ISSN No. 0976-5697

Volume 3, No. 7, Nov-Dec 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Performance Analysis of Different Classifiers to Build A Classification Model and to Improve The Vigilance Skills in Crime Detection Using Data Mining Techniques

R.Roseline Mary
Assistant Professor, Department of Computer Science
Christ University-Bangalore
roseline.mary@christuniversity.in

Abstract: Data mining is a powerful tool to mine knowledge from large amounts of data. It plays a vital role in the area of which enables fraud detectors who may lack extensive training as data analysts to explore large databases quickly and efficiently. Computer data analysts have started helping the criminal investigators and law enforcement officers to speed up the process of solving crimes by evaluating the crime data and studying the main attributes that lead to further investigation. This paper is an attempt to use the data mining processes particularly to analyse the classifiers and their performance which will help further in enhancing the quality of vigilance department to carry out their task in much faster and efficient way.

Keywords: Data mining, Classification, supervised learning

I. INTRODUCTION

Data mining is the process of extracting the implicit, unknown previously, and valuable knowledge and rules from vast, incomplete, noise, ambiguous and the practical application of random data[1]. Criminology is an area that focuses the scientific study of crime and criminal behaviour and is a process that aims to identify crime characteristics[2]. Data mining techniques have been applied in many application domains such as banking, health care, education, fraud detection, and telecommunications. Recently the data mining methodologies were used to enhance and evaluate the crime detection tasks. With the increasing use of the computerized systems to track crimes, computer data analysts have started helping the law enforcement officers and detectives to speed up the process of solving crimes. Here we will take an interdisciplinary approach between computer science and criminal justice to develop a data mining paradigm that can help solve crimes faster [3].

The data for crime often presents an interesting dilemma while some data is kept confidential, some becomes public information. The information about the sex offenders is made public to warn others in the area, but the identity of the victim is often prevented. Thus as a data miner, the analyst has to deal with all these public versus private data issues so that data mining modeling process does not infringe on these legal boundaries. The police departments use electronic systems for crime reporting that have replaced the traditional paper-based crime reports. These crime reports have the following kinds of information categories namely - type of crime, date/time, gender, location, region suspect's race, suspect's age, modus operandi etc. Then there is information about the suspect (identified or unidentified), victim and the witness. The police officers or detectives use free text to record most of their observations that cannot be included in checkbox kind of pre-determined questions. While the first two categories of information are usually stored in the computer databases as numeric, character or date fields of table, the last one is often stored as free text. The challenge in data mining crime data often comes from the free text field. While free text fields can give the newspaper columnist, a great story line, converting them into data mining attributes is not always an easy job. The classes into which the criminals are classified are as follows offenses against the person, gainful offenses against property with violence, gainful offenses against property without violence, offenses against chastity.

II. THE PROPOSED FRAMEWORK

There are different data mining methodologies like CRISP methodology, CIA Intelligence methodology, Van der Hulst's methodology, AMPA methodology. Each of these methods has been designed specifically for analysis of criminal records. Criminal Investigative analysis is an investigative tool used within the law enforcement community to help solve violent crimes[2].To build a reliable classification model, the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining) is adopted[4]. The methodology consists mainly of five steps: Collecting the relevant features of the problem under study, preparing the data, building the classification model, evaluating the model using one of the evaluation methods, and finally using the model for future prediction. These steps are presented in the next subsections.

A. Collecting the Relevant Features:

In this step the relevant features are collected from the database. Initially attributes have been collected and some of the attributes have been manually eliminated since they are considered as irrelevant to the study. Finally only conditional attributes and one class attribute have been considered.

B. Preparing the Data and Selecting the Relevant Attributes:

For this step, the collected data were prepared in tables in a format that it is suitable for the used data mining system. The data are cleansed by removing the various inconsistent values using the same standard value for all the

data. The cleaning also includes filling out the missing values using the most majority data approach. Since the collected attributes may have some irrelevant attributes that may degrade the performance of the classification model, a feature selection approach is used to select the most appropriate set of features. For this purpose the WEKA toolkit is used and the attributes are ranked and then attributes are eliminated by the feature selection approach.

C. Building the Classification Model:

For the past few years number of works have focussed on the use of data mining techniques and the widespread techniques are classification algorithms and association rules of data mining[5]. The next step is to build the classification model using the decision tree method. Classification finds common properties among officers and detectives to speed up the process of solving crimes by evaluating and studying the main attributes that lead to further investigation into predefined classes which are often referred to as supervised learning technique for knowledge discovery from the crime records and to help increase the predictive accuracy[6]. The decision tree is a very good and practical method since it is relatively fast, and can be easily converted to simple classification rules. For the investigation of crime under study in this paper, the attribute that has the highest gain ratio was the gender to first identify the criminal. This attribute is considered as the root node of the decision tree. The process is repeated for the remaining attributes to build the next level of the tree. After building the complete decision tree, the set of classification rules are generated by following all the paths of the tree where the decision tree has generated classification rules. The classes into which the criminals are classified are offenses against the person, gainful offenses against property with violence, gainful offenses against property without violence, offenses against chastity.

III. DATA SET

To analyse the various classifiers and to find which is suitable to study the crime data, the data set is collected from the UCI Machine Learning Repository. The data set collected are cleansed and all the relevant attributes are considered and the set is ready for building the classification model. Data classification is a two-step process. in the first step, a model is built describing a predetermined set of data classes or concepts. The model is analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class as determined by one of the attributes, called the class label attribute. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population.

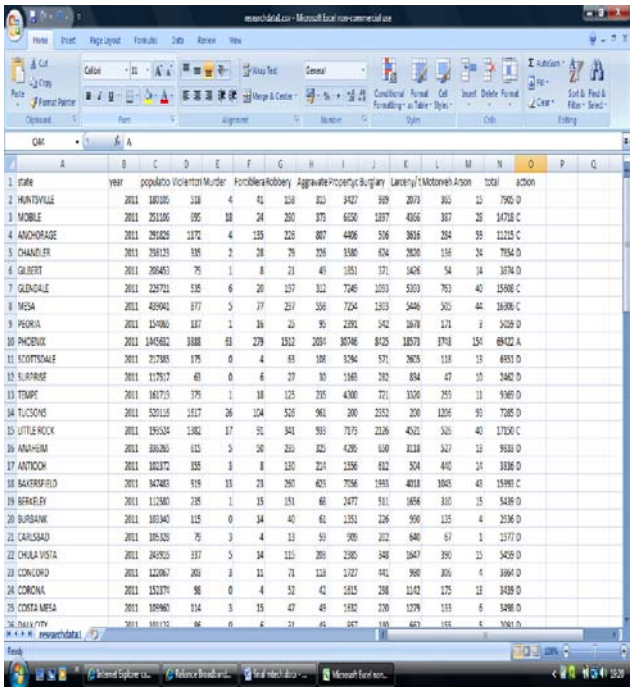

Figure 1 Data Set

IV. WEKA

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state of-the-art facility for developing machine learning (ML)techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules; It also includes visualization tools. The new machine learning schemes can also be developed with this package. WEKA is an open source software issued under General Public License [7].

V. DIFFERENT CLASSIFIERS USED

A. DNTB Classifier:

This DNTB classifier is for building and using adecision table/naive bayes hybrid classifier. The algorithm checks the merit of dividing the attributes into two disjoint subsets: one for the decision table, the other for naive Bayes. At each step the algorithm considers dropping an attribute entirely from the model.

B. PART Classifier:

This is a class for generating a PART decision list. It uses divide and conquer technique and builds a partial C4.5 decision tree in each iteration and makes the best leaf into a rule.

C. ZEROR Classifier:

ZEROR classifier is used to test the results of the other learners and chooses the most common category all the time.

D. ONER Classifier:

One learns a one-level decision tree, i.e. generates a set of rules that test one particular attribute. OneR algorithm

creates a single rule for each attribute of training data and picks up the rule with the least error rate.

E. J48:

The J48 decision tree classifier first creates a decision tree based on the attribute values of the available training data. Whenever it comes across a set of items it identifies the attribute that and tells us about the data instances so that we can classify them and the best is said to have the highest information gain.

F. MultiClass:

Many classifiers in weka handle multi-class problems directly or transform multi-class problems into multiple binary class ones (in various ways) internally. There is also a MultiClassClassifier meta classifier that does this for any binary class classifier.

VI. PERFORMANCE ANALYSIS

The performance of classification algorithms is usually examined by evaluating the accuracy of the classification. Classification accuracy is usually calculated by determining the percentages of tuples in the correct class. This ignores the fact that there also may be a cost associated with an incorrect assignment to the wrong class. This perhaps should also be determined. Given a specific class, Cj and a database tuple ,ti that tuple may or may not be an assigned to that class while its actual membership may or may not be in that class. This assigns the four quadrants which can be described in the following way.

True positive (TP): ti predicted to be in Cj and is actually in it.

False positive (FP): ti predicted to be in Cj but it is not actually in it.

True negative (TN): ti not predicted to be in Cj and it is not actually in it.

False negative (FN): ti not predicted to be in Cj but it is actually in it.

An OC (operating characteristics) curve or ROC (receiver operating characteristic) curve shows the relationship between false positives and true positives. In the OC curve the horizontal axis has the percentage of false positives and the vertical axis has the percentage of true positives for a database sample.

A. Confusion Matrix:

A confusion matrix contains information about actual and predicted classifications done by a classification system[8]. Given m classes, a confusion matrix is an mxm matrix where entry Ci, j indicates the number of tuples from D that were assigned to class Cj but where the correct class is Ci. Obviously the best solutions will have only zero values outside the diagonal.

VII. EXPERIMENTS AND EVALUATION

In order to measure the performance of a classification model on the test set, the classification accuracy or error rate are usually used for this purpose. The classification accuracy is computed from the test set where it can also be used to compare the relative performance of different classifiers on the same domain. However, in order to do so, the class labels of the test records must be known. Moreover an evaluation methodology is needed to evaluate the classification model and compute the classification accuracy. Mainly there are two methods for the evaluation named: The Holdout method and the K-Cross-Validation method. To obtain the accuracy of the classification model the WEKA toolkit is used.

Table I Results of Different Classifiers

Weka ML Classifier	Mean Abs Error	Mean Sqr error	Correctly classified instances	Incorrectly classified instances	Time taken
Decision Table	0.0625	0.0133	96.202	3.7975	0.05
DNTB	0.0394	0.1238	94.9367	5.0633	0.76
ZEROR	0.1845	0.2966	79.7468	20.253	0
ONER	0.1076	0.328	78.481	21.519	0
Simple Cart	0.0181	0.0951	97.4684	2.5316	0.29
J48	0.0485	0.2163	89.8734	10.126	0
Multiclass	0.3	0.3464	100	0	0.36

VIII. CONCLUSION AND FURTHUR WORK

The experimental study is done on the crime dataset. The dataset was trained and tested using the above classifiers. This experiment implies a very commonly used indicator which are mean of absolute errors and root mean squared errors[9]. In WEKA Decision Table classifier out of 79 records and 15 attributes the correctly classified instances were 96% with 0.05 seconds. DNTB classifier correctly classified instances were only 94% in 0.76 seconds. Although it makes little sense to use this scheme for prediction, it can be useful for determining a baseline performance as a benchmark for other learning schemes. ZEROR and ONER classifiers correctly classified instances were only 79% and 78% with 0 and 0.29 seconds. Simple cart classifier correctly classified instances were 97% with 0.29 seconds.J48 pruning tree classifier's correctly classified instances were 89.8734 with 0 seconds and Multiclass classifier's correctly classified instances were 100% with 0.36 seconds .Multiclass classifier shows true positive rate value as 1 and precision value 1. From the above table Multiclass classifier performance is better than the other methods and the accuracy rate is very high for Multiclass classifier. In order to be able to detect newer and unknown patterns in future, clustering techniques work better. As further work, unsolved crimes can be clustered based on the significant attributes and the result is given to detectives for inspection. Many future directions can be explored in this still young field. Visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern[10].

IX. REFERENCES

[1]. Jiang Xianhai, Xie Cunxi, "Home Health Monitoring System Based on Data Mining,"International Forum on Information Technology and Applications IEEE 2009

[2]. K.Zakir Hussain, .Durairaj,G.Rabia Jahani Farzana, "Criminal Behaviour Analysis by using data mining techniques," International Conference on Advances in Engineering, Science and Management(ICAESM-2012)March 30,31,2012

[3]. Shyam Varan Nath, Crime Pattern Detection Using Data Mining, Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology , 41-44, 2006

[4]. Chapman P, Clinton J, Kerber R, KhabazaT, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0: Step-by-step data mining guide, 2000.

[5]. Cesar Vialardi,Javier Bravo,Leila Shafti,Alvaro Ortigosa,"Recommendation in Higher Education using Data Mining Techniques," Educational DataMining ,2009

[6]. Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime Data Mining: An Overview and Case Studies", AI Lab, University of Arizona, proceedings National Conference on Digital Government Research,2003,

[7]. WEKA at http://www.cs.waikato.ac.nz/~ml/weka

[8]. Anshul Goyal and Rajni Mehta," Performance Comparison of Naïve Bayes and J48 Classification Algorithms", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012)

[9]. Mohd Fauzi bin Othman,Thomas Moh Shan Yau," Comparison of Different Classification Techniques Using WEKA for Breast Cancer" Biomed 06, IFMBE Proceedings 15, pp. 520-523, 2007

[10]. Malathi,Dr.S.Santhosh Babu,"An Enhanced Algorithm to predict a future crime using Data mining","International Journal of Computer Applications(0975-8887) Vol 21-No.1,May 2011