ORIGINAL RESEARCH PAPER

Volume-8 | Issue-6 | June-2019 | PRINT ISSN No. 2277 - 8179

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

A RANDOM FOREST ALGORITHM FOR CRIME PREDICTION OF BIGDATA ANALYSIS IN R

Computer Science

Selva Priya. T Student Department of Computer Science and Engineering Lord Jegannath College of Engineering and Technology

M. Thuraipandian* Associate Professor Department of Computer Science and Engineering Lord Jegannath College of Engineering and Technology *Corresponding Author

ABSTRACT

Random forest is an Ensemble Classifier made using many Decision Tree models. Finding the place of criminal activity is vast amount to prevent it. The law enforcement agencies can work effectively and respond faster if they have good knowledge about crime pattern in different geological points of a city. Crime prediction using Random Forest Algorithm of Machine Learning with Supervised Learning Approach (Regression and Classification), Decision Tree Classifier and Big Data Analysis. We have collected the crime description dataset and experiment result shows that the ensemble Random Forest Algorithm outperformed the prediction crime with decision tree classifier and other classification algorithm in both performance and accuracy within the Bigdata. And the forecast results are visualized using R programming language. Bigdata is the collection of dataset.so large and complex difficult to handle hand, on database. It is useful for improving the accuracy and speed crime prediction.

KEYWORDS

INTRODUCTION

A crime is an unlawful act punishable by a state or any other authority. The term "crime" in modern criminal law, have any simple and universally accepted definition though statutory definitions have been provided for certain purposes. The most popular view is that crime is a category created by law; in other words, something is a crime if declared as such by the relevant and applicable law.

One proposed definition is that a crime or offence is an act harmful not only to some individual but also to a community, society. Such acts are banned and punishable by law. Random is subsets which are selected by random. Forest is a collection of tree. Random forest is a versatile algorithm capable of performing both regression and classification. It is the type of Ensemble Learning Method. It is commonly used to predict modelling with accuracy. Ensemble learning refers to the algorithms that produce collections or ensembles of classifiers which learn to classify by training individual learners and fusing their predictions. Growing an ensemble of trees and getting them vote for the most popular class has provided a good enhancement in the accuracy of classification. Often, random vectors are built that control the growth of each tree in the ensemble.

The ensemble learning methods can be divided into two main groups: bagging and boosting. In bagging, models are fit in parallel where successive trees do not depend on previous trees. Each tree is independently built using bootstrap sample of the dataset. A majority vote determines prediction. In boosting, models are fit sequentially where successive trees assign additional weight to those observations poorly predicted by previous model. A weighted vote specifies prediction. The main classification of the random forest methods are machine learning technique and supervise learning model. Machine learning is a type of artificial intelligence that recognizes that patterns using data analysis. A computer can learn and make predictions from data through machine learning without being explicitly programmed. Machine learning methods attempt to find the class of new sample using the way that was learned from classification. The machine learning can be divided into three main categories: Supervised Learning, Unsupervised Learning and Reinforced Learning. In this paper, Supervised Learning methods are used to predict crime categories.

Supervised learning is a machine learning model that can predict an output from a set of inputs. In supervised learning, the output labels are specifically defined. Input object contains various numbers of features and usually is represented in a vector form. In the training dataset, each input object is paired with a specific output object. A supervised learning algorithm develops a predictive model using training data helps supervised learning models to avoid over fitting. The label of new information is predicted by the algorithm. Supervised learning models can be implemented on both classification and regression problems.

RELATED WORKS

In [1] Weiweilin, (2017) "**An Ensemble Random Forest Algorithm For Insurance Big Data Analysis**" which deals with the imbalanced distribution of business data, missing of user features and many other reasons, directly using big data techniques on realistic business data tends to deviate from the business goals. It is difficult to model the insurance business data by classification algorithms like Logistic Regression and SVM etc. In this paper, we exploit a heuristic bootstrap sampling approach combined with the ensemble learning algorithm on the large-scale insurance business data mining, and proposes an ensemble random forest algorithm which used the parallel computing capability and memory-cache mechanism optimized by Spark. We collected the insurance business data from China Life Insurance Company to analyze the potential customers using the proposed algorithm. We use F-Measure and G-mean to evaluate the performance of the algorithm. Experiment result shows that the ensemble random forest algorithm outperformed SVM and other classification algorithms in both performance and accuracy within the imbalanced data, and it is useful for improving the accuracy of product marketing compare to the traditional artificial approach.

In [2] Richard a. Bauder, (2018)"**Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data**",which describes the healthcare industry generates a substantial amount of data. This big data includes information such as patient records and provider payments. The use of big data is often considered the best way to produce effective models in areas such as fraud detection. In this study, we demonstrate that the use of more highly imbalanced big data does not produce acceptable fraud detection results. We use random under sampling to generate seven different class distributions and compare performance results. We use the 2012-2015 Calendar Year Medicare Provider Utilization and Payment Data mapping actual fraud labels from the List of Excluded Individuals/Entities (LEIE). Our results, based on building Random Forest models using 5-fold cross-validation, demonstrate that 90:10 is the best class distribution with a 0.87302 AUC, whereas the balanced and two highly imbalanced distributions produced the worst fraud detection performance. Furthermore, we show that the commonly used ratio of 50:50 (balanced) was not significantly better than using a 99:1 (imbalanced) class distribution. Our study clearly demonstrates the need to apply at least some sampling to big data with class imbalance and suggests the 50:50 class distributions do not produce the best Medicare fraud detection results.

In [3] Jianguochen, (2016)" **A Parallel Random Forest Algorithm For Big Data In A Spark Cloud Computing Environment**" deals with the emergence of the big data age, the issue of how to obtain valuable knowledge from a dataset efficiently and accurately has attracted increasingly attention from both academia and industry. This paper presents a Parallel Random Forest (PRF) algorithm for bigdata on the Apache Spark platform. The PRF algorithm is optimized based

on a hybrid approach combining data-parallel and task-parallel optimization. From the perspective of data-parallel optimization, a vertical data-partitioning method is performed to reduce the data communication cost effectively, and a data-multiplexing method is performed is performed to allow the training dataset to be reused the training process of RF, and a task Directed Acyclic Graph (DAG) is created according to the parallel training process of PRF and the dependence of the Resilient Distributed Datasets (RDD) objects. Then, different task schedulers are invoked for the tasks in the DAG. Moreover, to improve the algorithm's accuracy for large, high-dimensional, and noisy data, we perform a dimension-reduction approach in the training process and a weighted voting approach in the prediction process prior to parallelization. Extensive experimental results indicate the superiority and notable advantages of the PRF algorithm over the relevant algorithms implemented by Spark MLlib and other studies in terms of the classification accuracy, performance, and scalability.

In[4] Nejdetdogru (2018),"**Traffic Accident Detection Using Random Forest Classifier**" ,The Internet of Things (IoT) has been growing in recent years with the improvements in several different applications in the military, marine, intelligent transportation, smart health, smart grid, smart home and smart city domains. Although IoT brings significant advantages over traditional information and communication (ICT) technologies for Intelligent Transportation Systems (ITS), these applications are still very rare. Although there is a continuous improvement in road and vehicle safety, as well as improvements in IoT, the road traffic accidents have been increasing over the last decades. Therefore, it is necessary to find an effective way to reduce the frequency and severity of traffic accidents. Hence, this paper presents an intelligent traffic accident detection system in which vehicles exchange their microscopic vehicle variables with each other. The proposed system uses simulated data collected from vehicular ad-hoc networks (VANETs) based on the speeds and coordinates of the vehicles and then, it sends traffic alerts to the drivers. Furthermore, it shows how machine learning methods can be exploited to detect accidents on freeways in ITS. It is shown that if position and velocity values of every vehicle are given, vehicles' behavior could be analyzed and accidents can be detected easily. Supervised machine learning algorithms such as Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Random Forests (RF) are implemented on traffic data to develop a model to distinguish accident cases from normal cases. The performance of RF algorithm, in terms of its accuracy, was found superior to ANN and SVM algorithms. RF algorithm has showed better performance with 91.56% accuracy than SVM with 88.71% and ANN with 90.02% accuracy.

In [5] A.z. kouzani, s. Nahavandi and k. Khoshmanesh (2007)," **Face Classification by a RandomForest**" this presents a random forest-based face image classification method. The random forest is an ensemble learning method that grows many classification trees. Each tree gives a classification. The forest selects the classification that has the most votes. Three experiments are performed. The random forest-based method together with several existing approaches are trained and evaluated. The experimental results are presented and discussed.

In [6] Jianwuzhang, (2018) "**Face Recognition Model Based on Privacy Protection and Random Forest Algorithm**" With the rapid development of face recognition technology, people's awareness of privacy protection is getting higher and higher. Arnold transform because of its simple and easy to be applied to digital watermarking, security and other fields, has achieved good results. In this paper, Arnold transform is applied to face images, and propose scrambled face recognition of random forest (SFR-RF) model based on random forest (RF) algorithm. The model extracts the features of face database, makes label classification data, forms training set and testing set, and recognizes scrambled face by RF classification algorithm. The experimental results show that the SFR-RF model achieves good results in the area of face image scrambling. It improves the recognition rate by 20% compared with the traditional decision tree algorithm, making the proposed method a good candidate for face privacy protection and identification.

In [7]Junshixia member (2018),"**Random Forest Ensembles And Extended Multiextinction Profiles For Hyperspectral Image Classification**", Classification techniques for hyperspectral images based on random forest (RF) ensembles and extended multiextinction profiles (EMEPs) are proposed as a means of improving performance. To this end, five strategies—bagging, boosting, random subspace, rotation-based, and boosted rotation-based—are used to construct the RF ensembles. EPs, which are based on an extrema-oriented connected filtering technique, are applied to the images associated with the first informative components extracted by independent component analysis, leading to a set of EMEPs. The effectiveness of the proposed method is investigated on two benchmark hyperspectral images: the University of Pavia and Indian Pines. Comparative experimental evaluations reveal the superior performance of the proposed methods, especially those employing rotation-based and boosted rotation-based approaches. An additional advantage is that the CPU processing time is acceptable.

In[8]ArnuPretorius,SuretteBierman and Sarel J. Steel, (2016)"**A Meta-Analysis of Research in Random Forests for Classification**" Random Forests (RFs) have successfully been employed in a vast array of application areas. Fairly recently, a number of algorithms that are related to Breiman's original Forest-RI algorithm have been proposed in the literature. In this paper we conduct a meta-analysis of all (34) 2001-2015 papers that could be found in which a novel RF algorithm was proposed and compared to already established RF algorithms. The analysis revealed several limitations regarding the choice of performance measures, the way in which these measures are estimated, and the methodology for comparisons of multiple algorithms over multiple data sets. In fact, it is shown that in almost a third of the results from RF research papers, a significant improvement over the performance of Forest-RI is not found when comparisons are made using appropriate statistical tests.

EXISITNG SYSTEM

The large number of crimes happening in the globe. This problem has been resolved to provide some concrete solution. The comprehensive analysis for crime prediction in smart city using R programming is focused upon how data science could help too gain important insight from a crime data. The R programming has shown that how these insights discovered from the historical crime data which can be used to potentially predict crimes by combining with other relevant data sources.

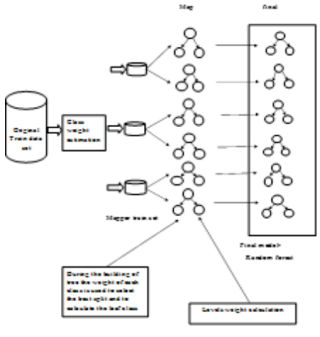
DISADVANTAGES

Random forest has been observed to over fit for some datasets with noisy classification/regression task. For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Therefore the variable importance scores from random forest are not reliable for this type of data. The large number of trees may make the algorithm slow for real-time prediction.

PROPOSED SYSTEM

In the proposed system the data related to the crime details are collected and stored in the database. The random forest algorithm of machine learning with SVM, decision tree can be producing the prediction of crime big data. We have collected the crime description dataset and experiment result shows that the random forest algorithm outperformed the prediction with accuracy. And it is useful for improving the speed to predict the best data. The large number of trees (Decision tree) may make the algorithm slow in real time prediction which can be overcome and speedup the prediction of data which is the major proposal of this system. If any crime occurred in the particular area, when they search the database, the details related to that crime are displayed. By using that details they can easily identified the crime occurred in that particular area.

SYTEM ARCHITECTURE



SYSTEM IMPLEMENTATION

System Module

- Data Processing
- Mapping Training Set
- Feature Selection
- Final Prediction

MODULES DESCRIPTION

DATA PROCESSING

Data processing is generally the collection and manipulation of items of data to produce meaningful information. In this sense it can be considered a subset of information processing, the change of information in any manner detectable by an observer.

This simple observation led to the idea that it would be useful to use only some selected trees in classification. The selection of trees was based on their performance on similar instances, but without success.

The step toward the analysis is preprocessing. If the data is dirty, it will generate incorrect visualizations, hence leading towards the incorrect conclusions. The crime data collected also has some level of dirtiness. It contains some null values, inconsistent date formats, and some outliers.

Using R, an exploratory analysis is carried out to identify this dirtiness in the data. Then, by using some data cleaning procedures follows a well formatted and clean dataset which will be used to carry out the analysis (i.e. Crime hours, dates, times, description, weapon, district, neighborhood, year, month, day).

Mapping Training Set

Considering the different random subsets of features to split on at each tree node. Apart from these randomizations, decision tree training is done in the same way as for individual decision trees. Training and test data can be supplied in two forms. Mapping is used to allow the work on small subset and work parallel. The original train data set divided by the class estimation weight of data.

Feature selection

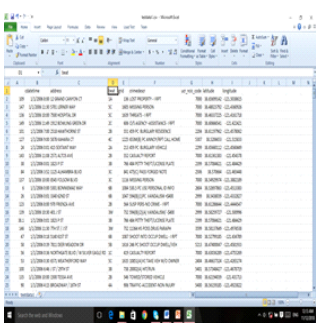
Feature selection is also known as variable selection. It is the automatic selection of attributes in data that are most relevant to the predictive modeling problem.Random split selection does better than bagging; introduction of random noise into the outputs also does better; but none of these do as well as adaboost by adaptive reweighting (arcing) of the training set. The importance of each feature variable in a training subset refers to the portion of the gain ratio of the variable compared with the total feature variables. The value of all feature variables are sorted in descending order and the top variable values are selected. Thus the number of dimensions of the dataset is reduced from feature variables in each sample to the number of the selected feature variables.

Final Prediction

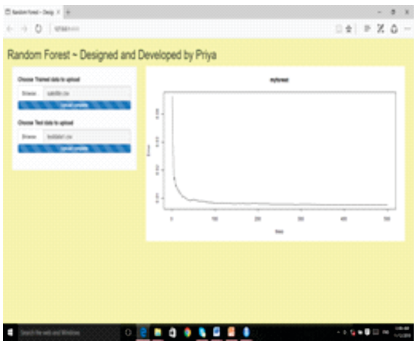
To make a prediction on a new instance, a random forest must aggregate the predictions from its set of decision trees. This aggregation is done differently for classification and regression.The data gets split into many subsets and it compares the train and test data to find the best one. This process gets repeatedly on each subset and find out the best prediction on each mapping. According to this process, each subset has its own predicted class. And comparing all the predicted class of its produce the final prediction based on training data.

RESULT

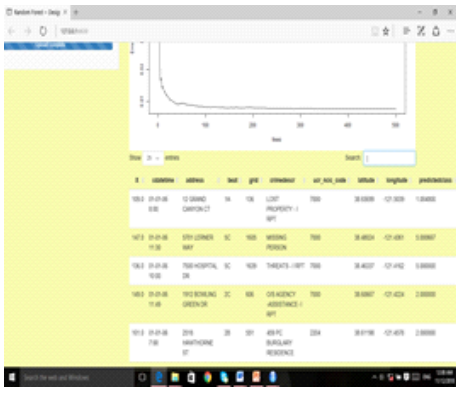
Dataset



Random Forest Graph



Prediction



CONCLUSION

With the advancements in the technology, we are stepping forward to the future of smart cities. But still we have large number of hurdles to achieve that vision. One such problem faced by our society is the large number of crimes happening across cities. So it is very cardinal that we find solution to this problem. However in the recent years, a comprehensive analysis for crime prediction in smart city has proved effectively. In this paper, we have used random forest algorithm to build our predictive model and our model has produced really impressive results with accuracy. Using social networking to predict crimes is a latest problem being worked upon. It is not an easy to achieve, working with the crime data offers a lot of challenges. The data usually is full of error and incomplete. And also there is a lack of proper standards across the globe to record the crimes that happen. Not having a proper infrastructure that could accommodate the various aspects of crime analysis is a major issue. Even if the data is available, it is not easy to accommodate that data with another framework to achieve more concrete predictions. These are challenges for crime analysis to handle. In future, manage the large number of trees which may make the algorithm very slow for real time prediction.

REFERENCES

1. Weiweilin,"An Ensemble Random Forest Algorithm For Insurance Big Data Analysis",IEEE Access, vol. 2, no. 7, July 2017
2. Richard A. Bauder,"Medicare Fraud Detection Using Random Forest With Class Imbalanced Big Data",2018 IEEE International Conference on Information Reuse and Integration for Data Science.
3. JianguoChen,"A Parallel Random Forest Algorithm For Big Data In A Spark Cloud Computing Environment",IEEE Transaction on Parallel and Distributed Systems.
4. NejdetDogru,"Traffic Accident Detection Using Random Forest Classifier",978-1-5386-2659-7/18/$31.00 ©2018 IEEE.
5. A.Z. Kouzani, S. Nahavandi, K. Khoshmanesh," Face Classification By A Randomforest", 1-4244-1272-2/07/$25.00 ©2007 IEEE
6. JianwuZhang,"Face Recognition Model Based On Privacy Protection And Random Forest Algorithm",The Wireless and Optical Communications Conference(WOCC2018)
7. JunshiXia, Member,"Random Forest Ensembles And Extended Multi extinction Profiles For Hyperspectral Image Classification", IEEE Transactions On Geoscience And Remote Sensing, Vol. 56, No. 1, January 2018
8. Arnu Pretorius, Surette Bierman And Sarel J. Steel,"A Meta-Analysis Of Research In Random Forests For Classification",2016 International Conference(PRASA-RobMech)
9. AngshumanPaul,"Improved Random Forest For Classification",2018 IEEE Transactions on Image Processing.
10. Simon Bernard, Laurent Heutte And SebastienAdam,"On The Selection Of Decision Trees In Random Forests",Proceedings of International joint Conference on Neural Networks, Atlanta, Georgia,USA, June 14-19,2009.