CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

The amount of raw data stored in the databases is exploding due to 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 plays an important role in knowledge extraction. It 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 and summarization.

Association rule mining is concerned with extracting a set of highly correlated features shared among a large number of records in a given database. It uses unsupervised learning where no class attribute is involved in finding the association rules. On the other hand, classification uses supervised learning where class attribute is involved in construction of the classifier to classify the new instance. Both, association rule mining and classification are significant and efficient data mining techniques. Associative classification is a recent and rewarding technique that applies the methodology of association rule mining into classification and achieves higher classification accuracy. This research work is mainly concerned with introducing new techniques in associative classification and thereby enhancing accuracy and reducing the computing time.
1.2 PROBLEM DEFINITION

Associative classification achieves high classification accuracy but generating minimal number of high quality rules and constructing the accurate associative classifier is a challenging task. This research work aims to overcome the above challenges by introducing new computational techniques in eager and lazy learning associative classification.

1.3 DATASETS USED FOR EXPERIMENTATION

The proposed systems are tested using benchmark datasets from the University of California at Irvine Repository (UCI Repository) (Blake and Merz 1998). The datasets are preprocessed and converted to a general format. A brief description about the datasets is presented in Table 1.1. The experiments are carried out on a PC with Intel Core 2 Duo CPU with a clock rate of 1.60 Ghz and 2 GB of main memory.

Table 1.1 Dataset Description

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Transactions</th>
<th>Number of Attributes</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balance Scale</td>
<td>625</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>286</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Breast –w</td>
<td>699</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Car</td>
<td>1728</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Credit – a</td>
<td>690</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Diabetes</td>
<td>768</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Ecoli</td>
<td>336</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Flare</td>
<td>1389</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>35</td>
<td>2</td>
</tr>
<tr>
<td>Iris</td>
<td>150</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Mushroom</td>
<td>8124</td>
<td>23</td>
<td>2</td>
</tr>
</tbody>
</table>
To estimate the performance of the classifier random sub sampling approach is used where 90% of the data is randomly chosen from the dataset and used as training dataset and remaining 10% is used as the testing dataset (Han and Kamber, 2001). The training dataset is used to construct a model for classification. After constructing the classifier, the testing dataset is used to estimate the classifier performance. The overall accuracy is obtained by taking the average of the accuracy values obtained from ten different runs.

1.4 PARAMETERS USED TO EVALUATE THE PERFORMANCE

The proposed eager and lazy learning methods are evaluated using the following parameters.

1. Accuracy.
2. Total number of rules generated.
3. Number of rules used in the classifier.
5. Sensitivity and specificity
6. Precision and recall

**Definition 1.1: Accuracy**

Accuracy measures the ability of the classifier to correctly classify an unlabeled data.

\[
\text{Accuracy} = \frac{\text{Number of Correctly Classified Data}}{\text{Total Number of Test data}} \tag{1.1}
\]
It is the ratio of number of correctly classified data over total number of
given test dataset.

Definition 1.2: Total Number of rules generated

Total number of rules generated is defined as the total number of class
association rules generated by the associative classifier with respect to the dataset.

Definition 1.3: Number of rules used in the classifier

Number of rules used by the classifier is defined as the total number of rules
used by associative classifier to classify the new data instances.

Definition 1.4: Computation time

Computation Time is the total time taken to complete the classification
process. Computation time has two parts 1. Training time and 2. Testing time.

\[ \text{Computation Time} = \text{Training Time} \uparrow \text{Testing Time} \quad (1.2) \]

In eager associative classification, training time is the time taken to build the
associative classifier and testing time is the time taken to classify the testing instances.
Lazy learning associative classification classifies the instances without constructing the
class. So there is no training time. Therefore, here computation time is the time taken to
classify the testing instances.
**Definition 1.5: Sensitivity and specificity**

Imbalance in class occurs when one of the classes in a dataset is represented by a small number of samples compared to the other classes (Barandela et al., 2003; Weiss, 2004). The accuracy measure is used extensively to compare the performance of classifiers but may not be suitable for evaluating imbalanced datasets (Tan, 2006). Class imbalance issue might skew the prediction accuracy of classification models (Guha and Schurer, 2008; Hsieh et al., 2008), resulting in a weakened performance of machine learning algorithms (Kang and Cho, 2006). So other metrics like sensitivity, specificity, precision and recall are used.

To use these metrics, the following terminologies are generally used.

i) The **rare class** is denoted as **positive class** and the **majority class** is denoted as **negative class**.

ii) **True Positive** (TP) corresponds to the number of positive examples correctly predicted by the classifier.

iii) **False Negative** (FN) corresponds to the number of positive examples wrongly predicted as negative class by the classifier.

iv) **False Positive** (FP) corresponds to the number of negative examples wrongly predicted as positive class by the classifier.

v) **True Negative** (TN) corresponds to the number of negative examples correctly predicted by the classifier.

\[
Sensitivity (True \ Positive \ Rate) = \frac{TP}{TP+FN} \quad (1.3)
\]

Sensitivity is defined as the fraction of positive examples correctly predicted as positive by the classifier.
Specificity (False Negative Rate) = \( \frac{TN}{TN + FP} \) \hspace{1cm} (1.4)

Specificity is defined as the fraction of negative examples correctly predicted as negative by the classifier.

**Definition 1.6: Precision and Recall**

Sensitivity and specificity measures are used in binary class dataset whereas precision and recall measures are used in multiclass problem.

The definition of precision and recall is given below.

\[
Precision (P) = \frac{TP}{TP + FP} \hspace{1cm} (1.5)
\]

\[
Recall(R) = \frac{TP}{TP + FN} \hspace{1cm} (1.6)
\]

Some combinations of precision and recall are more effective in measuring the classifier’s performance. F1-measure is a measure of a test’s accuracy. It reaches its best value at 1 and the worst score at 0. F1-measure is determined as the harmonic mean of precision (P) and recall (R) as shown in Equation (1.7)

\[
F1 = 2PR / (P + R) \hspace{1cm} (1.7)
\]
1.5 PROPOSED METHODOLOGIES

This research work proposes the following innovative and eminent ideas in eager and lazy learning associative classification methods which improves the performance of associative classification.

- Genetic Network Programming (GNP) is an evolutionary optimization approach in which it constructs a directed graph structure to analyze the decision making mechanism. Random genetic network programming based associative classifier construction method leads to high computation cost. To overcome, this research work applies the methodology of information gain in GNP and formulates a novel associative classification method called genetic network programming with information gain (GNP – IG AC). This method first generates class association rules by selecting the information gain attribute followed by rule evaluation. The proposed GNP-IG method can be categorized into two types depending on the measures used for rule evaluation. They are Chi – square and Lift. After rule evaluation, crossover and mutation operations are performed to generate new population. The rule set thus obtained are optimal and can be used for classification.

- Rule generation phase in associative classification is a hard step that requires a large amount of computation time. Generating all the frequent class association rules leads to high computation cost and memory exception. Therefore, Closed Associative Classification (CAC) is proposed. The CAC extends closed itemset mining to extract class association rules. This forms the informative associative classifier with smaller number of ruleset and thereby reducing the computation cost. To evaluate the efficiency of CAC, two classifier construction methods are
proposed. They are 1. Genetic Ruleset Selection algorithm (GRS) and 2. Eager closed associative classifier. GRS method constructs the associative classifier by randomly selecting the rules from the rule pool and shows improved classification accuracy over traditional associative classifier, Class Based Association (CBA) but with the existing L3 algorithm around 5% drop in classification accuracy is observed. This is due to the presence of more number of conflicting rules in the classifier. To overcome this, eager closed associative classifier construction method is proposed. This method constructs the classifier by eliminating conflict rules thereby improving the classification accuracy.

- Weighted association rule mining reflects the semantic significance of an item by considering its weight. Classification constructs the classifier and classifies the new data instance. Compact Weighted Associative Classification (CWAC) applies weighted association rule mining into classification and constructs an efficient weighted associative classifier. CWAC chooses one non class informative attribute from the dataset and all the weighted class association rules are generated based on that attribute. Therefore, a compact ruleset with efficient rules are generated. The weight of the item which is the most important parameter in generating the weighted class association rules is calculated using the HITS model. After generating weighted class association rules, user defined weighted support and weighted confidence threshold measures are applied to remove infrequent rules. The remaining class association rules are called as frequent weighted class association rules. This frequent weighted class association rules are ranked based on weighted confidence, weighted support, and lexicographical order of items from which compact weighted associative classifier is constructed.
• Eager associative classification has improved classification accuracy but constructing the classifier is a challenging task. Lazy learning associative classification eliminates the need of constructing the classifier. It delays the processing of the data until a new sample needs to be classified. This research work discusses about the three different approaches proposed in lazy learning associative classification methods. They are Highest Subset Confidence (HiSC), Compact Highest Subset Probability (CHiSP) and Compact Highest Subset Confidence (CHiSC). HiSC takes the testing dataset as input and calculates the confidence and support for each combination of class values and class labels are assigned based on the highest subset confidence and support values. It has 1.11% and 0.63% improved accuracy when compared to CBA and HiSP-GC respectively. HiSC improves the classification accuracy but generates more subset of rules that leads to high computation cost. To reduce the computation cost, CHiSP and CHiSC methods are proposed and these two methods use Information Gain attribute to generate the subset. So this method reduces the number of subset generation.

1.6 ORGANIZATION OF THE THESIS

This section gives an overview of the chapters in the thesis. This chapter introduced the associative classification and its challenges, and dataset used for the research.

Chapter 2 discusses relevant prior work and positions the contribution of this research in the context of existing approaches. It surveys the works on various associative classification methods, like genetic network programming based associative classification, weighted association rule mining, closed association rule mining, and lazy learning methods.
Chapter 3 describes the genetic network programming and its implementation for associative classification. Genetic network programming constructs a directed graph type gene structure which helps in directly generating associative classification rules. The performance of the proposed system is tested with chi-square and lift measures.

Chapter 4 proposes closed associative classification method. This method generates the minimum number of high quality closed class association rules. Here the classifier is constructed using genetic algorithm and eager method.

Chapter 5 proposes compact weighted class association rule mining method where weighted association rule mining is applied in classification to obtain an efficient weighted associative classifier. This method chooses one non class informative attribute from the dataset and all the weighted class association rules are generated based on that attribute. The weight of the item is considered as one of the parameter in generating the weighted class association rules and it is calculated using HITS model which do not require any preassigned weights.

Lazy learning associative classification eliminates the need for constructing the classifier. Chapter 6 reviews the existing lazy learning method and proposes three lazy learning associative classification methods.

Chapter 7 compares the various techniques used in this research. It also concludes the thesis with a discussion about the findings of the research and future work directions opened up by this research work.