniques in itemset mining. Section 2 discusses theoretical definitions related to Data Mining, Itemset Utility Mining and Temporal Mining. Section 3 discusses different soft computing techniques and a literature survey of usage of different soft computing methods in data mining, itemset mining and temporal mining. Section 4 highlights the general findings of the exploratory survey of various soft computing tools in itemset mining. Section 5 concludes the chapter.

2. DATA MINING DEFINITION OF TERMS

Data mining is the technique of automatic finding of hidden valuable patterns, relationships and information elicitation from huge volume of data stored in data bases, data warehouses and other data repositories, in order to help make better business decisions, finding sales trends, in developing smarter marketing campaigns and in predicting customer loyalty. Two categories of Data mining tasks are; Descriptive Mining and Predictive Mining. The Descriptive Mining techniques include Clustering, Association Rule Discovery and Sequential Pattern Discovery, are used to find human-interpretable patterns that describe the data in the form of clusters, itemsets, association rules and sequential patterns. The Predictive Mining techniques such as Classification, Regression, Deviation Detection, are used to classify objects or to predict future values of other variables. The most frequently used Data Mining techniques are Classification, Prediction, Clustering and ARM.