IMPLEMENTATION OF MACHINE LEARNING ALGORITHMS FOR AGRICULTURE

PROF.KALAM NARRE
[1]Assistant Professor, CSE,Chalapathi Institute of Technology,Guntur, India
[2]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[3]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[4]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[5]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

ABSTRACT : In this paper author proposed that fraud detection is a critical problem affecting large financial companies that have increased due to the growth in credit card transactions. This paper presents detection of frauds in credit card transactions, using data mining techniques of Predictive modeling, logistic Regression, and Decision Tree. The data set contains credit card transactions in September 2013 by European cardholders. This data set present transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The data set is highly unbalanced, the positive class(frauds) Account for 0.172% of all transactions.

1. INTRODUCTION

Credit card fraud is a major problem that involves payment card like credit card as illegal source of funds in transactions. Fraud is an illegal way to obtain goods and funds. The goal of such illegal transaction might be to get products withoutpaying or gain an unauthorized fund from an account. Identifying such fraud is a troublesome and may risk the business and business organizations. In the real world FDS [1], investigator are not able to check all transactions. Here the Fraud Detection System monitors all the approved transactions and alerts the most suspicious one. Investigator verifies these alerts and provides FDS with feedback if the transaction was authorized or fraudulent. Verifying all the alerts everyday is a time consuming and costly process. Hence investigator is able to verify only few alerts each day. The rest of the transactions remain unchecked until customer identifies them and report them as a fraud. Also the techniques used for fraud and the cardholder spending behavior changes over time. This change in credit card transaction is called as concept drift [1] [7]. Hence most of the time it is difficult to identify the credit card fraud. Machine Learning is considered as one of the most successful technique for fraud identification. It uses classification and regression approach for recognizing fraud in credit card. The machine learning algorithms are divided into two types, supervised [14][18] and unsupervised [16] learning algorithm. Supervised learning algorithm uses labeled transactions for training the classifier whereas unsupervised learning algorithm uses peer group analysis [23] that groups customers according to their profile and identifies fraud based on customers spending behavior.Many learning algorithm have been presented for fraud detection in credit card which includes , Logistic Regression [3], decision tree, Naive Bayes [6], Support Vector Machines [5], K-Nearest Neighbors [6] and Random Forest. This paper examines the performance of above algorithms based on their ability to classify whether the transaction was authorized or fraudulent and then compares them. The comparison

is made using performance measure accuracy, specificity and precision. The result showed that Random Forest algorithm showed better accuracy and precision than other techniques.

2. LITERATURE REVIEW

The author [1] has proposed a paper where they have first explained the proper performance measures which is used for fraud identification. The authors have structured a novel learning technique that can solve concept drift, verification latency, and class imbalance issues. The paper also showed effect of above issues in true credit card transactions.

Here in paper [2] authors presented two types of classifier using random forests which are used to train the behavior features of transactions. The authors have compared the two random forests and have analyzed their performance on fraud identification in credit card.

In paper [3] authors presented a FDS for credit card using Artificial Neural Network and Logistic Regression. The system used to monitor each transaction separately using classifier and then classifier would generate score for each transaction and label this transaction as legal or illegal transaction. A decision tree method was proposed

In paper [4]. The method decreased overall misclassification costs and selected splitting property at each node. The author also compared the decision tree method for fraud identification with other models and proved that this approach performs well using performance measure like accuracy and genuine positive rate.

The author [5] developed a FDS for credit card transaction using support vector machines and decision tree. This study built seven alternative models that were created using support vector machines and decision tree. The author also compared this classifiers performance using performance measureaccuracy. The study also showed that as size of training dataset increases the number of fraud detected by SVM are less than fraud identified by decision tree method.

Here in [6] author presented fraud detection system using a Naive Bayes K-Nearest Neighbors method. The main aim of proposed system was to improve accuracy. Naive Bayes Classifier predicts probabilities of fraud in transaction while KNN classifier predicts how near the undefined sample data is to kth training dataset. The author compared both this classifier and showed that both work differently for given dataset. Most of predictive model used for detecting fraud in credit card transaction faces the issue of concept drift.

The author [7] presented two FDS based on sliding window and ensemble learning and showed that classifier need to be trained separately using feedback and delayed samples. The outcome of the two was than aggregated to improve the alert precision in FDS. Thus the author showed that to solve the issue of concept drift, the feedback and delayed samples are to be handled separately.

3. EXISTING SYSTEM

Credit card frauds are easy targets. Without any risks, a significant amount can be withdrawn without the owner's knowledge, in a short period. Fraudsters always try to make every fraudulent transaction legitimate, which makes fraud detection very challenging and difficult task to detect.

4. PROPOSED SYSTEM

Users can unfamiliarity is a very difficult problem in real-world when are called concept drift problems. Concept drift can be said as a variable which changes over time and in unforeseen ways. These variables

cause a high imbalance in data. The main aim of our research is to overcome the problem of Concept drift to implement on real-world scenario.

5. IMPLEMENTATION

Preprocessing of Data:

Following are the Preprocessing steps that have been carried out:

Importing Data

Importing Data set in CSV format file.

Checking the Missing Values in Data set

Balanced Data Set

Here it can be observed that the dataset is highly inbalanced, and thus for accurate ML predictions and training, a balanceddataset has to be created.

Feature Scalling

Train data is fitted to a suitable classifier upon feature extraction,then once the classifier is trained enough then we predict the results of the test data using the classifier, then compare the original value to the value returned by the classifier.

Modelling

Here We are applied Various Machine learning algorithms applied. Such as

KNN

LogisticRegression

DecisionTree

Random Forest

Navie Bayes

SVM

Result Analysis:

Here the comapre of different classifiers are shown among which the best classifier with highest accuracy percent is the chosen. Some factors such as f1-score, recall,precision. etc., also accounts for consideration of the classifiers.

Visual Representation:

Our final results are plotted as charts which contains different fields such asGenunie, Fraud inanalysis. Thus it is chosen ML Models

6. ALGORITHMS

Decision tree classifiers

Decision tree classifiers are used successfully in many diverse areas**.** Their mostimportant feature is the capability of capturing descriptive decisionmaking knowledge from the supplied data**.** Decision tree can be generated from training sets**.** The procedure for such generation based on the set of objects **(**S**)**, each belonging to one of the classes C1, C2, **…**, Ck is as follows**:**

Step 1**.**If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2**.**Otherwise, let T be some test with possible outcomes O1, O2,**…**, On**.** Each object in S has one outcome for T so the test partitions S into subsets S1, S2,**…**Sn where each object in Si has outcome Oi for T**.** T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si**.**

Gradient boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

- Simple, but a very powerful classification algorithm
- Classifies based on a similarity measure

- Non-parametric
- Lazy learning
- Does not "learn" until the test example is given
- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Example

- Training dataset consists of k-closest examples in feature space
- Feature space means, space with categorization variables (non-metric variables)
- Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training dataset

Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name multinomial logistic regression is usually reserved for thecae when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables.Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

Naïve Bayes

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis:it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature .

Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast

even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier.In the first part of this tutorial, we present some theoretical aspects of the naive bayesclassifier.Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminate analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others.In the second part, we use various tools on the same dataset (Weka 3.6.0, R 2.9.2,Knime 2.1.1,Orange 2.0b and Rapid Miner 4.6.0). We try above all to understand the obtained results.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance. The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's"bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance.

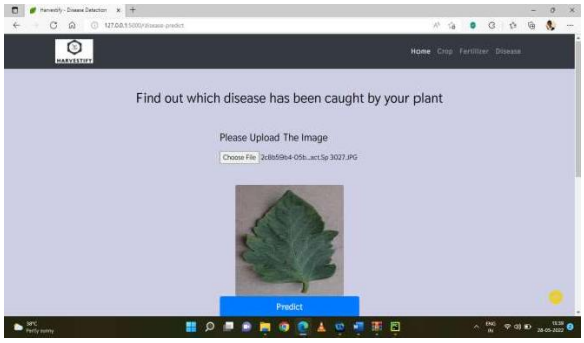
Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an independent and identically distributed (iid) training dataset, a discriminant function that can correctly predict labels fornewly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifieris equivalent to finding the equation for a multidimensional surface that best separates the different classesin the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to genetic algorithms **(**GAs**)** orperceptrons, both of which are widely used for classification in machine learning**.** For perceptions, solutions are highly dependent on the initialization and termination criteria**.**For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GAclassifier models are different each time training is initialized**.** The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyper planes' meeting this requirement**.**

7. RESULT



8. CONCLUSION

A new method for identifying fraud in credit card transactions has been proposed. The strategy was tested using a real-world dataset. The results demonstrate the possibility of fusing. many simple classification algorithms to generate a reliablesolution.The need for a very low false positive rate as well as the This is a fairly busy site with a lot of valid transactions. a challenging issue The outcomes demonstrate the capabilities of the approach proposed to save at least 30% of the money fraudulent credit card activities have put you at risk.

REFERENCES

[1] Jalinus, N., Nabawi, R. A., &Mardin, A. (2017). The Seven Steps of Project-Based Learning Mod‹l to Enhance Productive Competences of Vocational Students. In 1st International Conference on Technology and Vocational Teacher (ICTVT 2017). Atlantis Press. Advances in Social Science, Education and Humanities research (Vol. 102, pp. 251-256).

[2] Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippiand GianlucaBotempi, ‖Credit card Fraud Detection : A realistic Modeling and a Novel Learning Strategy‖, IEEE Trans. on Neural Network and Learning system,vol.29,No.8, August 2018.

[3] ShiyangXuan,GuanjunLiu,ZhenchuanLi, LutaoZheng,Shuo Wang, Jiang,‖Random Forest for credit card fraud detection‖,Int.conf.on Networking,Sensing and control,2018.

[4]Y. Sahin , and Duman,E.,(2011) —Detecting credit card fr ɛud by ANN and logistic regression.‖ In Innovations in Intelligent Systems and Applications(INISTA),201 1 international Symposium on (pp.315-319).IEEE

[5]Y. Sahin, S.Bulkan, and E. Duman, —A cost-sensitive decision tree approach for fraud detection,‖ Expert Syst. Appl., vol. 40, no. 15,pp. 5916–5923, 2013

[6]Sahin Y.and Duman E. (2011),‖Detecting Credit Crad Fraud by Decision Trees and Support Vector Machines‖, International Multi-Conference Of Engineers and Computer Scientists(IMECS 2011),Mar 16-18,Hong Kong,Vol.1,pp.1-6

[7]Sai Kiran, Jypti Guru, Rishabh Kumar, Naveen Kumar, Deepak Katariya,Credit card fraud detection usingNaïve Bayes model based and KNN classifier‖, Int. Journal of Adv. Research , Ideas and Innovations in Technology,vol.4,2018.

[8] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, ―Credit card fraud detection and concept-drift adaptation with delayed supervised information, in Proc. Int. Joint Conf. Neural Netw., 2015,pp. 1–8.
[9] A. C. Bahnsen, D. Aouada, and B.Ottersten, ―Example-dependent cost-sensitive decision trees,‖ Expert Syst. Appl., vol. 42, no. 19,pp. 6609–6619, 2015
[10] A. Dal Pozzolo, O. Caelen, and G. Bontempi, ―When is undersampling effective in unbalanced classification tasks?‖ in Machine Learning and Knowledge Discovery in Databases. Cambridge, U.K.: Springer, 2015
[11] N. Mahmoudi and E. Duman, ―Detecting credit card fraud by modified fisher discriminant analysis,‖ Expert Syst. Appl., vol. 42, no. 5,pp. 2510–2516, 2015
[12] A. C. Bahnsen, D.Aouada, A.Stojanovic, and B. Ottersten, ―Detecting credit card fraud using periodic features,‖ in Proc. 14th Int. Conf. Mach. Learn. Appl., Dec. 2015, pp. 208–213.
[13] AnuruddhaThennakoon, Chee Bhagyani, SasithaPremadasa, ShalithaMihiranga, Realtime Credit Card Fraud Detection Using Machine Learning ,‖Int. Conf. on Cloud Computing, Data Science & Engineering,2019.
[14] S. Wang, L.L.Minku, and X. Yao, ―Resampling-based ensemble methods for online class imbalance learning.‖ Trans, Knowl., Data Eng., vol 27, no. 5, pp. 1356-1368, May 2015.
[15] Jan may Kumar Behera, SuvasiniPanigrahi, ―Credit Card Fraud Detection: A Hybrid Approach using Fuzzy Clustering and Neural Network‖, 2015 IEEE Second International Conference on Advances in Computing and Communication Engineering.
[16] M. Carminati, R.Caron, F.Maggi, I.Epifani, and S. Zanero, BankSealer: ―A Decision Support System for Online Banking Fraud Analysis and Investigation‖, Berlin, Germany: Springer, 2014, pp. 380–394
[17] R. J. Bolton and D. J. Hand, ―Unsupervised profiling methods for fraud detection, in Credit Scoring Credit Control VII. London, U.K.: Imperial College London, 2001, pp. 235–255
[18] R. Elwell and R. Polikar, ―Incremental learning of concept drift in nonstationary environments,‖ Trans. Neural Netw., vol. 22, no. 10,pp. 1517–1531, 2011.
[19] S. Bhattacharyya, S.Jha, K.Tharakunnel, and J. C. Westland, ―Data mining for credit card fraud: A comparative study,‖ Decision SupportSyst., vol. 50, no. 3, pp. 602–613, 2011.
[20] Tao Guo ,Gui-Yang Li ,‖Neural data mining for credit card fraud detection‖,Int.Conf. on Machine Learning and Cybernetics, Sept 2008
[21] J. Gao, B.Ding, W.Fan, J.Han, P.S.Yu, ―Classifying data streams with skewed class distributions and concept drifts‖, IEEE internet comput., vol.12, no. 6, pp. 37-49, Nov 2008
[22] E. Aleskerov, B. Freisleben, and B. Rao, ―CARDWATCH: A neural network based database mining system for credit card fraud detection,‖ in Proc. IEEE/IAFE Computat. Intell. Financial Eng., Mar. 1997, pp. 220–226.
[23] J.R. Dorronsoro, F.Ginel, C. Sgnchez and C.S. Cruz, ―Neural fraud detection in credit card operations‖, IEEE transaction neural network vol. 8, no. 4, pp. 827-834, Jul.1997.
[24] D. J. Weston, D. J. Hand, N. M. Adams, C. Whitrow, and P. Juszczak,―Plastic card fraud detection using peer group analysis,‖ Adv. Data Anal.Classification, vol. 2, no. 1, pp. 45–62, 2008.