Exploring Data Mining Classification Techniques

Manish Kumar Shrivastava
M.Tech. C.S.E. Scholar
Department of Computer Science & Engineering
Dr. C.V. Raman Institute of Science & Technology
Bilaspur, India

Praveen Chouksey
Assistant Professor
Department of Computer Science & Engineering
Dr. C.V. Raman Institute of Science & Technology
Bilaspur, India

Rohit Miri
Assistant Professor
Department of Computer Science & Engineering
Dr. C.V. Raman Institute of Science & Technology
Bilaspur, India

Abstract

Data Mining is an analytical process of discovering interesting patterns from large amount of data. Data mining performs several tasks one of its major task is classification. Classification maps data into predefined groups or classes that is why it is often referred to as supervised learning. This paper discusses few of data mining classification techniques and algorithm. In this research work three different data mining classification techniques known as ANN, SVM, DT are applied to classify data of three different datasets: the Vote dataset , Breast-cancer(w) dataset and KDD dataset (Intrusion detection) obtained from UCI repository site.

1. Introduction

Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses [1].

Classification in data mining is a form of data analysis that can be used to extract models to describe important data classes or to predict future data trends (Han & Kamber, 2006). The classification process has two phases; the first phase is learning process, the training data will be analyzed by the classification algorithm. The learned model or classifier shall be represented in the form of classification rules. Next, the second phase is classification process where the test data are used to estimate the accuracy of the classification model or classifier. If the accuracy is considered acceptable, the rules can be applied to the classification of new data.

Classification techniques used in this research work described as below.

Multilayer Perceptron

Multilayer Perceptron (MLP) network models are the popular network architectures used in most of the research applications in medicine,engineering, mathematical modeling, etc.. In MLP, the weighted sum of the inputs and bias term are passed to activation level through a transfer function to produce the output, and the units are arranged in a layered feed-forward topology called Feed Forward Neural Network (FFNN). The schematic representation of FFNN with '*n*' inputs, '*m*' hidden units and one output

Unit along with the bias term of the input unit and hidden unit is given in Figure 1. [5]

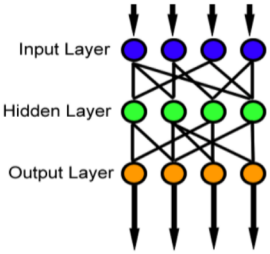


Figure 1. Feed forward neural network.

Decision Trees (DT's)

A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Choice of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached. Decision trees can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules. The J48 decision tree in WEKA is based on the C4.5 decision tree algorithm. The C4.5 algorithm is a part of the multi-way split decision tree. C 4.5 yields a binary split if the selected variable is numerical, but if there are other variables representing the attributes it will result in a categorical split. That is, the node will be split into C nodes where C is the number of categories for that attribute.

Support Vector Machine (SVM)

Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features.

2. Related Work

Many others have worked on different domain to design and develop classification models using data mining techniques.

A. Soltani Sarvestani et al.[2,3] provided a comparison among the capabilities of various neural networks such as Multilayer Perceptron (MLP), Self Organizing Map(SOM), Radial Basis Function (RBF) and Probabilistic Neural Network(PNN) which are used to classify WBC and NHBCD data. The performance of these neural network structures was investigated for breast cancer diagnosis problem.

Dr. Medhat Mohamed Ahmed Abdelaal et al.[2,4] investigated the capability of the classification SVM with Tree Boost and Tree Forest in analyzing the DDSM dataset for the extraction of the mammographic mass features along with age that discriminates true and false cases.

J. Padmavati[5] performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. Logistic regression was performed using logistic regression in SPSS package and MLP and RBF were constructed using MATLAB. It was observed that neural networks took slightly higher time than logistic regression but the sensitivity and specificity of both neural network models had a better predictive power over logistic regression. When comparing RBF and MLP neural network models, it was found that RBF had good predictive capabilities and also time taken by RBF was less than MLP.

Heba Ezzat Ibrahim et al.[6,7] proposed a multi-Layer intrusion detection. There experimental results showed that the proposed multi-layer model using C5 decision tree achieves higher classification rate accuracy, using feature selection by Gain Ratio, and less false alarm rate than MLP and naïve Bayes. Using Gain Ratio enhances the accuracy of U2R and R2L for the three machine learning techniques (C5, MLP and Naïve Bayes) significantly

3. System Implementation

Proposed research work introduces a framework to develop a classifier based on data mining techniques. Another objective is to perform cross validation of different framework designed for different category of data. In this frameworks dataset is given to Pre-processing stage which further classified by selected

classifier. Machine learning tools WEKA are used to analyze the performance of datasets. This approach involve three major steps-

1. Data Pre-processing:
   • Data preparation (load data) e.g.
      -Vote data
      - Breast-cancer data set
      - KDD data set (for intrusion detection)
   • feature reduction (attribute analysis) if needed
2. Data Mining: Classify datasets
   • Select classifier e.g.
      - MLP
      - SVM
      - DT (J48)
3. Data Post-processing:
   • Result Interpretation

System Architecture

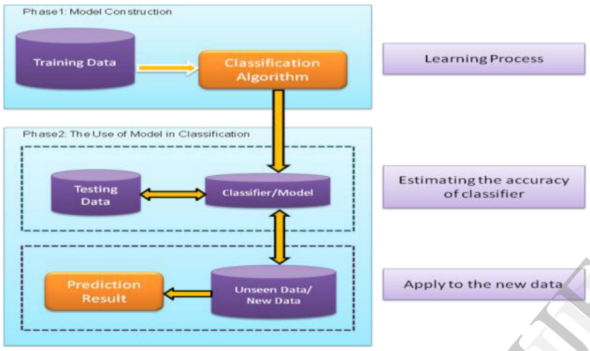


FIG [A]: Classification in Data Mining

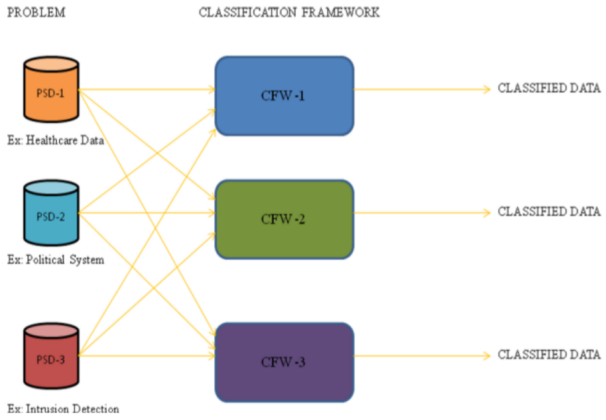


FIG [B]: External Architecture of Cross Validation

4. Experimental Methodology

The experimental methodology followed in this research includes data sets and classification technique. The descriptions of these methodologies are described below.

Data Description

Required data sets for experiment collected from following sources-

- UCI data repository
  - Vote dataset
  - KDD dataset
  - Breast cancer dataset

The data sets used for experimental purpose is downloaded from university of California of Iravin (UCI) repository site (web source http://www.archive.ics.uci.edu/ml/datasets.html). There are three different data sets which belongs to different domains. These datasets are Vote data set which has 435 instances from which 236 belongs to no category while 187 belongs to yes category with 17 features (attribute), Breast Cancer dataset which 699 instances from which 458 benign and 241 malignant with 11 features another data set is KDD data set (for intrusion detection) which has 2519 instances from which 1338 normal and 1181 anomaly with 42 features. The detail of data set is shown in table 1.

Table 1: Datasets Description

Data Set Name	No. of Instances	No. of Class	Name of Classes
Breast-Cancer (Wisconsin)	699	2	Benign, Malignant
Vote	435	2	N(no), Y(yes)
KDD data set	2519	2	Normal Anomaly

Weka as a Data Miner Tool

In this paper we have used WEKA (to find interesting patterns in the selected dataset), a Data Mining tool for classification techniques.. The selected software is able to provide the required data mining functions and methodologies. The suitable data format for WEKA data mining software are MS Excel and ARFF formats respectively. WEKA is developed at the University of Waikato in New Zealand. "WEKA" stands for the Waikato Environment of Knowledge Analysis. The system is written in Java,

An object-oriented programming language that is widely available for all major computer platforms, and WEKA has been tested under Linux, Windows, and Macintosh operating systems. Java allows us to provide a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. WEKA expects the data

to be fed into be in ARFF format (Attribution Relation File Format). [8]

Classification in WEKA

The basic classification is based on supervised algorithms. Algorithms are applicable for the input data. Classification is done to know exactly how the data is being classified. The Classify Tab is also supported which shows the list of machine learning tools. These tools in general operate on a classification algorithm and run it multiple times to manipulating algorithm parameters or input data weight to increase the accuracy of the classifier. Two learning performance evaluators are included with WEKA. The first simply splits a dataset into training and test data, while the second performs cross validation using folds. Evaluation is usually described by the accuracy. The run information is also displayed, for quick inspection of how well a classifier works.

Learning Algorithms

This paper consists of three different supervised machine learning algorithms derived from the WEKA Data mining tool. Which include:

- MLP
- SVM
- J48 (C4.5)

Model Evaluation [9]

Based on data mining techniques as explained above all the developed models are evaluated in terms of following error measures –

Accuracy: Is a percentage of samples that are classified correctly .It is calculated as follows:
Accuracy = (TP + TN) / (P + N)..... (1)

Sensitivity: Is also known as true positive rate (TPR) which can be calculated as follows:

Sensitivity = TP/ (TP+FN)..... (2)

Specificity: Is also known as true negative rate (TNR). It is calculated as follows:
Specificity = TN/ (TN +FP)..... (3)
Where TP, TN, FP and FN are true positive, true negative, false positive and false negative respectively.

5. Results

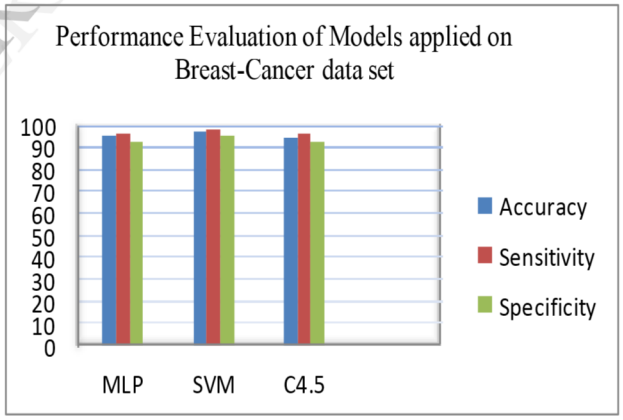
[A] Breast Cancer diagnosis Experiment result:

Table 2 : Confusion Matrix for various predictive models

Predictive Model	Target class	Experiment Result	
		Benign	Malignant
Multilayer Perceptron	Benign	440	18
	Malignant	15	226
Support Vector Machine	Benign	446	12
	Malignant	9	232
Decision Tree	Benign	438	19
	Malignant	17	224

Table 3 : Error measures of various predictive models

Predictive Model	Accuracy	Sensitivity	Specificity
Multilayer Perceptron	95.3	96.7	92.6
Support Vector Machine	**97.0**	**98.0**	**95.1**
Decision Tree	94.6	96.2	92.2

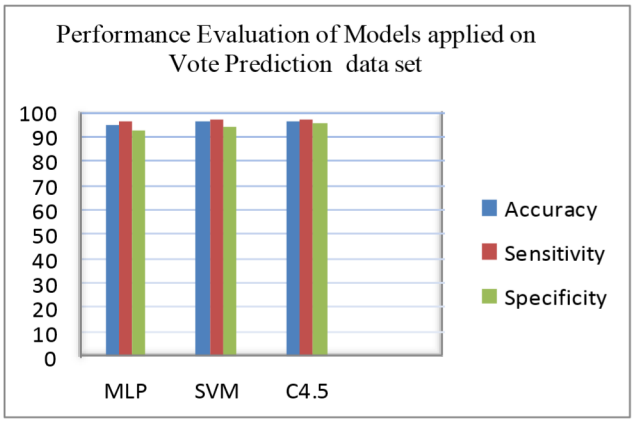


[B] Vote Prediction Experiment result:

Table 4 : Confusion Matrix for various predictive models

Predictive Model	Target class	Experiment Result	
		n	y
Multilayer Perceptron	n	254	13
	y	10	158
Support Vector Machine	n	257	10
	y	7	161
Decision Tree	n	259	8
	y	8	160

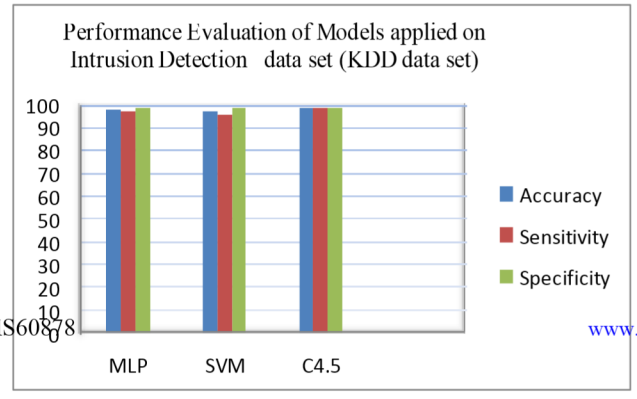
Table 5 : Error measures of various predictive models			
Predictive Model	**Accuracy**	**Sensitivity**	**Specificity**
Multilayer Perceptron	94.7	96.2	92.4
Support Vector Machine	96.1	97.3	94.2
Decision Tree	**96.3**	**97.0**	**95.2**



[C] Intrusion Detection (KDD dataset) Experiment result:

Table 6 : Confusion Matrix for various predictive models			
Predictive Model	**Target class**	**Experiment Result**	
		Normal	Anomaly
Multilayer Perceptron	Normal	1319	19
	Anomaly	38	1143
Support Vector Machine	Normal	1318	20
	Anomaly	58	1123
Decision Tree	Normal	1323	15
	Anomaly	16	1165

Table 7 : Error measures of various predictive models			
Predictive Model	**Accuracy**	**Sensitivity**	**Specificity**
Multilayer Perceptron	97.7	97.2	98.4
Support Vector Machine	96.9	95.8	98.3
Decision Tree	**98.8**	**98.8**	**98.7**



6. Conclusion

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Further predictive models are evaluated as discussed in model evaluation section using equations 1,2 and 3 and calculated results are presented in table 3,5 and 7 in terms of accuracy, sensitivity and specificity. From table 3 it is clear that in breast cancer diagnosis SVM performs well as compared to other two techniques Accuracy whereas from table 5 and 7 it is clear that Decision tree technique (J48) performs well in vote and kdd data set. A comparative Bar Chart showing Error Measures of all classifiers.

7. References

[1] Bharati M. Ramageri / Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305

[2] Shelly Gupta et al./ Indian Journal of Computer Science and Engineering (IJCSE)

[3] Sarvestan Soltani A. , Safavi A. A., Parandeh M. N. and Salehi M., "Predicting Breast Cancer Survivability using data mining techniques," *Software Technology and Engineering (ICSTE), 2nd International Conference*, 2010, vol.2

[4] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed, "Using data mining for assessing diagnosis of breast cnacer," *in Proc. International multiconfrence on computer science and information Technology*, 2010, pp. 11-17. ".

[5] Padmavati J., "A Comparative study on Breast Cancer Prediction Using RBF and MLP," International Journal of Scientific & Engineering Research, vol. 2, Jan. 2011..

[6] Poonam Gupta / International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 5, May - 2013 ISSN: 2278-0181

[7] Heba Ezzat Ibrahim,Sherif M. Badr, Mohamed A. Shaheen," Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems", *International Journal of Computer Applications (0975 – 8887)*,Volume 56– No.7, October 2012.

[8] T. Balasubramanian/European Journal of Scientific Research ISSN 1450-216X Vol.78 No.3 (2012), pp.384-394 © EuroJournals Publishing, Inc. 2012

[9] H.S. Hota/International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-3, January 2013.

Manish Kumar Shrivastava is Persuing M.Tech (CSE) in the department of CSE from Dr. C.V. Raman University , Bilaspur. He recevied his M.C.A. from Gurughasidas University, Bilaspur Chhattisgarh. His interestes area includes Data Mining and Neural Network.

Praveen Chouksey is Currently Assistent Professor in the department of CSE, Dr. C.V. Raman University , Bilaspur. His interestes area includes Data Mining and Neural Network.

Rohit Miri is Currently Assistant Professor in the department of CSE, and Pursuing Ph.D from Dr. C.V. Raman University, Bilaspur, Chhattisgarh, India. He received his M.Tech(CSE) form GEC Pune and BE(CSE) from GEC, Raipur, His interest area includes Application of Soft Computing.