CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

This chapter reviews the prior work relevant to the research in data mining.

2.2 DATA MINING

The amount of raw data stored in databases is exploding due to the drop in the price of data storage. In today's fiercely competitive business environment, everyone wants to rapidly turn the stored terabytes of raw data into significant information. Data mining or knowledge discovery is the process of digging out information from the enormous amount of data. Some of the primary data mining tasks are association rule mining, classification, regression, clustering, outlier analysis and summarization.

Data mining tasks can be classified into Predictive and Descriptive data mining. The goal of predictive data mining is to construct a model which can be used to perform classification, prediction, estimation, and other similar tasks. Decision tree, Bayes classification are the best example for predictive data mining. On the other hand, descriptive data mining produces new, nontrivial information based on the available data set. Association rule mining is the best example for descriptive data mining.
2.3 ASSOCIATION RULE MINING

Association Rule Mining (ARM) (Agrawal, R. and Srikant, 1994) is one of the important data mining tasks. ARM is an unsupervised data mining technique, which works on variable length data, and produces clear and understandable association rules. The basic task of association rule mining is to determine the correlations between the items belonging to a transactional database.

In general, every association rule must satisfy two user specified constraints, one is support and the other is confidence.

Support of a rule $X \rightarrow Y$ ($X$ and $Y$ are disjoint items) determines how often a rule is applicable to a given dataset.

$$Support = \frac{\sigma(X \cup Y)}{N}$$  \hspace{1cm} (2.1)

Where $X$ and $Y$ are itemset and $N$ is total number of transaction.

Confidence of a rule determines how frequently items in $Y$ appears in transactions that contain $X$.

$$Confidence = \frac{\sigma(X \cup Y)}{\sigma(X)}$$  \hspace{1cm} (2.2)

The target of association rule mining is to dig out all association rules that satisfy the user specified minimum support and confidence values.
2.4 CLASSIFICATION

Classification is another important technique in data mining. It uses supervised learning technique in which a classifier is constructed for classifying new data instances. The most commonly used classification algorithms are:

- Decision Tree
- Bayesian Networks

2.5 ASSOCIATIVE CLASSIFICATION FRAMEWORK AND METHODOLOGY

Combining different problem solving methods is a very active research area in data mining. Associative classification integrates two well known data mining tasks namely association rule mining and classification. The integration is focused on a special subset of association rules whose right hand side is always restricted to a class attribute.

For example, consider the rule R: X \rightarrow C, C is the class label.

Figure 2.1 illustrates the framework of the associative classification. The system comprises a training phase and a testing phase. In the training phase, associative classification algorithm starts to search for a set of associated, frequently occurring rules. The generated ruleset is pruned by which only those classifiers with good quality are kept for classification. In the testing phase, the system identifies the corresponding class labels for the testing dataset using an associative classifier. Finally, the performance of the whole system is validated to evaluate how accurately the system performs the classification.
Figure 2.1 Framework of Associative Classification
Let D be the database that contains instances and each instance is represented as \(<a_1, a_2 \ldots a_m, C>\), where \(a_1, a_2 \ldots a_m\) are attributes that contains a set of items (i.e. itemset) (I) and \(C\) is the class label. An itemset with \(k\) items is called as a \(k\)-itemset. An Associative Classification is an implication of the form \(a_i \rightarrow C\), where \(a_i\) is attribute value and \(C\) is class label and this rule holds support \(S\%\), where \(S\%\) is the probability that \(a_i\) and \(C\) hold together among all the transactions; and \(C\%\), where it is the conditional probability that \(a_i\) is true under the condition of \(C\).

Support of associative classification rules can be calculated as

\[
Support(X \rightarrow C) = \frac{\sigma(X \cup C)}{N}
\]  \hspace{1cm} (2.3)

Confidence of associative classification rule can be calculated as

\[
Confidence(X \rightarrow C) = \frac{\sigma(X \cup C)}{\sigma(X)}
\]  \hspace{1cm} (2.4)

Where \(X\) is the itemset and \(Y\) is class label.

Merschmann and Plastino (2007) categorized associative classification methods into two types namely:

1. Eager associative classification
2. Lazy learning associative classification.

Eager associative classification methods construct the generalized model (classifier) to classify the data instance whereas, Lazy learning associative classification does not build the generalized classifier to classify an instance.
2.6 EAGER ASSOCIATIVE CLASSIFICATION

Eager Associative classification method works in two stages namely

(i) Rule Generation:

In this phase, a complete set of class association rules are generated from the training dataset. This results in huge number of class association rules.

(ii) Classifier Construction:

In this phase, a classifier is constructed using the subset of class association rules (CAR).

The following tasks are involved in this process:

- Ranking of the Class Association Rules.
- Selecting the subset of CARs to form the classifier.
- Measuring the quality of the classifier on test data objects.

2.6.1 Rule Generation

Eager associative classification methods either adopts Apriori candidate generation (Agrawal and Srikant, 1994), or Frequent Pattern (FP) growth (Han et al, 2000) association rule generation algorithm to generate the class association rules. Figure 2.2 shows the steps involved in the eager associative classification method.
Figure 2.2 Eager Associative Classification

For example, the Class Based Association (CBA) method employs Apriori candidate generation algorithm (Liu et al, 1998). Other associative classification methods such as Classification Based on Multiple Class-Association Rules (CMAR) (Li et al., 2001), Lazy rule pruning algorithm (Baralis and Torino, 2002,

2.6.2 Rule Ranking and Rule Pruning

The rule generation step in associative classification results in large number of rules (Topor and Shen, 2001), (Li et al., 2001). Accuracy of the classifier may be high if all the generated frequent class association rules are used in the classifier but the process of classification will be very slow. Experimental results reported by Baralis et al., (2008) shows that CBA method generates more than 80,000 rules for some datasets that lead to memory exceptions and severe other problems, such as over-fitting.

Therefore the generated rules are ranked based on parameters and interestingness measures such as confidence, support, lexicographical order of items, etc. Then, few top ranked rules are chosen to build the classifier. This has made the classification process more effective and accurate (Thabtah, 2006, Thabtah and Abdeljaber, 2007).

In CBA, initially the rules are ranked based on their confidence value. If two rules have the same value for the confidence measure then the rules are sorted based on their support. If both confident and support values are same for the two rules then the rules are sorted based on their length. Even after considering confidence, support, and cardinality measures, if there exists some rules with the same values for all three measures then the rules are sorted based on its lexicographic order in Lazy pruning (Baralis et al., 2008) method.
To construct a classifier Liu et al., (1998), Li et al., (2001) and Thabtah et al., (2005) used database coverage method. The database coverage chooses top ranked rules and checks it against the training data set. If it covers at least one training data element, then, it is considered for the construction of the classifier. If the rule is unable to predict even a single training instance, then the rule is discarded. The process is repeated until all the sorted rules or training objects are covered.

One of the well known discrete data hypothesis testing is the Chi-square test (Snedecor. W and Cochran 1989). This method evaluates the nature of correlation between two items i.e. it checks whether the two items are positively correlated or negatively correlated. Only positively correlated rules are used in the classifier (Wu et al., 2004).

\[ \chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \]  

(2.5)

where \( e_i \) represents the expected frequencies and \( f_i \) represents the observed frequencies.

CMAR (Li et al., 2001) follows Chi square method to prune the negatively correlated rules. Here, Chi-square test is performed to identify the correlation between the items and the class. Only positively correlated rules are used for classification and discarded otherwise.

CPAR (Yin and Han, 2003) uses Laplace accuracy measure (Clark and Boswell, 1991) to estimate the expected error rate for each rule. Only the rules with less error rates are selected and used for constructing the classifier.
A recent approach is lazy pruning (Baralis, E. and Torino, 2002, Baralis et al., 2004, Baralis et al., 2008) where a rule is pruned only if it misclassifies the data. The entire ruleset is segregated into three sets namely 1. Useful rules, 2. Harmful rules and 3. Spare rules. A rule which classifies at least one data item correctly is said to be a useful rule and these rules will be used by the classifier. A rule that misclassifies a data item is said to be harmful and pruned from the rules pool. Spare rules are not pruned but used when the useful rule is unable to predict an instance. Though lazy pruning strategy works well for small datasets, there exist constraints in memory space and ruleset quality for larger datasets.

Su et al., (2008) proposed entropy based associative classification. This method first generates all the class association rules. Then informative rules are identified using information gain. This formed efficient associative classifier.

2.7 LAZY LEARNING ASSOCIATIVE CLASSIFICATION

Generating, ranking and selecting a small subset of high-quality rules for constructing the classifier is indeed a challenging task. Therefore, Merschmann and Plastino (2006, 2007 and 2010), Veloso et al., (2006), Li et al, (2004) proposed Lazy learning associative classification. This method delayed the processing of data until a new sample needs to be classified. Lazy learning associative classification method eliminates the need for constructing the classifier. Lazy learning consists of the following steps as shown in Figure 2.3.
Figure 2.3 Lazy Learning Associative Classification

2.7.1 Subset Generation

In subset generation stage, testing dataset instances are taken as input and all the combination of subsets are generated.
2.7.2 Subset Evaluation

After generating all the subsets, each subset is evaluated. Merschmann and Plastino (2007) proposed Bayes Theorem based Highest Subset Probability (HiSP) algorithm to classify the protein sequence. Here, motif structures were evaluated based on training dataset and protein was classified.

Recently, Merschmann and Plastino (2010) proposed a lazy approach for general classifier. In their method, HiSP-GC algorithm was used to evaluate the subsets.

2.8 SYSTEM PERFORMANCE VALIDATION

To evaluate how accurately the proposed associative classification assigns class labels, this research work compares the result of the proposed methods based on accuracy (Written and Frank, 2000, Liu et al., 1998, Baralis et al. 2008, Li et al., 2001, Yin and Han, 2003, Antonie and Zaiane, 2004), number of rules generated, number of rules used in the classifier and computation time.

Suppose that the training dataset T and testing dataset T’ comprises S and S’ records respectively. C’ is the original class label in the Testing dataset and C’’ is the predicted class label. Then the accuracy is computed as following:

\[
\text{Accuracy} = \frac{\sum_{i=1}^{S'} V_S}{|S'|}
\]

\[
V_S = \begin{cases} 
1 & \text{if } C' = C'' \\
0 & \text{Otherwise}
\end{cases}
\]

Where \(V_S = \text{No of correctly predicted class}\).
2.9 INFORMATION GAIN (IG)

Information gain is a measure which is used in information theory to quantify the ‘information content’ of messages (Han and Kamber, 2001). In ID3 decision tree algorithm information gain is used to choose the best split attribute (Quinlan, 1986).

Gain based Associative Rule Classification (GARC) (Chen et al., 2006) and Gain based Effective Association Rule classification (GEAR) (Zhang et al., 2011) were proposed based on information attribute which generates compact rule set.

2.10 GENETIC NETWORK PROGRAMMING

Genetic Network Programming (GNP) (Hirasawa et al., 2001, Eguchi et al., 2006, Mabu et al., 2007) is one of the evolutionary optimization techniques, that uses the directed graphs to solve the problems. GNP consists of three nodes: Start node, Judgment node and Processing node. Start node defines the start of the genetic tree. Judgment are decision making nodes and are denoted as J₁, J₂, ……, Jₙ. Processing nodes are results of the decision and are denoted as P₁, P₂, ……, Pₙ.

Shimada et al., (2005, 2006) used this genetic network programming and constructed a new associative classification method. The initial genetic network was constructed randomly. Then the significance of each rule was evaluated using chi-square test.

Gonzales et al., (2009) proved that Lift and Chi-square measures enhances classification accuracy and also showed that the single correlation measure is not suitable for all the datasets.
2.11 WEIGHTED ASSOCIATION RULE MINING

Classical Association Rule Mining framework assumes that all items have the same significance or importance i.e. their weight within a transaction or record is same (weight=1 for every item) which is not always the case. In the supermarket context, some items like jewellery, designer clothes, etc., are much significant in terms of revenue or profit of the store compared with other items. Although classical association rule mining algorithm has reflected the statistical relationship between the items, it does not reflect the semantic relation between the items (Yao et al., 2006).

Hence, it is important to include weight as a parameter to generate association rules (Sun et al, 2008). This method is known as weighted association rule mining.

Table 2.1 Transaction Table

<table>
<thead>
<tr>
<th>TID / Attributes</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2.2 Weighted items table

<table>
<thead>
<tr>
<th>Item</th>
<th>Profit/Unit Sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.8</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
</tr>
<tr>
<td>D</td>
<td>0.6</td>
</tr>
</tbody>
</table>
In the sample database given in Table 2.1, sample association rule is given by A \rightarrow B (support: 60\%) and A \rightarrow D (support 20\%). If the minimum support threshold is 25\% then the important rule A \rightarrow D is not obtained by classical association rule mining but purchase of product A leads to purchase of highly profitable product D. To meet the user objective and business value, weighted association rule mining was introduced, which works on the some preassigned weight. Here, weight is assigned mostly based on the utility of the item such as profit.

Weighted association rules are generated based on user specified minimum weighted support and minimum weighted confidence thresholds. The use of weighted-support and weighted-confidence helps to prioritize the rule according to their importance, instead of their support and confidence alone.

Wang et al., (2000) proposed an algorithm for Weighted Association Rules (WAR). WAR generates association rules according to the weight of individual item, which leads to the invalidation of downward closure property. Ramkumar (1998) addressed this issue by introducing a relaxation factor but it is very difficult to determine the relaxation factor. The problem of invalidation of downward closure property is solved in Weighted Association Rule Mining (WARM). WARM algorithm (Tao et al., 2003) was developed based on the improved model of weighted support and weighted confidence measures.

Some of the datasets such as web site click-stream does not come with preassigned weights. Therefore, Sun et al., (2008), proposed weighted association rule mining without preassigned weight to mine weighted association rules. This algorithm generates w-support and w-confidence from hub weight of transaction by extending Kleinberg’s HITS (1999) model. The itemset is generated using the w-support. The basic idea of w-support is that a frequent itemset may not be as important as it appears, due to the difference in weights of transactions. The weights are derived
from the internal structure of the database based on the assumption that good transactions consist of good items.

If an itemset appears in a very few transactions but in a large quantity, then it is possible that buying of this itemset may lead more profit, but it is not a frequent itemset based on user defined minimum support. This results in a loss of information. Consider the Table 2.1, purchase of item A in two quantity leads to purchase of 10 quantity of item C, but classical nor weighted association rule mining do not consider the above scenario.

To overcome the above problem, Sulaiman Khan et al., (2008) proposed Weighted Utility Association Rule Mining (WUARM). WUARM considered the varied significance and different frequency values of individual items as their weights and utilities. Weighted utility mining focused on identifying itemsets with high weighted utilities. Weighted association rules are generated by considering utility weight of each item.

2.12 CLOSED ITEMSET

"An itemset is closed if none of its immediate supersets has the same support as the itemset . Most of the association rule mining methods are a variant of Apriori algorithm. Apriori algorithm employs breadth-first search to generate huge number of rules (Zaki et al, 2000, 2003, 2005).

Apriori algorithm used the downward closure property to prune the infrequent itemsets. According to this property, the subset of a frequent itemset must be frequent. Thus, Apriori based algorithms extend the itemset level by adding highly frequent items to the frequent itemset and pruning the infrequent itemset. Apriori
based algorithms has shown good performance with sparse datasets and poor performance with dense datasets (Zaki et al., 2005).

To overcome the problem of mining dense dataset, maximal frequent itemset mining was proposed (Agrawal et al., 2000, Bayardo et al., 1998, Burdick et al., 2001, Gouda and Zaki, 2001, Lin and Kedem, 1998). Maximal itemset mining algorithms generated fewer rules than all frequent itemset methods. However, subset frequency is not retrieved by these algorithms which leads to loss of information.

Closed itemset mining has generated minimal number of itemset with its frequency of occurrence even on dense databases. So, closed itemset mining identifies frequent itemset with its exact frequency (Bastide et al, 2000, Pasquier et al., 1999, Pei et al., 2000, Wang et al., 2003, Zaki, 2003).

2.13 MOTIVATIONS FOR THE RESEARCH

Associative classification achieves higher classification accuracy than traditional classifiers like decision tree (Baralis et al., 2008). However, it is a known fact that associative classification typically yields a large number of rules in the training phase. Experimental results reported by Baralis et al., (2008) has shown that Class Based Association (CBA) (Liu et al., 1998) which follows Apriori association rule mining algorithm (Agarwal et al., 1994) generates more than 80,000 rules for some datasets. This large fraction of rules are useless in the classification process and also it leads to memory exception, over-fitting and consumes more time to classify an instance. So, several rule ranking and rule pruning methods are applied for selecting high quality rules in the classifier construction.
Hence, generating high quality rules and constructing the efficient associative classifier without jeopardizing the classification accuracy is of prime importance but indeed a challenging task. This research work aims to improve the performance of associative classification by generating minimal number of efficient class association rules, constructing an optimal associative classifier and also by introducing computationally efficient lazy learning associative classification methods.

The rest of the thesis is organized as follows: Chapter 3 gives brief introduction about Genetic Network Programming and how associative classification rules are extracted in GNP Network. Chapter 4 deals with Associative Classification with Closed Itemset where it explains how to generate closed itemset and closed associative classification rules. Chapter 5 deals with Compact Weighted Associative Classification which includes weight calculation and constructing compact weighted class association classifier. Various Lazy Learning associative classification algorithms are explained in Chapter 6 followed by result and discussion in Chapter 7.

2.14 CONCLUSION

This chapter gave a short walkthrough of the research works carried out in eager associative classification, lazy learning associative classification, genetic network programming, weighted association rule mining and closed itemset mining. These works tends to improve the performance of the associative classification.