Abstract— Data mining is the discovery of knowledge and useful information from the large amounts of data stored in databases. It is referred to as knowledge discovery from databases (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

Classification techniques are widely used in data mining to classify data among various classes. Classification techniques are being used in different industry to easily identify the type and group to which a particular tuple belongs. In this paper, different classification techniques are summarized. These techniques are applied on XML data to analyze their advantages and disadvantages. Information coded in XML is easy to read and understand, plus it can be processed easily by computers. It’s open and extendible i.e. there are no fixed set of tags. New tags can be created as they are needed. It contains machine-readable structured information. This is a major advantage over HTML or plain text. XML is self-descriptive and XML documents can be stored without such definitions, because they contain metadata in the form of tags and attributes. It separates content from presentation.

Keywords— Data mining, classification, XML data, Support vector machines.

1. INTRODUCTION

The title Data mining is the process of extracting patterns from data. As more data are gathered, with the amount of data doubling every three years [1]. Data mining is the process of discovering knowledge from large amounts of data stored either in databases or warehouses [2]. Data mining is becoming an increasingly important tool to transform these data into information. Data mining can also be referred as knowledge mining or knowledge discovery from data. Many techniques are used in data mining to extract patterns from large amount of database [1]. Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”. Classification is a twostep process. 1st step is Model Construction [3].

Model Construction: It describes a set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction is called training set. The model is represented as classification rules, decision trees, or mathematical formulae [2].

Model usage: this is the 2nd step in classification. For classifying future or unknown objects, this is used. This model estimates the accuracy of the model. The known label of test sample is compared with the classified result from the model. Test set is independent of training set [2]. For e.g.: if some tuples with certain data is given in the training dataset, in which these tuples are distributed among different classes, then this dataset is used to further determine the class of new tuple arrived for classification. There are many algorithms which are used for classification in data mining shown above. Following are some algorithms [2]:

1) Rule based classifier
2) Decision tree induction [2][1]
3) Nearest neighbour classifier [2]
4) Bayesian classifier
5) Artificial neural network[5][2]
6) Support vector machine[2][6][4]
7) Ensemble classifier
8) Regression trees

2. CLASSIFICATION SCHEMES

2.1 Decision tree induction

Decision tree classification is the learning of decision trees from class labeled training tuples [2]. A decision tree is a flowchart like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. [2].

Advantages: Amongst other data mining methods, decision trees have various advantages [1]. Decision trees are simple to understand and interpret. They require little data and are able to handle both numerical and categorical data. It is possible to validate a model using statistical tests. They are robust in nature,
therefore, they perform well even if its assumptions are somewhat violated by the true model from which the data were generated. Decision trees perform well with large data in a short time. Large amounts of data can be analyzed using personal computers in a time short enough to enable stakeholders to take decisions based on its analysis.

Limitations: The problem of learning an optimal decision tree is known to be NP-complete. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. Decision-tree learners create over-complex trees that do not generalize the data well. This is called over fitting. Mechanisms such as pruning are necessary to avoid this problem.

2.2 Nearest neighbor classifier
The k-nearest neighbor’s algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning. It can also be used for regression. The k-nearest neighbor algorithm is amongst the simplest of all machine-learning algorithms. The space is partitioned into regions by locations and labels of the training samples. A point in the space is assigned to the class c if it is the most frequent class label among the k nearest training samples. Usually Euclidean distance is used as the distance metric; however this will only work with numerical values. In cases such as text classification another metric, such as the overlap metric (or Hamming distance) can be used.

2.3 Artificial neural network
Neural Networks are analytic techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data. Neural Networks is one of the Data Mining techniques. The first step is to design a specific network architecture (that includes a specific number of "layers" each consisting of a certain number of "neurons"). Network is then subjected to the process of "training." In that phase, neurons apply an iterative process to the number of inputs to adjust the weights of the network in order to optimally predict the sample data on which the "training" is performed. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions. The resulting "network" developed in the process of "learning" represents a pattern detected in the data.

2.4 Support vector machines
Support Vector Machines were first introduced to solve the pattern classification and regression problems by Vapnik and his colleagues [8]. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression [2]. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyper-plane in that space, which one which maximizes the margin between the two data sets [2]. To calculate the margin, two parallel hyper-planes are constructed, one on each side of the separating hyper-plane, which are "pushed up against" the two data sets [2]. A good separation is achieved by the hyper-plane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier [2]. This hyper-plane is found by using the support-vectors and margins.

2.4.1 Two class problem
Considering the classes which are linearly separable. Let the data set D be given as (X1, Y1), (X2, Y2)…….. (X|D|, Y|D|), where Xi is a set of training tuples with associated class labels Yi. Each Yi can take only two values, either +1 or -1, corresponding to the classes buys computer=yes and buys computer=no respectively. A separating hyper-plane can be written as

\[ W \cdot X + b = 0 \]  

Where W is a weight vector i.e. \( w=(w_1, w_2, \ldots, w_n) \). Here, n is the number of attributes and b is a scaler.

![Fig. 1 Hyper-planes separating objects and dividing them into two classes](image)
As shown in the above diagrams, there can be number of hyper-planes among objects. These hyper-planes can separate the objects among classes. The hyper-planes with the maximum margin among themselves is considered to the best hyper-plane. Considering two attributes $A_1$ and $A_2$ (as shown in fig 3 and 4), training tuples are considered to be 2D, eg, $X=(x_1, x_2)$ are the values of the attributes $A_1$ and $A_2$. Taking $b$ as an additional weight, $w_0$, equation of separating hyper-plane is

$$w_0 + w_1x_1 + w_2x_2 = 0 \quad (2)$$

Any training tuples that fall on the hyper-planes $H_1$ and $H_2$ are called support vectors.

3. EXPERIMENTS FOR OBTAINING THE PERFORMANCE INFORMATION

A crucial point is how to get the performance information of each algorithm. Here, a data mining tool implemented in java programming language is used [7]. Weka includes tools for data preprocessing, classification, clustering, regression, association rule mining and visualization. It has implemented most of the main stream algorithms. The performance information is obtained by evoking an algorithm included in Weka on benchmark datasets. Case study was done with some datasets. Four classified datasets i.e. CBF, CC, Trace, Gun downloaded from [19] and one time series data ECG from [20] are used in the experiments. Experiments are done with the four commonly used classification algorithms, i.e., decision tree, Support vector machine (SVM), 1-nearest neighbor (1-NN), and Naïve bayes. The classification errors obtained with the developed system are shown in table 1 [21]. The classification execution time is presented in table 2 [21].

Table 1. The classification error for five datasets using four algorithms

<table>
<thead>
<tr>
<th></th>
<th>CBF</th>
<th>CC</th>
<th>Trace</th>
<th>Gun</th>
<th>ECG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree</td>
<td>9.06</td>
<td>7.06</td>
<td>12.00</td>
<td>8.40</td>
<td>65.19</td>
</tr>
<tr>
<td>SVM</td>
<td>1.95</td>
<td>0.92</td>
<td>19.55</td>
<td>7.6</td>
<td>66.46</td>
</tr>
<tr>
<td>1-NN</td>
<td>5.68</td>
<td>2.95</td>
<td>7.6</td>
<td>4.35</td>
<td>55.82</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>14.53</td>
<td>5.56</td>
<td>10.70</td>
<td>27.8</td>
<td>71.52</td>
</tr>
</tbody>
</table>

Table 2. The classification execution time for five datasets using four algorithms

<table>
<thead>
<tr>
<th></th>
<th>CBF</th>
<th>CC</th>
<th>Trace</th>
<th>Gun</th>
<th>ECG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree</td>
<td>5.153</td>
<td>6.427</td>
<td>6.217</td>
<td>2.071</td>
<td>2.731</td>
</tr>
<tr>
<td>SVM</td>
<td>4.579</td>
<td>15.790</td>
<td>6.910</td>
<td>1.815</td>
<td>3.635</td>
</tr>
<tr>
<td>1-NN</td>
<td>2.575</td>
<td>2.914</td>
<td>1.536</td>
<td>0.896</td>
<td>0.287</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.629</td>
<td>0.687</td>
<td>0.765</td>
<td>0.323</td>
<td>0.246</td>
</tr>
</tbody>
</table>

If the execution time is taken into consideration, Naïve Bayes is the best choice. Generally speaking, the 1-NN algorithm has the highest accuracy than the other here algorithms and execution time is shorter than the decision tree and SVM.

4. XML DATA

Information coded in XML is easy to read and understand, plus it can be processed easily by computers [9]. It’s open and extendible i.e. there are no
fixed set of tags. New tags can be created as they are needed. It contains machine-readable structured information [9]. This is a major advantage over HTML or plain text. XML is self-descriptive [10] and XML documents can be stored without such definitions, because they contain metadata in the form of tags and attributes. It separates content from presentation. XML tags describe meaning not presentation [10]. The motto of HTML is: "I know how it looks", whereas the motto of XML is: "I know what it means, and you tell me how it should look." The look and feel of an XML document can be controlled by XSL style sheets. It supports multilingual documents and Unicode [9]. This is important for the internationalization of applications and facilitates the comparison and aggregation of data[4].

The tree structure of XML documents allows documents to be compared and aggregated efficiently element by element. It can embed multiple data types [10] and XML documents can contain any possible data type - from multimedia data (image, sound, and video) to active components (Java applets, ActiveX). Existing data can be embedded [11] and mapping existing data structures like file systems or relational databases to XML is simple. It provides a 'one-server view' for distributed data. XML documents can consist of nested elements that are distributed over multiple remote servers. XML is currently the most sophisticated format for distributed data - the World Wide Web can be seen as one huge XML database [11]. It is being rapidly adopted by industry [4] i.e. IBM, Sun, Microsoft, Netscape, Data Channel, SAP and many others have already announced support for XML.

5. CONCLUSION

In this paper, we discussed the classification techniques in data mining. Various classification techniques are discussed with algorithms. Classification techniques are used to classify data into classes with the help of the training datasets. Discussed above classification techniques can be analyzed on XML data.

6. REFERENCES


[5] Some Techniques and Applications in Data Mining by Tzung-Pei Hong. 2006


[15] “Research of Web mining Technology based on XML” Li Lan. Bibliometric Analysis of Data Mining in the Chinese Social Science Circle Chao Li, Zi.

[16] “Research and Implement of Classification Algorithm on Web Text Mining” Shiqun Yin Gang Wang Yuhui Qiu Weiquan Zhang, Faculty of Computer and Information Science Southwest University 2007 IEEE


