A Study on Classification Algorithms for Crime Records

Conference Paper · August 2016

DOI: 10.1007/978-981-10-3433-6_104

CITATIONS
3

READS
1,766

2 authors:

Sundharakumar Kb
Sri Sivasubramaniya Nadar College of Engineering
11 PUBLICATIONS **106** CITATIONS
SEE PROFILE

Bhalaji Natarajan
Sri Sivasubramaniya Nadar College of Engineering
70 PUBLICATIONS **650** CITATIONS
SEE PROFILE

A Study on Classification Algorithms for Crime Records

K.B. Sundhara Kumar and N. Bhalaji(✉)

SSN College of Engineering, Chennai, India
{sundharakumarkb, bhalajin}@ssn.edu.in

Abstract. Data mining has its popularity among crime data analysis significantly due to increasing crime rates across the globe. In this research, classification methods are applied for predicting the nature of a crime that is whether the crime is a violent crime or a non-violent crime. In this work, we present two classification algorithms – Gradient Boosting algorithm and Random Forest algorithm for predicting the crime as a violent or non-violent crime and analyze the accuracy, precision and recall values of these algorithms for the crime records. The dataset is taken from the Communities and Crime data from UCI repository for processing. Further, to improve the accuracy of the predicted results, we use Boruta algorithm which is primarily a wrapper-algorithm for all relevant feature selections. The study finds that Boruta algorithm performs better in feature selection than the Chi-Square feature selection algorithm.

Keywords: Classification · Crime data mining · Gradient boosting · Random forest · Chi-square test · Boruta

1 Introduction

The goal of any country is to provide security to its citizens. Reducing crime in the nation is one solution to overcome the security concerns in the country. In order to predict and prevent a crime, it is very critical to understand the underlying patterns of a crime. There has been a significant rise in crime pattern mining, predictive policing, hotspot analysis and spatial and temporal analysis of crime. However, due to inconsistency in the crime data which vary drastically with time, it is a herculean task to choose appropriate methods for mining the crime pattern. Machine learning and data mining are widely used techniques in crime pattern detection. These are very apt techniques that provide some fairly interesting results.

Though there are various widely used techniques, like K-Nearest Neighbor, Naïve Bayesian, decision trees, support vector machines, etc., with the crime data growing exponentially, these techniques doesn't prove to provide accurate results. With the organized crime involving various gangs are also increasing, in order to understand the organizational structure, the data is modeled as a graph. This can provide some interesting results and can be useful for gang network analysis.

In this work, crime dataset from the UCI machine learning repository – the communities and crime data is taken for analysis. Two new algorithms, Random Forest and

A. Unal et al. (Eds.): SmartCom 2016, CCIS 628, pp. 873–880, 2016.
DOI: 10.1007/978-981-10-3433-6_104

Gradient Boosting Method are taken and the accuracy, precision and recall values of these algorithms are formulated. This work focuses on feature extraction using Boruta algorithm [1]. It is available as a package in R studio, an efficient tool for data mining. The random forest and GBM methods are performed using R to predict the status of the crime. The parameters such as precision, recall and accuracy are all calculated. R is a free software environment [9] for statistical computing. It is also used for predictive analytics.

The objective of this paper is to classify the crime status into one of the two classes - violent crime or non-violent crime based on the different parameters from the dataset. From the dataset using Boruta algorithm, the features are selected and the random forest and GBM methods are applied to the crime data for predicting the crime status (violent or non-violent crime). From these experimental analysis, parameters such as accuracy, precision and recall values are calculated based on the predicted crime status.

2 Related Work

Techniques such as Association mining, clustering, classification, prediction, machine learning are applied to predict crime status. This section elaborates on the closely related work to mining, machine learning, feature selection [10] and crime status prediction techniques.

Wortley and Mazerolle [2] explained the concept of crime pattern theory in a more suitable form that is straightforward. The authors suggested that there have to be simple rules and a procedure to join those rules in order to understand a complex pattern problem. Starting with individual offenders and moving on to a network of offenders, the authors suggested guiding rules for obtaining crime templates. They defined patterns as a recognizable inter-connectiveness of objects and concluded that crimes occur in patterns; the decision to commit crimes have patterns and the methods by which a crime is committed also occur in patterns.

Nath [3] presented the idea of applying clustering techniques to solve crimes faster by identifying crime patterns. The author suggested that k-means with some enhancements can be used to understand the crime patterns. The paper also had an implementation of semi-supervised learning along with the geo-spatial plot. The main contribution of this paper is applying data mining and machine learning techniques to the crime data and help police solve crime faster.

Buczak and Grifford [4] suggested that manually inspecting each and every crime is infeasible as the amount of data is too huge to be processed in a limited time frame. The research suggests the use of fuzzy association rule mining on crime dataset. The metrics and rules had helped them in identifying interesting and novel rules to extract the essence of crime patterns. The authors had also given a list of trends in crime analysis such as applying mining and machine learning, geo-spatial mechanisms, finding crime hot-spots, predicting gang criminal offenses, etc. The main contribution of the work by these authors lies in fuzzification of the data sets and the application of fuzzy apriori algorithm on the dataset to generate new rules. Some rare and interesting results were found on applying this algorithm. The work concludes on discovering certain crime patterns that are consistent across all regions and states.

The authors of [5] developed an incremental mining algorithm to discover the crime patterns through temporal association rules. Honk Kong district has been taken into consideration and the incremental algorithm called as ITAR has been developed by understanding the Modus Operandi of the criminals. The authors had applied an incremental algorithm as the size of the database keeps increasing dynamically with the happenings of a new crime. So, applying mining techniques on these, leads to an understanding of the crime patterns.

Friedman [6] explained the effect of randomization on gradient boost procedures. The sample size and distribution function taken are the most important characteristics that affect the performance of the problem. The gradient boosting technique constructs an additive regression model by fitting a function sequentially using least-square principles in each iteration. To incorporate randomness as a significant feature few minor modifications were made to the gradient boosting algorithm and henceforth called as stochastic gradient boosting. The performance metrics also depends on the error rate of the derived estimate. The author concluded that the accuracy of the procedure could be improved by increasing the randomness by taking few smaller samples in each iteration.

Flaxman [7] proposed a fully bayesian spatio-temporal model that incorporates Gaussian model to forecast the crime. The author used a relative risk service as the mean in a Poisson likelihood estimation. The crime data of Chicago city were used and grouped based on the type of crime, community area and week of the year. The author had chosen types of crime that are prone to happen frequently. Other than forecasting the crime, the model also is suitable for statistical inference like how the crime rates varied over a decade. The results of this model were compared with kernel intensity estimation approach which is the most widely used tool by the police department.

Somayeh Shojaee, Aida Mustapha, Fatimah Sidi, Marzanah A. Jabar [8] has done an extensive work in applying various classification algorithms for predicting the crime status using the crime and communities dataset from the UCI machine learning repository. They had used two different feature selection techniques - one is the manual technique that is the features are selected based on the human understanding of the domain. The second is a Chi-Square distribution which offers a feature selection package. They had applied the classification algorithms like Naive Bayesian, k-Nearest neighbor, Neural networks, Decision trees and support vector machines (SVM) and calculated the accuracy, precision, recall and area under the curve. Their worked comprised of pre-processing techniques and feature selection by the two methods as explained above. They had concluded that k-Nearest neighbor algorithm is an efficient algorithm to predict the crime status by way of visualizing the ROC curve.

In this study, we select two prospective algorithms namely Gradient Boosting and Random Forest algorithm, that were not discussed related to crime records. Also, we select a wrapper feature selection algorithm - Boruta algorithm another looming feature selection algorithm. The following section discusses these selected algorithms.

3 Analysis of Selected Algorithms

3.1 Boruta Feature Selection Algorithm

Boruta [1] is an all-relevant feature selection wrapper algorithm available as a package with R. It finds relevant features by comparing the importance of original attributes with the importance that is achieved at random, estimated using the permuted copies. It is available in the CRAN repository licensed by GPL. It has an important method called as attstats which show the summary of a Boruta run in an attribute-centered way. Boruta iteratively compares the importance of attributes by shuffling principle.

Those attributes which have more importance than the shadow attributes are labeled as confirmed and those attributes that are of low importance than the shadow attributes are labeled as rejected. Few attributes which have very close values to the shadow values are labeled as tentative. This happens iteratively till only the list of accepted attributes remain or it runs till a threshold value called as MaxRun is reached. Usually, MaxRun is set as 100. The object of Boruta class has various components like finalDecision that describes whether the attribute is confirmed, rejected or tentative, ImpHistory that projects the importance of attributes, timetaken – the time taken for a single run of the algorithm, etc.

3.2 Random Forest

The base of Random Forest is Classification and Regression Trees. Random Forest performs classification and regression based on a forest of trees that uses random inputs [11]. The main purpose of using a Random Forest algorithm is to improve the accuracy of prediction [12]. By growing an ensemble of trees drastic improvements in classification accuracy is found and then voting for the most popular class occurs. This procedure is referred as Random Forest procedure.

These forests use randomly selected inputs or combinations of split inputs at each node to grow the tree. It is found to be robust to outliers and noise. It is very simple to implement and can be easily parallelized. It also gives useful estimates of errors, strength, correlation and variable importance.

3.3 Gradient Boosting

Gradient boosting is a machine learning technique [13] that is also used for Classification and Regression. It produces an ensemble of prediction models. This algorithm works with a variety of loss functions. This algorithm builds additive tree-models in logistic regression. The size of the tree used in this method is a very significant factor. It inherits all the good features of the trees and improves on the weaker features such as the prediction performance.

This algorithm consists of two steps - a dictionary is constructed in the first step using the weak learners. The second step details about fitting a model [6]. This ensemble is grown in an adaptive fashion which is then averaged at the end.

4 Dataset Description

The Communities and Crime dataset [14] from UCI machine learning repository is considered for this study. The dataset comprises of 2215 instances with 147 attributes of which 125 attributes are predictive attributes, 4 attributes are non-predictive and 18 attributes are the goal attributes. This data set focuses on United States communities linked to the socio-economic status, Law enforcement Management administrative Statistics (LEMAS) and the crime statistics report of FBI [4]. The dataset was submitted to the UCI repository in July 2009. This work is only a study on the classification algorithms. We have used these algorithms for the crime records.

4.1 Data Preprocessing

The dataset contains lots of attributes with missing and inconsistent data. There are various methods for handling the missing values. Some of the widely used techniques are removing the missing values, manually entering the values, creating a prediction model which provides values to the missing data. In this work, we remove the attributes containing missing and inconsistent values. This is a very important step as completely removing an attribute may lead to a wrong result even if it contains lots of missing values. So when a particular attribute is completely removed, proper justification has to be maintained. This step is followed by a feature selection method which has two methods – one is the manual method of feature extraction and the other is using the Boruta wrapper algorithm.

4.2 Feature Selection

As mentioned above, the feature selection is done in two methods. The feature selection is mainly done in order to increase the performance and at the same time avoid overfitting of data. The two methods used are discussed below:

- Manual Method – In this method, based on the views from [4, 8] along with the human understanding of the scenario, we derived 38 attributes which gave better results while calculation of accuracy, precision and recall on applying random forest and gradient boosting ensemble methods.
- Boruta Algorithm – The Boruta wrapper runs till it reaches the maximum threshold and provides a result plot that suggests the important, tentative and rejected attributes. The green color represents the important attributes. Red color represents the rejected attributes and yellow suggests tentative attributes which may be considered as important.

The Boruta algorithm is run on the dataset with all the attributes and it yielded 55 attributes as important. With these 55 attributes, the accuracy precision and recall were calculated.

5 Results and Discussions

In this experiment, a comparison between the random forest and gradient boosting is performed over the crime dataset. The accuracy, precision and recall (in %) are calculated for the dataset using both these algorithms. Accuracy denotes the percentage of instances that are classified correctly. Precision shows what ratio of the data is classified correctly. Recall shows the percentage of information relevant to the class that is correctly classified.

The existing scheme in Table 1 denotes the results from the authors of [8]. That is we have assumed the existing scheme to be the best method from the discussions in [8]. As discussed above, they had used Chi-Square algorithm for feature selection and with that the accuracy, precision and recall were calculated. However, we have used the Boruta algorithm for feature selection.

A confusion matrix is built using both the algorithms. The confusion matrix consists of the following parameters: True positives (TP) - correctly identified instances, True Negatives (TN) - correctly rejected instances, False Positives (FP) - incorrectly identified instances and False Negatives (FN) - incorrectly rejected instances.

Table 1. Accuracy, Precision and Recall

Algorithm	Manual feature selection			Algorithmic feature selection		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Existing scheme	87.51	86.90	87.50	88.00	87.30	88.00
Random forest	95.60	97.20	96.66	98.00	98.66	97.44
Gradient boost method	97.80	99.72	98.66	98.40	99.17	98.62

The accuracy, precision and recall parameters used in this work are calculated using the below given formulas

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{P} + \text{N}) * 100. \tag{1}$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) * 100. \tag{2}$$

$$\text{Recall} = \text{TP}/\text{P} * 100. \tag{3}$$

where

$$\text{P} = \text{TP} + \text{FP}(\text{Total Positives}). \tag{4}$$

and

$$N = TN + FN(\text{Total Negatives}). \tag{5}$$

The Table 1 shown above shows the comparison between the accuracy, precision and recall values for the crime and community dataset taken under consideration. The existing scheme mentioned in the above table is the k-Nearest Neighbor algorithm taken from the work by the authors of [8].

The Fig. 1 shown above describes the graphical comparison between the two algorithms and the existing scheme for both the manual feature selection and algorithmic approach. The goal attribute here is whether the crime is a violent crime or nonviolent crime. It is clear from Fig. 1 that our work outperforms the existing scheme by a great margin.

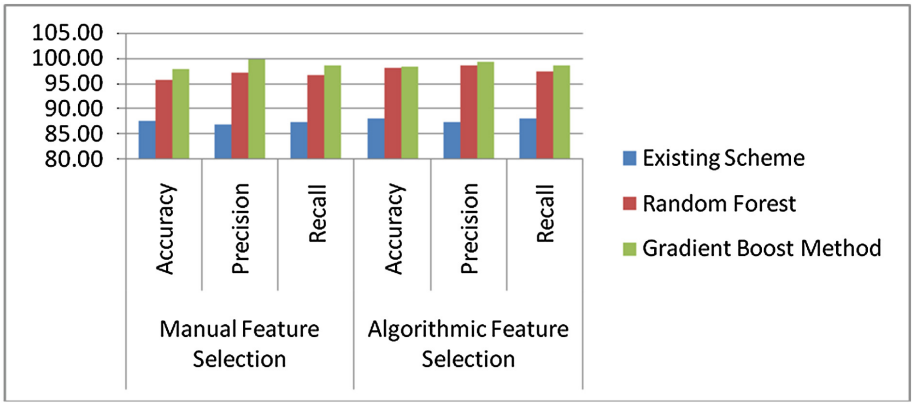

Fig. 1. Comparison of random forest and gradient boosting with the existing scheme

A classifier with high accuracy is termed as a good classifier. There is almost 8–10% increase in the accuracy percentage of our algorithms. Also using the Boruta feature selection method the number of attributes was also reduced significantly compared to chi-square method. Thus, with a limited number of parameters through the Boruta feature selection algorithm and with the random forest and gradient boosting algorithm, we have achieved higher accuracy and precision. Feature selection is proved to be one of the important aspects of data mining. It is very evident from the Table 1 that with the manual feature selection strategy we had 95.6% accuracy while the algorithmic approach to feature selection resulted in an accuracy of 98%.

6 Conclusion

The aim of this paper is to classify the crime as violent crime or non-violent crime. This work presents a detailed analysis of the two classification algorithms viz random forest and gradient boosting for predicting the crime status. The experimental results

suggested that the two algorithms we have considered performed better than the k-Nearest Neighbor method which was proposed as one of the best classification methods. Also, this work suggested using the Boruta wrapper algorithm that can be used for feature selection. The results also indicated that using this feature selection algorithm is also better than choosing the attributes manually with less domain knowledge in this area.

Acknowledgments. We extend our heartfelt gratitude towards SSN institutions for providing us with the necessary infrastructure and funding in carrying out this project.

References

1. Kursa, M.B., Jankowski, A., Rudnicki, W.R.: Boruta–a system for feature selection. Fundamenta Informaticae **101**(4), 271–285 (2010)
2. Wortley, R., Mazerolle, L. (eds.): Environmental Criminology and Crime Analysis. Willan, London (2013)
3. Nath, S.V.: Crime pattern detection using data mining. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, WI-IAT 2006 Workshops. IEEE (2006)
4. Buczak, A.L., Gifford, C.M.: Fuzzy association rule mining for community crime pattern discovery. In: ACM SIGKDD Workshop on Intelligence and Security Informatics. ACM (2010)
5. Ng, V., et al.: Incremental mining for temporal association rules for crime pattern discoveries. In: Proceedings of the Eighteenth Conference on Australasian Database, vol. 63. Australian Computer Society, Inc. (2007)
6. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**(4), 367–378 (2002)
7. Flaxman, S.R.: A general approach to prediction and forecasting crime rates with Gaussian processes (2014)
8. Shojaee, S., et al.: A study on classification learning algorithms to predict crime status. Int. J. Digit. Content Technol. Appl. **7**(9), 361 (2013)
9. Ihaka, R., Gentleman, R.: R: a language for data analysis and graphics. J. Comput. Graph. Stat. **5**(3), 299–314 (1996)
10. Tsymbal, A., Puuronen, S., Patterson, D.W.: Ensemble feature selection with the simple Bayesian classification. Inf. Fusion **4**(2), 87–100 (2003)
11. Random Forest. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
12. Random Forest. http://perso.math.univ-toulouse.fr/motimo/files/2013/07/random-forest.pdf
13. Gradient Boosting. https://cran.r-project.org/web/packages/gbm/gbm.pdf
14. Crime dataset. http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime