Appling Data Mining Technique for Crime Prevention: The Case of Hossaena Town Police Office

Fantaye Ayele
School of Informatics, Wolaita Sodo University, PO box 138, Wolaita Sodo, Ethiopia

Abstract
The Law enforcement agencies like that of police today are faced with large volume of data that must be processed and transformed into useful information and hence data mining can greatly improve crime analysis and aid in reducing and preventing crime. The purpose of this study is to construct predictive models that could help in the effort of crime pattern analysis with the aim of supporting the crime prevention activities at the Hossaena town police office. For this study, a six-step hybrid knowledge discovery process model is followed, due to the nature of the problem and attributes in the dataset. The classification technique such as J48 decision tree and Naive Bayes used to build the models. Performance of the models is compared using accuracy, True Positive Rate, False Positives Rate, and the area under the Relative Optical character curve. J48 decision tree registers better performance with 96.34% accuracy. Lastly for extracting the knowledge the researcher develop the prototype for the user for support the decision which crime is assigned under the serious, medium or low for this purpose the researcher generate the prototypes.

Keywords: Classification, Crime, Data Mining, Hybrid, WEKA
DOI: 10.7176/CEIS/11-1-03
Publication date: January 31st 2020

1. Introduction

Data Mining or Knowledge Discovery in Databases (KDD) in simple words is nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Jiawei. H and Micheline. K, 2006). It deals with the discovery of hidden knowledge, unexpected patterns and new rules from large databases. According to Berry and Linoff (1997) Data mining is one such tool that has evolved to play a role as an instrument to discovery patterns buried in large databases. Data mining is the "exploration and analysis of large quantities of data in order to discover meaningful patterns and rules". Along with the prevention and investigation of crime, police makes use of previous crime reports and data as an input for the formulation of crime prevention policies and strategic plans (Wilson,1963). It is obvious from the outset that to make use of data and records, relevant data have to be kept and managed properly. For this reason the Hossaena town police office have been collecting criminal records since its establishment and have maintained numerous criminal records consisting of fingerprints, names, photographs, and general descriptions of criminals. Human beings want to live and labor in a place where they are safe. They want to ensure that there is a worried body that defends their lives as well as their belongings from possible risks. According to Wilson (1963) In fact, one of the core purposes of any government is to ensure that law and instruction for the security of its citizens are put in place. In other words, faraway from creation and performing of laws for the avoidance of crime, governments must start agencies and establishments, which apply these laws.

2. Statement of the problem

Crime is a complex social phenomenon and its cost is increasing due to a number of societal changes and the like, and hence, law enforcement organizations like that of police need to learn the factors that constitute higher crime trends (Wilson, 1963). To curb this social evil there is always a need for prudent crime prevention strategies and policies. Understanding and processing of criminal records is one method to learn about both crime and individuals who involve in misdeeds so that police can take crime prevention measures accordingly (Brown, 2003). However, in the case of Hossaena town police office there are no modern tools and techniques that can support in managing crime records properly and efficiently. As a result almost all the decision-making processes of the office are not supported by tools and techniques that could extract patterns from previous crime records. Consequently, training programs, resource deployment, crime prevention and investigation strategies are being pursued on the basis of crime incidents rather than crime patterns and trends. Thus, one can observe the cost of those entire activities that do not rely on sound justifications. Therefore, this study is launched to identify appropriate tools as well as to develop models that could extract crime patterns from the criminal database which supports the decision making process of crime prevention.

3. Objective of the study

3.1. General objectives

The general objective of the study is to construct predictive models that could help in the effort of crime pattern analysis with the aim of supporting the crime prevention activities at the Hossaena town police office.

3.2. Specific objectives

To accomplish the above stated general objective, the following specific objectives are developed:

- ✓ To understand the area via an extensive literature review.
- ✓ To find out classification algorithm that more suitable to build predictive model for crime prevention.
- ✓ To evaluate the performances of the model.
- ✓ To develop a prototype

4. Research Methodology

This paper was used a Hybrid data mining model which is a six step knowledge discovery process model. Due to the nature of the problem and attributes in the dataset, classification mining task were selected to build the predictive model. Hybrid data mining methodology basically follows an iterative process consists of: Business understanding, Data understanding, Data preparation, model building, evaluation and use discover knowledge. Fig. 1

5. Experimentation and Analysis of Result

In this study an attempt is made to explore crime data to identify regular patterns in order to determine crime level. The purpose of experiments in classification is to find model that is able to predict the cream level of crime as low, medium and serious by taking selected variables as inputs. This paper incorporated the typical stages that characterize a data mining process.

5.1. Experimental Design

In this study, all experiments are done based on the final processed dataset which contains 5,000 instances and 12 attributes. The algorithms used during predictive model building experimentations are found in Weka 3.8 version. This version works on many file formats than its antecedents and it is compatible with CSV file format. Thus, changes the dataset from excel to ".arff" file format which is necessary in the previous versions. The prepared dataset is saved using CSV file extension format. Fig. 2

5.2. Modeling Building

Model building is one of the major tasks which are undertaken under the phase of data mining in Hybrid data mining methodology. To build the predictive model, J48 and naive Bayes are trained and evaluated. For training and testing the classification model the researcher used two methods. The first method is percentage split method, where 75% of the data used as training and the remaining 25% testing. The second method is K-fold cross validation methods the data was divided into 10 folds, some fold is used as testing and the remaining folds are used as training. When we compare the result of experiment 1 with all attributes and 5, 000 instances are used.

Experimentation I: J48 decision tree

In the experiment I; the first scenario #2 of this experiment, 10-fold cross validation registered the best performance with 96.34% accuracy and 0.963 % TP rate. The Accuracy and WTPR all models indicate the performance of the model in accurately classifying new instances in classes of crime level and it is calculated to be: 96% with misclassification of 4%. In the second scenario #1; 75/25 percentage split registered the performance with 96.24% of correct classification with error rate of 3.4% and 0.962 for WTPR. Tab. 1

Finally, 10–fold cross validation test option the J48 learning algorithm is the best model scored an accuracy of 96.34%. This result shows that out of the total training datasets 4817 (96.34%) records are correctly classified, while only 183 (3.66%) of the records are incorrectly classified.

Experimentation II: Naive Bayes

In the experiment II; the first scenario #1 of this experiment, 10-fold cross validation registered the best performance with 83.6% accuracy and 0.836 % TP rate. The Accuracy and WTPR all models indicate the performance of the model in accurately classifying new instances in classes of crime level and it is calculated to be: 83% with misclassification of 17%. In the second scenario #2; 75/25 percentage split registered the performance with 82.34% of correct classification with error rate of 17.7% and 0.823 for WTPR. Tab. 2

Finally, 75/25 percentage split test option the Naive Bayes learning algorithm is the best model scored an accuracy of 83.6%. This result shows that out of the total training datasets 1045 (83.6%) records are correctly classified, while only 205 (16.4%) of the records are incorrectly classified.

5.3. Model Comparison

In this research work, several experiments had been carried out with two classification algorithms, i.e. J48 decision tree algorithm and Naive Bayes classifier to build a predictive model that predicts the Cream Level in crime Dataset. From the experiments all attributes were identified to make sound rule and better accuracy. Selecting a better classification technique for building a model, which performs best in handling the prediction and identifying significant attribute of crime level of crime prevention is one of the aims of this study. Finally, the accuracy achieved on selected feature was 83.6%, 96.34% for Naïve Bayes and J48, respectively

5.4. Evaluation of the discovered knowledge

At this stage in the data mining task a model was built to have high quality from a data analysis perspective. Besides, it is important to thoroughly evaluate the model and review the steps executed to construct the model and to be certain that it achieves the business objectives. At the end of this phase, a decision on the use of the data mining results is reached. This is performed based on the domain expert's advice and the parameters set and the researcher's personal judgment. It is good to see the meaning of the patterns generated by decision tree.

Generating Rules from Decision Tree

The model developed by J48 classifier was selected as the best model for this study. The generated rules were evaluated by the domain expert. The domain expert agreed on the relevance of the rules, but suggested that further analysis should be performed. The domain expert selected 20 rules that used to develop prototype.

5.5. Prototype development

The final objective of this study was developing a prototype interface that assists physician easy access to the identified knowledgebase. The final selected if-then rules are used to implement the selected best models. Therefore, only twelve rules which are suggested to be important by domain experts are placed in to this prototype which means all the rules for predicting tumor states of a patient can't be answered by this prototype.

6. Conclusion

The purpose of this study was to explore the applicability of data mining techniques in the process of crime prevention for the Hossaena Police office. Hybrid data mining methodology basically follows an iterative process consists of: Business understanding, Data understanding, Data preparation, model building, evaluation and use discover knowledge. The models were built on the preprocessed crime prevention dataset with two different supervised machine learning algorithms i.e. J48 Classifier and Naïve Bayes using Weka 3.8 machine learning software. Although both techniques have shown promising results, the decision tree data mining technique was found more appropriate to the crime prevention problem as the accuracy rate was relatively higher in both experiments. Moreover, decision tree seems applicable due to the fact that in contrast to neural networks, it expresses the rules explicitly. These rules can be expressed in human language so that anyone can easily understand how and why a classification of instances is made. The most effective model to predict prevention of crime with crime level appears to be a J48 classifier implemented on 10-Fold Cross Validation with a classification accuracy of 96.34% and still much remains to fill the gap of 3.66% misclassified cases. This means the selected model can also predict crime level correctly medium as medium or vice versa wrongly with a rate of 3.66%.This has its own implication in reality. Misclassifying medium as low/Serious means leaving infected person to transmit the disease where as that of low/Serious as medium is adding tension to crime. Finally, prototype interface develop are developed and the performance of the system.

Reference

Jiawei Han and Micheline Kamber, (2006), Data Mining: concepts and Techniques 2nd ed , Morgan Kaufmann

Berry, M. and Linoff, G. (1997). Data mining techniques: For marketing, sales and customer support. New York. John Wiley and Sons, Inc.

Wilson , O.W. (1963). Police Administration. USA, McGraw Hill Company.

Brown, D. (2003). The Regional Crime Analysis Program (RECAP): A Framework for Mining

Cios, K. and Kurgan, L., (2005), Trends in Data Mining and Knowledge Discovery in Advanced Techniques in Knowledge Discovery and Data Mining, London: Springer, Pp. 1–26.ata to Catch Criminals.

Figures and Tables

Figure 1: Hybrid-DM Process Model [cios, kand kurgan, L, (2005)]

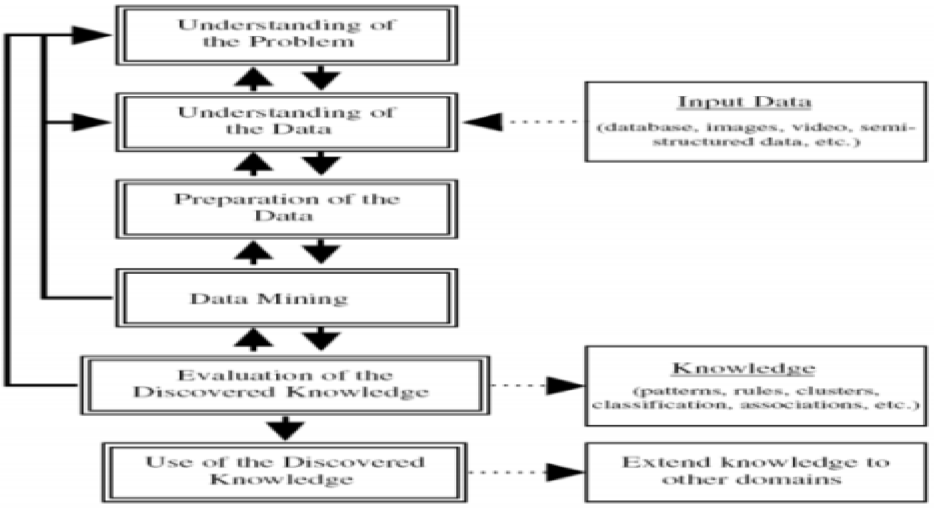
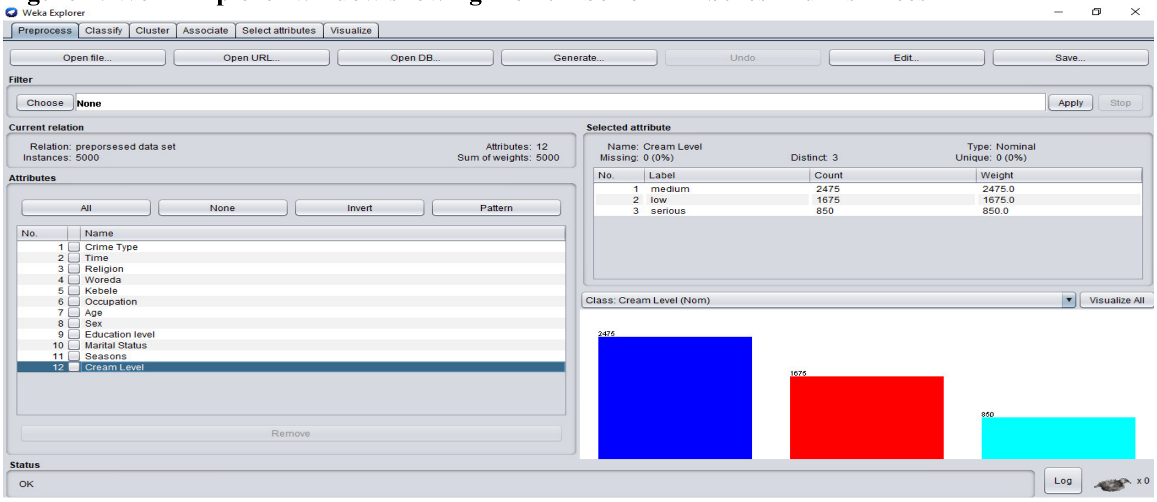

Understanding of the Problem
Understanding of the Data
Input Data
(database, images, video, semi-structured data, etc.)
Preparation of the Data
Data Mining
Evaluation of the Discovered Knowledge
Knowledge
(patterns, rules, clusters, classification, associations, etc.)
Use of the Discovered Knowledge
Extend knowledge to other domains

Figure 2: Weka Explorer window showing the number of attributes and instances



Filter

Choose None

Current relation

Relation: preporsesed data set
Instances: 5000
Attributes: 12
Sum of weights: 5000

Attributes

No.	Name
1	Crime Type
2	Time
3	Religion
4	Woreda
5	Kebele
6	Occupation
7	Age
8	Sex
9	Education level
10	Marital Status
11	Seasons
12	Cream Level

Selected attribute

Name: Cream Level
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	medium	2475	2475.0
2	low	1675	1675.0
3	serious	850	850.0

Class: Cream Level (Nom)

Status

OK

Figure 3: prototypes graphic user interface (GUI)

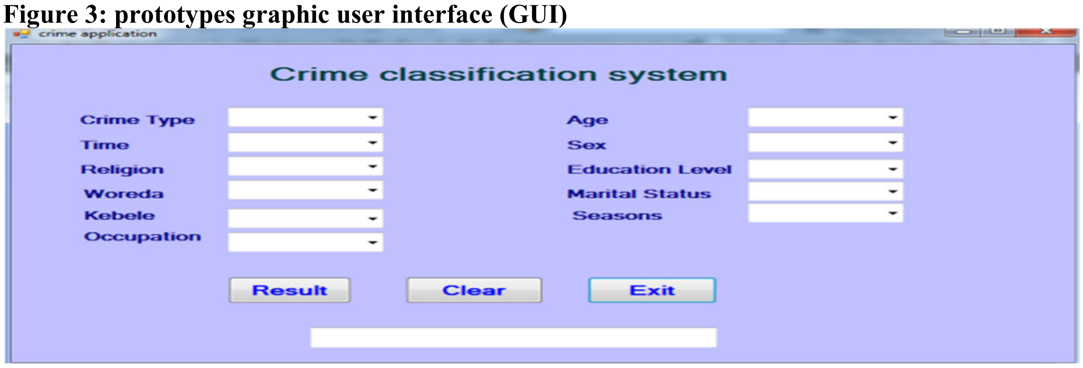


Table.1. Summary of Experiments I with J48 decision tree

Exp I **(test model)**	**Accuracy**	**Time Taken**	**Tree Size**	**Leaf Size**	**W TPR**	**W FPR**	**W PR**	**W RR**	**W ROC**	**CCI**	**ICI**
J48 75/25 percentage split	96.24	0.01	59	46	0.962	0.034	0.964	0.961	0.996	1203	47
J48 10-fold cross validation	**96.34**	0.03	59	46	**0.963**	**0.030**	**0.964**	**0.963**	**0.995**	4817	183

Table.2. Summary of Experiments I with Naive Bayes

Experiment II	**Accuracy**	**Time Taken**	**Av TPR**	**Av FPR**	**Av PR**	**Av RR**	**Av ROC**	**CCI**	**ICI**
Naive Bayes **75/25 percentage split**	**83.6**	0	**0.836**	**0.119**	**0.837**	**0.836**	**0.920**	1045	205
Naive Bayes 10-fold cross validation	82.32	0.01	0.823	0.129	0.824	0.823	0.911	4116	884

Table 3: The selected Models Comparison

Selected Model	**Accuracy**	**Time Taken**	**Av TPR**	**Av FPR**	**Av PR**	**Av RR**	**Av ROC**	**CCI**	**ICI**
J48 10-fold cross validation	**96.34**	0.03	**0.963**	**0.030**	**0.964**	**0.963**	**0.995**	4817	183
Naive Bayes 75/25 percentage split	83.6	0	0.836	0.119	0.837	0.836	0.920	1045	205