International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 5
ISSN: 2454-4248
19 – 23

A Comparative Study of Classification Techniques for Fraud Detection

Er. Monika[1]
Department of Computer Science and Engineering,
Chandigarh University
Gharuan, Mohali, India
monikachouhan472@gmail.com[1]

Er. Amarpreet Kaur
Department of Computer Science and Engineering,
Chandigarh University
Gharuan, Mohali, India
amarpreet844@gmail.com[2]

Abstract-There is large volume of data generated each day and the handling such large volume of data is very cumbersome. The generated data is stored in huge databases which can be retrieved as per the user. There are large sized repositories and databases generated in which the data can be stored. However, the retrieval of important data from such large databases is a major concern. There are numerous tools presented which can help in extracting useful information from the databases as per the requirement of users. The mechanism through which the data can be stored and extracted efficiently as per the requirement is known as data mining. This review paper studied about the classification techniques on the basis of different types of algorithms like Decision tree, Naïve bayes, Rule based, K-NN(K Nearest Neighbour), Artificial Neural Network. It describe the uses of various classification algorithm for develop a predictive model which is useful in different fields like Software fault prediction , credit card fraud analytics, and intrusion detection, medical and so on with respect to accuracy during the past few years.

Keywords— *Classification Approaches, Decision Tree, Naïve Bayes, Rule based, K-NN and Neural* Network

*****

I. INTRODUCTION

Because of headway in social sites, online business, business demonstrating and different colossal innovations, different enormous datasets are accessible which can be utilized for prescient displaying. Headway in correspondence and data, scholastics partners are seeing distinct fascination in prescient demonstrating of datasets. Prescient displaying investigation is performed to assist scholastics associations with driving advancement by securing new knowledge into their customers[1]. Today's datasets comprises of very huge and fast values of formats to process using existing algorithms. Such huge datasets are available in both structured and unstructured format generated from vide set of scenarios devices and applications. The process of capturing, analysis, storage and visualization of this enormous data is done by various existing data mining techniques[2].

In today's era, such huge data needs a system for processing across various domains which require real time response on datasets for faster decision making. There are various examples where data mining algorithms can be implemented such as network fault prediction from sensor data, credit card fraud analytics, and intrusion detection and so on. If decisions such as these are not taken in real time, the opportunity to mitigate the damage is lost. The heart of any prediction system is the model. There are several data mining and machine learning algorithms are available for different types of prediction system. Any prediction system is built to provide higher probability of correctness if it uses good training samples.

This investigation indicates how prescient examination can be actualized utilizing open source advances and different information mining calculations to process ongoing information in blame tolerant way in adaptable and effective

II. LITERATURE REVIEW

[3] Presented, a novel framework is purposed for settling on suitable choices identified with the endorsement or dismissal of load demand of a client based on different definite data. This paper dissected execution of different classifiers, for example, choice tree, bolster vector machine, versatile boosting and so forth. The execution and precision accomplished through the irregular woods arrangement is higher in contrast with rest of the calculation. [4] Introduced, a financial fraud has many effects in the financial industry and author also investigated the performance of different methods such as Naïve Bayes, K-nearest neighbour and Logistic regression models. They are utilized for the detection of the fraud in the financial industry. For the detection of credit card fraud in online transactions an essential role is played by the data mining. This fraud of credit card becomes a major challenge due to two major factors such as first, behaviour change in the profile of normal and fraudulent. Second factor is due to highly twisted data sets in the credit cards fraud. There are many factors that affect the performance of the fraud detection these issue are sampling approach on dataset, selection of variables and detection techniques. The performance of naïve bayes, k-nearest neighbour and logistic regression are investigated on the highly skewed credit card fraud data. These techniques are applied on the raw and pre-processed data. The work is implemented in Python and on the skewed data a hybrid technique was applied that carried out its functioning on the under-sampling and oversampling. Author concluded that performance of three techniques is measured on the basis of accuracy, sensitivity, specificity and precision. The comparative results show that k-nearest neighbour performs better than naïve bayes and logistic regression techniques. **[5]** Proposed, A RUSMRN algorithm by using machine learning methods. It is based on RUS data sampling

 ISSN: 2454-4248
19 – 23

technique and MRN algorithm for the prediction of the data payment. In order to improve the classification accuracy of unbalance characteristic data author proposed a RUSMRN method that is a combination of the boosting and data sampling. Today's, many enterprises mainly focused on the expenditure services through credit card broadly because it is convenient and quick to pay for products and services. Author emphasized on the fraud detection of credit card payment by using the machine learning technique called RUSMRN. The proposed method adopts three base classifiers which are MLP, NB and Naïve Bayes algorithms. In addition, it can analyse the correctness to work with the unbalance datasets. Author concluded that the proposed method can achieve the best classification performance in terms of accuracy and sensitivity. RUSMAN has highest sensitivity after training and testing by applying the propose method and it is appropriated for predicting the data.**[6]** Proposed to design a high quality software is difficult when the size and complexity is high. Various machine learning algorithms such as random forest, bagging and naïve bayes are used to predict the software quality in early phase on the bases of different metrics like C & K, Henderson & Sellers, McCabe etc. **[7]**, Presented, a genuine informational index from secondary school is utilized for filter desire datasets utilizing WEKA tool. The dataset of understudy scholarly records is tried on different characterization calculations, for example, multilayer recognition, Naïve bayes, SMO, J48 and REPTree utilizing WEKA device which utilized as a part of request to foresee the exactness and execution lattices. **[8]** Examined, A pattern recognization and data mining method is used in predictive modelling in the domain of cardiovascular diagnosis. The analysis carried out using classification algorithm such as naïve bayes, decision tree, knn and neural network and outcome proves that naïve bayes technique outperformed other used technique.**[9]**, Introduced that now a days, an email has become a medium of communication between users. When the volume of an email user is rise in the size has produce a considerable rise in spam mail. Classification approaches are used to classify the email as spam or non spam which generates categorical results. The analysis carried out using different type of decision tree algorithm such as ID3,J48,ADTree,and SimpleCART on the basis of spam email datasets. Author concluded using these classification approaches J48 results a good results rather than other in terms of accuracy using WEKA tool. **[10**], proposed for uses decision trees, naïve bayes, and neural network to predict heart disease with 15 popular attributes as risk factor listed in the medical data analysis when attributes considered to be as "without disease" and number of cases used 146,180 and 178 corresponding .

I.TABLE 1: COMPARISON OF DIFFERENT APPROACHES

Paper Title	Year	Tool used	Techniques used	Accuracy
Prediction analysis of risky credit using Data mining classification models [3]	2017	Python	**Predictive power**	
			Decision tree	5
			Random Forest	1
			Adaptive Boosting	4
			SVM	3
			Linear Regression	2
			Neural Network	6
Credit card fraud detection using machine learning techniques[4]	2017	Python	Naïve Bayes	97.92%
			KNN	97.62%
			Logistic Regression	54.86%
Credit card fraud detection using RUS and MRN algorithms[5]	2016	Matlab	RUSMRN	79.73%
			RUSBoost	77.8%
			AdaBoost	57.73%
			Naïve Bayes	70.13%
An empirical evaluation of classification algorithms for fault prediction in open source projects[6]	2016	PMD, Weka	Random Forest	-
			Bagging	
			Naïve Bayes	
Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector[7]	2015	WEKA	Multilayer Perception	75%
			Naïve Bayes	65.13%
			SMO	68.42%
			J48	69.73%
			REPTree	67.76%
An empirical study on prediction of heart disease using classification data mining techniques[8]	2012	WEKA	Naïve Bayes	83.70%
			Decision Tree	76.66%
			K-NN	75.18%
			Neural Network	78.148%

 ISSN: 2454-4248
19 – 23

A Comparative Study of Classification Algorithms for Spam Email Data Analysis[9]	2011	WEKA	ID3	89.11%
			J48	92.76%
			ADTree	90.91%
			SimpleCART	92.63%
Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques[10]	2010	WEKA	Decision Tree	89%
			Neural Network	83%
			Naïve bayes	83.53%

III. CLASSIFICATION APPROACHES

Classification approaches of data mining is the systematic approach for developing classification models for training and testing datasets is known as classification technique. Various classification techniques like decision tree classifier, rule based classifier, neural network classifier, naïve bayes classifier, neuro fuzzy classifier, support vector machine, regression etc are used for predictive modelling and classification of datasets. Classification is an information mining process used to arrange everything in an arrangement of information into one of predefined set of classes or gatherings[11][12]. For example:- a classification model used to discriminate loan applicants as low, high, medium credit risk. A classification task begin with the dataset in which the class assignments are known[13][14].

In this paper, we illustrate various role of classification methods such as Decision tree, Naive bayes, K-NN, Rule based , ANN.

3.1 Decision tree Classification Approach: Decision tree classifier used to produce classification models in the form of tree hierarchical structure. It splits a huge dataset into smaller subsets simultaneously to develop an associated decision tree[15]. The root node in a decision tree resembles or correlates the best predictor. Decision tree is intended to deal with numerical and unmitigated information. This approach is used to reconstruct the manual classification of the training document created in the form of true and false queries in hierarchy where the nodes represent the query and leaves represent the class of the document[16][17]. The tree design done by the model gives the combined view of the classification of the logic and tree techniques is known as over fitting. However, decision tree algorithm produce the effective performance with large number of records[18][19][20].

3.2 Naïve Bayes Classification Approach:- It is the system which was found by Cooper and Herskovits in the year 1992. In the information mining, these systems are considered as the measurable strategies [5]. At the point when the data is earlier known or uncertain for the approaching information accessible fractional all things considered these systems has been used keeping in mind the end goal to display circumstances. This strategy has been used for the precise expectation of the estimation of an assigned discrete class variable utilizing Baye rules **[21][22]**.

3.3 K-NN (K Nearest Neighbour) Classification Approach: K-NN is the classification of object performed on the basis of closeness of training data available within the feature based is called KNN algorithm. It is lazy learning algorithm known as instance based learning which utilize in order to perform regression. With the given labelled location of training data, the space is divided into regions. If there is most frequent class available amongst the knn to which class the points in space is assigned. If there is numerical values given the Euclidean distance is used to estimate distance metric[23]. KNN classifier are rely upon learning by relationship, it implies by contrasting a given test information and preparing test which is like it. We can portray preparing by n characteristics and each tuple speak to a point in n dimensional space to store all preparation information. When we are with an obscure tuple, classifier fines the example space for the k preparing tuple that are closes to the obscure tuple. The closeness is clarified regarding a separation metric, for example, Euclidean distance[24][25].

3.4 Rule Based Classification Approach: Rule based classifier is a model which is represented as IF-THEN rule format. It consists of two parts:-IF part known as rule antecedent and THEN part is rule consequent. Rules are a decent method for speaking to data or bit of information. In the event that THEN lead is a declaration of the frame IF condition THEN conclusion. The IF – THEN lead is composed from decision tree since rules are anything but difficult to decipher by people. This classifier is surveyed by scope and precision. Manage scope is the level of tuples secured by the govern and its precision is the level of accuracy in classification[25].

3.5 Artificial Neural Network Classification Approach: The artificial neural network is known to be analytical approach which is modelled after the process of learning in cognitive system and it is based on the neurological function of the brain and capable of predicting new observation from existing one after the execution of processed called learning from an existing data[26]. A neural network consists of interconnected processing elements also called units, notes, or neurons. The neurons collaboratively work within a network to produce output function to make the network robust and fault tolerance. Neural network can often produce very accurate prediction.

IV. CONCLUSION

The most valuable favorable position of information mining characterization approach is forecast investigation. In this survey work, it has been presumed that forecast examination is the system which can group the information as per the arrangement of principles. Prescient investigation can be actualized utilizing open source innovations and different

 ISSN: 2454-4248
19 – 23

information mining calculations to process continuous information in blame tolerant way in adaptable and proficient manner.Different sort of grouping calculations are utilized for forecast examination. The root node in a decision tree resembles or correlates the best predictor. Decision tree is intended to deal with numerical and unmitigated information. Naïve bayes classifier is used in less amount of training data to predict the parameters necessary for classification. K-NN is lazy learning algorithm known as instance based learning which utilize in order to perform regression If there is numerical values given the Euclidean distance is used to estimate distance metric. Rule based classifier is a model which is represented as IF-THEN rule format. Rules are a decent method for speaking to data or bit of learning. Neural Network classifier is used for predicting new observation from existing one after the execution of processed called learning from an existing data. Be that as it may, different grouping approaches are utilized for expectation investigation to enhance the exactness of the anticipated model yet all the examination in view of a specific dataset which is utilized to create an anticipated outcomes as far as Accuracy, Precision, Recall, F-measure, G-measure et cetera to discover the beatest order approach for forecast investigation.

Table 2: Comparison of various classification techniques

Techniques	Pros	Cons
Decision Tree	It is used to build an instinctive decision making rules and handle features whose output is not directly proportional to its input or an unpredictable features.	To design a decision tree with so many branches become a main cause of increasing cost and complexity.
Naive Bayes	It is cost effective when datasets are too small and simple to understand or implement. It is used to generate good results in categorical problems than numerical.	It performs badly as an independent predictor because in real life it is hard to find out independent set of predictor.
K-NN	It is simple to use and impose. It is used to handle unstructured data in short time and adaptable to range/feature selection.	It performs well in large dataset so that is why the storage and searching of data to find a nearly neighbour is difficult.
Rule Based	It follows simple IF-THEN rules which can be easily imposable and understandable. It is less costly or complex.	It does not perform well in multi class prediction.
Artificial Neural Network	It deals with complex and hidden problems. It is highly usable with non linear data modelling.	It is time consuming, difficult to understand and difficult to handle training dataset outcomes.

REFERENCES

[1] N. College, “Performance Analysis of Bayes Classification Algorithms in WEKA Tool using Bank Marketing Dataset,” vol. 5, no. 2, pp. 128–133, 2018.

[2] B. Kaur and G. Bathla, “Document Classification using Various Classification Algorithms : A Survey.”

[3] A. Gahlaut and P. K. Singh, “Prediction analysis of risky credit using Data mining classification models,” 2017.

[4] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis,” *2017 Int. Conf. Comput. Netw. Informatics*, pp. 1–9, 2017.

[5] A. Charleonnan, “Credit card fraud detection using RUS and MRN algorithms,” *2016 Manag. Innov. Technol. Int. Conf.*, p. MIT-73-MIT-76, 2016.

[6] A. Kaur and I. Kaur, “An empirical evaluation of classification algorithms for fault prediction in open source projects,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 30, no. 1, pp. 2–17, 2018.

[7] P. Kaur, M. Singh, and G. S. Josan, “Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector,” *Procedia Comput. Sci.*, vol. 57, pp. 500–508, 2015.

[8] T. J. Peter and K. Somasundaram, “An empirical study on prediction of heart disease using classification data mining techniques,” *IEEE Int. Conf. Adv. Egineering, Sci. Manag.*, pp. 514–518, 2012.

[9] A. K. Sharma, “A Comparative Study of Classification Algorithms for Spam Email Data Analysis,” *Int. J. Comput. Sci. Eng.*, vol. 3, no. May, pp. 1890–1895, 2011.

[10] K. Srinivas, G. R. Rao, and A. Govardhan, “Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques,” *2010 5th Int. Conf. Comput. Sci. Educ.*, pp. 1344–1349, 2010.

[11] P. Pahwa, M. Papreja, and R. Miglani, “Performance Analysis of Classification Algorithms,” vol. 3, no. 4, pp. 50–58, 2014.

[12] I. Charalampopoulos and I. Anagnostopoulos, “A comparable study employing weka clustering/classification algorithms for web page classification,” *Proc. - 2011 Panhellenic Conf. Informatics, PCI 2011*, pp. 235–239, 2011.

[13] D. Kabakchieva, “Student performance prediction by using data mining classification algorithms,” *Int. J. Comput. Sci. Manag. Res.*, vol. 1, no. 4, pp. 686–690, 2012.

[14] I. M. M. Mitkees, A. Ibrahim, and B. Elseddawy, “Customer Churn Prediction Model using Data Mining techniques,” pp. 262–268, 2017.

[15] A. B. Raut and A. A. Nichat, “Students Performance Prediction Using Decision Tree Technique,” *Int. J. Comput. Intell. Res. ISSN*, vol. 13, no. 7, pp. 973–1873, 2017.

[16] S. Ravichandran, V. B. Srinivasan, and C. Ramasamy, “Comparative Study on Decision Tree Techniques for Mobile Call Detail Record,” *J. Commun. Comput.*,

 ISSN: 2454-4248
19 – 23

vol. 9, pp. 1331–1335, 2012.

[17] H. D. Masethe and M. A. Masethe, "Prediction of Heart Disease using Classification Algorithms," *Proc. World Congr. Eng. Comput. Sci.*, vol. II, pp. 22–24, 2014.

[18] M. Computing and R. Kaur, "A Review - Heart Disease Forecasting Pattern using Various Data Mining Techniques," *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, no. 6, pp. 350–354, 2016.

[19] E. Suganthi and S. Prakasam, "PREDICTING SIGNIFICANT DATASETS USING DECISION TREE TECHNIQUES FOR SOFTWARE DEFECT ANALYSIS," vol. 5, no. 7, pp. 73–78, 2017.

[20] C. Paper, "Performance Analysis of DTT for Mobile CDR Performance Analysis of DTT for Mobile CDR," no. February, 2016.

[21] S. D. Jadhav and H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques," *Int. J. Sci. Res.*, vol. 14611, no. 1, pp. 2319–7064, 2013.

[22] S. Banu and S. Swamy, "Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics : A Survey," *Int. Conf. Electr. Electron. Commun. Comput. Optim. Tech.*, pp. 256–261, 2016.

[23] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the cognitive skill of students in education environment," *2014 IEEE Int. Conf. Comput. Intell. Comput. Res.*, pp. 1–4, 2014.

[24] M. N. Amin and A. Habib, "Comparison of Different Classification Techniques Using WEKA for Hematological Data," *Am. J. Eng. Res.*, no. 43, pp. 2320–847, 2015.

[25] I. Technology, "International Journal of Research in Computer & Information Technology (IJRCIT) Vol . 1 , Special Issue 1 , 2016 ISSN : 2455-3743 ' A SYSTEMATIC OVERVIEW ON DATA MINING : CONCEPTS AND TECHNIQUES ' International Journal of Research in Computer & Informat," vol. 1, no. 1, pp. 136–139, 2016.

[26] S. Banihashemi, G. Ding, and J. Wang, "Developing a Hybrid Model of Prediction and Classification Algorithms for Building Energy Consumption," *Energy Procedia*, vol. 110, no. December 2016, pp. 371–376, 2017.