COMPREHENSIVE RESEARCH IN EMERGING TECHNOLOGIES
https://choicemade.in/cret/ Volume 2 issue 1

Improved Performance of Machine learning Classifiers with Imbalanced Class Dataset

Deepti Jharbade[1], Vaibhav Patel [2] , Anurag Shrivastava[3]

[1]*M.Tech Scholar, Department of Computer Science, NIRT, Bhopal*
[2]*Assistant professor, Department of Computer Science, NIRT, Bhopal*
[3]*Associate professor, Department of Computer Science, NIRT, Bhopal*

Abstract: With the rapid growth of digital transitions and connecting things with internet huge amount of data collected every day. Classification of this large dataset with high accuracy and low error rate is also an issue. Sampling of data is one the solution of large dataset. An imbalanced class of dataset is another problem for improving classifiers performance. Many researchers have been used machine learning algorithms for improving the classifier performance. This work proposes a sampling technique for obtaining the balancedsampled data. Sampled dataset represent the whole dataset. Imbalanced Classesalso balanced by sampling techniques like oversampling and under-sampling. Majority class under-sampled and minority class oversampled for balancing the imbalanced classes. The main purpose of this work is to propose classification model. This model used the sampling technique for class balancing and machine learning algorithms to improve the classifier performance. The Proposed work is tested on basis of Accuracy and Error rate. Various dataset like KDD Cup 99, Bankruptcy and CDK dataset is used for this approach.

Keywords: Classifier, Machine learning techniques, Sampling, over sampling, under sampling

I. INTRODUCTION

We securing information either in private or government sector has become an essential requirement. System vulnerabilities and valuable information magnetize most attackers' attention. Traditional intrusion detection approaches such as firewalls or encryption are not sufficient to prevent system from all attack types. The number of attacks through network and other medium has increased dramatically in recent years. Efficient intrusion detection is needed as a security layer against these malicious or suspicious and abnormal activities. Thus, intrusion detection system (IDS) has been introduced as a security technique to detect various attacks. IDS can be identified by two techniques, namely misuse detection and anomaly detection. Misuse detection techniques can detect known attacks by examining attack patterns, much like virus detection by an antivirus application. However they cannot detect unknown attacks and need to update their attack pattern signature whenever there is new attacks .On the other hand, anomaly detection identifies any unusual activity pattern which deviates from the normal usage as intrusion. Although anomaly detection has the capability to detect unknown attacks which cannot be addressed by misuse detection, it suffers from high false alarm rate .In recent years, and interest was given into machine learning techniques to overcome the constraint of traditional intrusion techniques by increasing accuracy and detection rates. New machine learning based IDS with sampling is used in our detection approach. The advantage of IDS (Intrusion Detection system) can greatly reduce the time for system administrators/users to analyze large data and protect the system from illicit attacks. Improve the performance of IDS and the low false alarm rate.

A. Machine Learning Technique:

When a computer needs to perform a certain task, a programmer's solution is to write a computer program that performs the task. A computer program is a piece of code that instructs the computer which actions to take in order to perform the task. The field of machine learning is concerned with the higher-level question of how to construct computer programs that automatically learn with experience. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E Thus, machine learning algorithms automatically extract knowledge from machine readable information. In machine learning, computer algorithms (learners) attempt to automatically distill knowledge from example data. This knowledge can be used to make predictions about novel data in the future and to provide insight into the nature of the target concepts applied to the research at hand, this means that a computer would learn to classify alerts into incidents and non-incidents (task T). A possible performance measure (P) for this task would be the Accuracy with which the machine learning program classifies the instances correctly. The training experiences (E) could be labelled instances.

II. RELATED WORK:

The Autors [1] introduced a new method for improving detection rate to classify minority-class network attacks/ intrusions using cluster-based under-sampling with Random Forest classifier. The proposed method is a multi-layer classification approach, which can process the highly imbalanced big data to correctly identify the minority/ rare class-intrusions. Initially, the proposed method classify a data point/ incoming data is attack/ intrusion or not (like normal behavior), if it's an attack then the proposed method try to classify attack type and later sub-attack type. Authors used cluster-based under-sampling technique to deal with class imbalanced problem and popular ensemble classifier Random Forest for addressing overfitting problem. We have used KDD99 intrusion detection benchmark dataset for experimental analysis and tested the performance of proposed method with existing machine learning algorithms like: Artificial Neural Network (ANN), na¨ıve Bayes (NB) classifier, Random Forest, and Bagging techniques.

The Autors [2] use a data-driven approach based on an under sampling technique to evaluate the performance of classifiers in the detection of network intrusion. Authors applied an under-sampling technique in two IDS datasets and evaluated the performance of 5 (five) classifiers in a 5% dataset portion followed by a validation step in a 2% portion of the same dataset without overlaps or repetitions. The results indicate that the use of a stratified train/test into under-sampled datasets show stability in the results in relation to the validation subsampling. The use of this approach allows us to evaluate the classifiers in reduced time, including those considered computationally costly.

The Research [3] proposes an approach through machine learning, specifically the ensemble learning approach and the synthetic minority over-sampling technique (SMOTE) method as a method of detecting intrusions in the IoT system which is expected to produce better performance. The results of this study indicate that the proposed approachiis able to detect intrusion and classify into five types of intrusion including normal intrusion, probe, dos, r2l, u2r. Based on the evaluation results, the proposed approach can improve the performance of intrusion detection in terms of accuracy to 97.02%, detection rate of 97%, false alarm rate 0.16%, compared to base learning and approaches in previous studies used as intrusion detection methods, but in processing time performance have not shown satisfying results.

The Autors [4] propose six machine-learning-based IDSs by using K Nearest Neighbor, Random Forest, Gradient Boosting, Adaboost, Decision Tree, and Linear Discriminant Analysis algorithms. To implement a more realistic IDS, an up-to-date security dataset, CSE-CIC-IDS2018, is used instead of older and mostly worked datasets. The selected dataset is also imbalanced. Therefore, to increase the efciency of the system depending on attack types and to decrease missed intrusions andfalse alarms, the imbalance ratio is reduced by using a synthetic data generation model called Synthetic Minority Oversampling Technique (SMOTE). Data generation is performed for minor classes, and their numbers are increased to the average data size via this

technique. Experimental results demonstrated that the proposed approach considerably increases the detection rate for rarely encountered intrusions.

The authors [5] have proposed to use data mining technique including classification tree and support vector machines for intrusion detection. Utilize data mining for solving the problem of intrusion because of following reasons: It can process large amount of data. User's subjective evolution is not necessary, and it is more suitable to discover the ignored and unknown information. Machine learning based ID3 and C4.5 two common classification tree algorithms used in data mining. Author said C4.5 algorithm is better than SVM in detecting network intrusions and false alarm rate in KDD CUP 99 dataset.

In [6], the author said performance of a Machine Learning algorithm called Decision Tree is evaluated and compared with two other Machine Learning algorithms namely Neural Network and Support Vector Machines which has been conducted by A. The algorithms were tested based on accuracy, detection rate, false alarm rate and accuracy of four categories of attacks. From the experiments conducted, it was found that the Decision tree algorithm outperformed the other two algorithms. Compare the efficiency of Neural Networks, Support Vector Machines and Decision Tree algorithms against KDD-cup dataset.

According to the authors [16] Selection of dataset for training the prediction model, the Machine Learning tool used for prediction and various other factors are essential in building an efficient prediction model. The dataset includes financial ratios as attributes that are derived from the financial statements of various companies. The most influencing ratios that are required for predicting bankruptcy are selected on the basis of the Genetic Algorithm which filters out the most important ones from different existing bankruptcy models. These ratios of different companies are fed as an input to train the model being implemented in R. The prediction algorithm used is Random Forest, which will enable us to differentiate between bankrupt and non-bankrupt companies.

III. DATA SET AND SAMPLING:

A. KDD CUP 99 DATASET: Used in the evaluate machine learning technique. In practice, we recognize that this dataset is decade old and has many criticisms for Current research. But we believe that it is still sufficient for our experiment which aims to reflect the performance of distinct machine learning approaches in a general way and find out relevant issues. In addition, the full KDD99 dataset Contain 4,898,431 records and each record contain 41 features. Due to the computing power, we do not use the full dataset of KDD99 in the experiment but a 10% portion use of it. This 10% KDD99 dataset contains 494,021 records (each with 41 features) and 4 categories of attacks. The details of attack categories and specific types are shown in Table1. According to Table1, there are four attack categories in 10% KDD99 dataset.

B. CDK DATASET: In various research papers authors used the dataset for experiment purpose. CKD dataset from UCI ML repository [8] is used. The dataset includes 400 instances with 24 attributes and a class attribute. The CKD dataset used in this study is taken from the UCI Machine Learning Repository [8]. The data was donated by Soundarapandian et al. and collected for nearly 2-month period. The dataset comprise of 400 samples represented by 11 numeric and 10 nominal attributes and a class descriptor which is also nominal. Out of 400 samples, 250 samples belong to the CKD group, and the other 150 samples belong to the non-CKD group.

C. The dataset was imported from UCI Machine Learning Repository [33]. The dataset consists of 64 calculated ratios which are obtained from the companies' financial annual report, including profit and loss statement and income statement. The target value is categorical with 1 means "bankrupt" and 0 for "non-bankrupt". The data was also collected for surviving companies. The size of the files is different, as well as the percentage of the bankruptcy instances. For Example, year 1 consists of 5910 instances while bankruptcy makes only 6.9% of the data.

D. Sampling:

Data sets contain very large amount of data which is not an easy task for the user to scan the entire data set. The researcher's initial task is to formulate a rational justification for the use of sampling in his research. Sampling has been often suggested as an effective tool to reduce the size of the dataset operated at some cost to accuracy. It is the process of selecting representatives which indicates the complete data set by examining a fraction. Due to sampling we overcome the problems like; i) in research it is not possible to collect and test each and every element from the data base individually; and ii) study of sample rather than the entire dataset is also sometimes likely to produce more reliable results.

Feature selection

Due to the large amount of data flowing over the network real time intrusion detection is almost impossible. Feature selection can reduce the computation time and model complexity. Research on feature selection started in early 60s [9]. Basically feature selection is a technique of selecting a subset of relevant/important features by removing most irrelevant and redundant features [10] from the data for building an effective and efficient learning model [11].

Under sampling:

Under sampling is to select a portion of the majority class to achieve the distribution balance of the two classes. In Random under sampling the majority class is under-sampled by randomly removing samples from the majority class. Over sampling by SMOTE:Oversampling is to sample the minority class over and over to achieve the balanced distribution of the two classes. By applying a combination of under-sampling and over-sampling, the initial bias of the learner towards the negative (majority) class is reversed in the favor of the positive (minority) class. Classifiers are learned on the dataset perturbed by "SMOTING" the minority class and under-sampling the majority class.

III. PROPOSED CLASSIFICATION MODEL

In this research work a cloud based prediction model is proposed, to detect the possibilities of CKD and its progression in patients with some health issues like hypertension and diabetes, is proposed and implemented. The models are trained and tested on the CKD data provided on UCI repository [8] and it is deployed on Microsoft Azure ML platform. The proposed model offered in this work (refer Figure 3.1), actually employs Two class boosted decision tree and Two class deep support vector machine learning algorithm. The model built over Boosted decision tree algorithm has highest prediction accuracy. The model can also be used to test and predict risk on any unknown data. For faster evaluation and lesser overall time cloud platform is used. The dataset requires pre-processing for converting it into a suitable format for getting highly accurate results within smaller time. The pre-processing methods affect a lot in final evaluation results of ML model. It is a good practice to apply such processes on raw data. After applying suitable pre-processing techniques, a method is applied to overcome the missing values. The 'Missing Value Scrubber' is applied to deal with missing values.

In the next step dataset is split into two subsets known as training and testing set. Generally a small part of dataset is chosen to train the classifier/model. The ratio 40: 60 i.e. train: test is used for this work.

To build the model various ML algorithms are applied and tested iteratively in the next step and best model is determined. The ML methods involving mathematical models and statistical analysis like regression analysis or more complex approaches like Decision Trees and Neural Network algorithm to the data are to be applied to fulfil the purpose of Prediction. The best model based on ML methods is preferred by data scientist to decide many aspects to generate more useful results.

In the proposed Prediction model, the Boosted Decision Tree Algorithm along with other Modules is applied for better Prediction accuracy and faster evaluation. Application of Boosted Decision Tree Algorithm provides better data classification better Prediction accuracy than other models like LR. The Prediction classifier (model) is deployed and tested using test set.

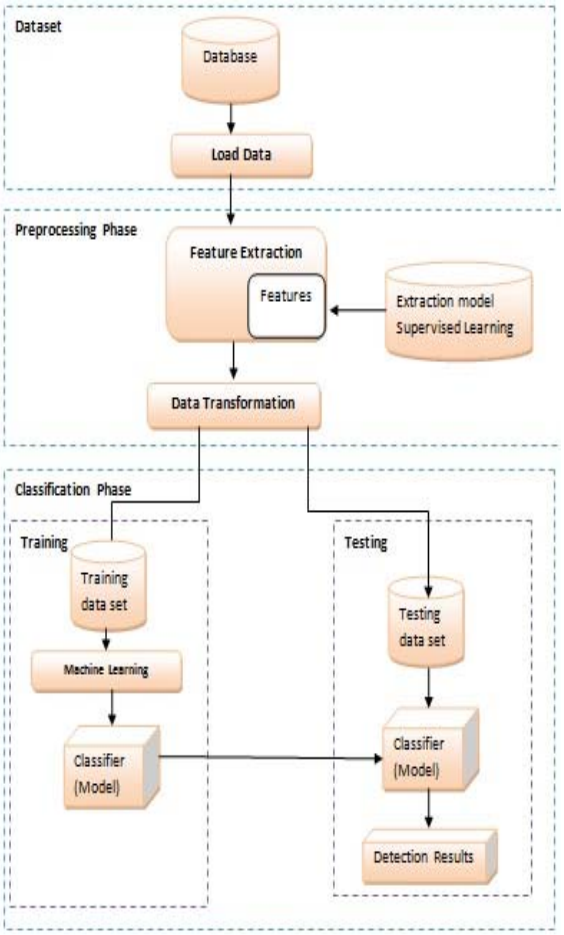

Figure: 3.1 Proposed Classification Model

V. RESULT ANALYSIS

In this research work the evaluation metrics that taken into account are Model's 'Prediction Accuracy' and Precision. The proposed models are achieving good prediction accuracy and precision. Among both the models Boosted decision tree algorithm is performing well.

	CDK Dataset		Bankruptcy		KDD Dataset	
Models	**Prediction Accuracy**	**Precision**	**Prediction Accuracy**	**Precision**	**Prediction Accuracy**	**Precision**
Boosted Decision Tree	*100*	*0.01*	*82.2*	*17.7*	*93.31*	*7.69*
Neural Network	*97.9*	*0.94*	*86.7*	*13.3*	*91.65*	*9.26*

Table 5.1: Results: Prediction Accuracy and Precision

CONCLUSION

In this paper review the machine learning based intrusion detection system which used sampling technique for improving the system performance. Under sampling and oversampling is used for the class balancing purpose. The main purpose for this paper is to find the various sampling and class balancing techniques which are used for improving the machine learning based IDS.

REFERENCES

[1] Md. Ochiuddin Miah, Sakib Shahriar Khan, Swakkhar Shatabda, and Dewan Md. Farid "Improving Detection Accuracy for Imbalanced Network Intrusion Classification using Cluster-based Under-sampling with Random Forests" 978-1-7281-3445-1/19/$31.00 c 2019 IEEE

[2] Bruno Silva, Manuel Silva Neto, Paulo Cortez and Danielo Gomes "Design of Network Intrusion Detection Systems with under-sampled datasets" 978-1-7281-3185-6/19/$31.00 c 2019 IEEE

[3] Aldhi Ari Kurniawan, Heru Agus Santoso "Intrusion Detection System as Audit in IoT Infrastructure using Ensemble Learning and SMOTE Method" 978-1-7281-2380-6/19/$31.00 ©2019 IEEE

[4] GOZDE KARATAS, ONDER DEMIR "Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset" ieee access 2020

[5] YU-XIN MENG "The Practice on Using Machine Learning For Network Anomaly Intrusion Detection" 2011 IEEE

[6] Jiwani, N., & Gupta, K. (2019). Comparison of Various Tools and Techniques used for Project Risk Management. International Journal of Machine Learning for Sustainable Development, 1(1), 51-58. Retrieved from https://ijsdcs.com/index.php/IJMLSD/article/view/119

[7] Chi Cheng, Wee Peng Tay and Guang-Bin Huang "Extreme Learning Machines for Intrusion Detection" - WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia

[8] Naeem Seliya , Taghi M. Khoshgoftaar "Active Learning with Neural Networks for Intrusion Detection" IEEE IRI 2010, August 4-6, 2010, Las Vegas, Nevada, USA 978-1-4244-8099-9/10/$26.00 ©2010 IEEE

[9] Kamarularifin Abd Jalill, Mohamad Noorman Masrek "Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion" 201O International Conference on Networking and Information Technology 978-1-4244-7578-0/$26.00 © 2010 IEEE

[10] Shingo Mabu, Member, IEEE, Ci Chen, Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa, Member, IEEE "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming" IEEE, JANUARY 2011

[11] Liu Hui, CAO Yonghui "Research Intrusion Detection Techniques from the Perspective of Machine Learning" - 2010 Second International Conference on MultiMedia and Information Technology 978-0-7695-4008-5/10 $26.00 © 2010 IEEE

[12] Jingbo Yuan , Haixiao Li, Shunli Ding , Limin Cao “Intrusion Detection Model based on Improved Support Vector Machine” Third International Symposium on Intelligent Information Technology and Security Informatics 978-0-7695-4020-7/10 $26.00 © 2010 IEEE

[13] Maria Muntean, Honoriu Vălean, Liviu Miclea, Arpad Incze “A Novel Intrusion Detection Method Based on Support Vector Machines” IEEE 2010.

[14] Jiwani, N., & Gupta, K. (2019). Comparison of Various Tools and Techniques used for Project Risk Management. International Journal of Machine Learning for Sustainable Development, 1(1), 51-58. Retrieved from https://ijsdcs.com/index.php/IJMLSD/article/view/119 [15] W. Yassin, Z. Muda, M.N. Sulaiman, N.I.Udzir, “Intrusion Detection based on K-Means Clustering and OneR Classification” IEEE 2011.

[15] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey “Intrusion Detection Using Data Mining Techniques” IEEE 2010.

[16] Bilal Khan 1, Rashid Naseem 2, Fazal Muhammad 3,Ghulam Abbas 4, (Senior Member, IEEE), and Sunghwan Kim 5 “An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy” Received March 2, 2020, accepted March 14, 2020, IEEE 2020

[17] Shanmukha Vellamcheti, Pradeep Singh “Class Imbalance Deep Learning for Bankruptcy Prediction” 978-1-7281-4997-4/20/$31.00 ©2020 IEEE

[18] Gupta, Ketan and Jiwani, Nasmin, Prediction of Insulin Level of Diabetes Patient Using Machine Learning Approaches (January 18, 2022). Ketan Gupta,Nasmin Jiwani, 'Prediction of Insulin Level of Diabetes Patient Using Machine Learning Approaches', International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.10, Issue 8, pp.c434-c441, August 2022, Available at :http:/&, Available at SSRN: https://ssrn.com/abstract=4205251

[19] Shuoshuo Fan, Guohua Liu, Zhao Chen “Anomaly detection methods for bankruptcy prediction” The 2017 4th International Conference on Systems and Informatics (ICSAI 2017) IEEE

[20] Shreya Joshi1, Rachana Ramesh 2, Shagufta Tahsildar3 “A Bankruptcy Prediction Model Using Random Forest” (ICICCS 2018) IEEE Xplore