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

Improved Performance of Classifiers Using Sampling Techniques

Ashish Anand[1], Anurag Shrivastava[2], Vaibhav Patel [3]

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

Abstract- The development and popularization of information system, classification of various dataset is become important task now days. Data Mining and Machine Learning techniques proved useful attention in security research area. Recently, many machine learning methods have also been applied by researchers, to obtain highly accuracy. The problem of all those methods is that how to classify classes effectively. Looking at such inadequacies, the machine learning technique is used for obtain the high accuracy. Also, use of internet is increasing progressively, so that large amount of data and it size is also an issue. Sampling technique is one the solution of large dataset. This work proposes a sampling technique for obtaining the sampled data. Sampled dataset represent the whole dataset with proper class balancing. Class imbalanced can be balanced by sampling techniques. In this paper propose a classification framework model based on proposed sampling, class balancing and machine learning technique. This model improves the classification performance. The Proposed work is tested on basis of Accuracy and Error rate with KDD Cup 99 and bankruptcy dataset.

Keywords—Class Balancing, Sampling, Classification, Machine learning technique

1. INTRODUCTION

Information security 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.

Data Mining

Data Mining is defined as the technique of extracting information or knowledge from huge amount of data. In other words, we can say that data mining is mining knowledge from large data.

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 labeled instances.

2. 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 approach is 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.

In [5] authors said Data sets contain very large amount of data which is not an easy task for the user to scan the entire data set. 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. This paper focuses on different types of sampling strategies applied on neural network. Here sampling technique has been applied on two real, integers and categorical dataset such as yeast and hepatitis data set prior to classification. Authors give the comparison of different sampling strategies for classification which gives more accuracy.

The work [7] discusses imbalanced dataset. A dataset is imbalanced if the classification categories are not approximately equally represented. Authors discuss some of the sampling techniques used for balancing the datasets, and the performance measures more appropriate for mining imbalanced datasets. Over and under-sampling methodologies have received significant attention to counter the effect of imbalanced data sets. Sampling methods are very popular in balancing the class distribution before learning a classifier.

3. DATA SET AND SAMPLING:

KDD Cup 99 dataset: KDD99 dataset Contain 4,898,431 records and each record contain 41 features. Due to the computing power, researchers 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. This dataset is used for the cyber security and intrusion detection system.

For experimentation in stock market analysis the NSE datasets is used. In this data two popular financial organizations (companies) are captured from the NSE website. The companies are Reliance and HDFC bank. The data set encompassed the trading days of nearly 20 years i.e. from 03rdJan 2003 to 30th May 2020. The snapshot of the first five rows of the HDFC and Reliance dataset is shown in Figures 4.1 and 4.2 respectively. The dataset comprises timestamps and features that have numerical values.

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.

Class Imbalanced:

A Dataset is imbalanced if the Classification categories are not just about equally represented. Over and under-sampling methodologies have received attention to counter the effect of imbalanced data sets[10].

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 Population until the majority class becomes up to minority class or other class.

Over sampling:

Oversampling is to sample the minority class over and over to achieve the balanced distribution of the two classes.

4. EXPERIMENTAL SETUP

Weka 3.6.11 is used as DM tool for simulation purpose. Weka is installed over Windows 7 Operating System. Two different dataset is used for this experiment purpose. 10% portion of KDD99 dataset and another dataset consists 1000 Polish companies. 19.4% companies went bankrupt during 2000-2012. Both dataset is obtained from UCI Machine Learning Repository. Firstly, we are applying sampling and class balancing technique and get balanced sampled dataset now we are using preprocessing technique in sampled and balanced dataset and applying feature selection method. Now going to classification part and determine the training and testing data in very short period after that applying classification technique in trained data and evaluate the result. Same procedure is applying in different machine learning classifier using both dataset and measure result. Also measure the different classifier performance with un-sampled and imbalanced dataset with both datasets. Accuracy and error rate two parameter is used in these experiments.

5. EXPERIMENTAL RESULT

- Balanced sampled KDD'99 and bankruptcy dataset, obtain from sampling technique. Result shows the performance of the proposed approach classifier in terms of accuracy, time taken to build model and error rate on sampled KDD Cup 99 and bankruptcy dataset. Result also shows comparison of performance of the un-sampled, imbalanced dataset in terms of the same parameter.

Decision Tree (J48) Classifier				Naïve Bayes Classifier			
Dataset KDD Cup 99	Accuracy	Error Rate	Time Taken to Build Model	Dataset Banckruptcy	Accuracy	Error Rate	Time Taken to Build Model
Balanced sampled Dataset	99.75	0.24	**7.33 Second**	Balanced sampled Dataset	92.4704	08..4701	**1.21 Second**
Imbalanced un-sampled Dataset	**99.6**	**0.39**	**11.93 Second**	Imbalanced un-sampled Dataset	**90.564**	**10.12**	**3.5 Second**

6. CONCLUSION

In this research paper, Machine Learning technique have been proposed in terms of accuracy, error rate, and tine taken to build model for KDD Cup 99 and bankruptcy dataset. The purpose of this proposed method efficiently classify abnormal and normal data by using very large data set and classify large datasets with short training and testing times. Most importantly when using this method redundant information, complexity with abnormal behaviors are reduced. With proposed method we get high accuracy error rate. The proposed method results compare with imbalanced un-sampled dataset. Experimental results and analysis shows that the proposed system gives better performance in terms of high accuracy, low error rate.

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]. Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey "Intrusion Detection Using Data Mining Techniques", 978-1-4244-5651-2/10/$26.00 ©2010 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]. YU-XIN MENG," The Practice on Using Machine Learning For Network Anomaly Intrusion Detection" Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, 978-1-4577-0308-9/11/$26.00 ©2011 IEEE

[8]. 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

[9]. 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

[10]. 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

[11]. Devendra kailashiya, Dr. R.C. Jain "Improve Intrusion Detection Using Decision Tree with Sampling" Vol 3 (3), 1209-1216 ijcta 2012

[12]. Megha Aggarwal, Amrita "Performance Analysis Of Different Feature Selection Methods In Intrusion Detection" INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 2, ISSUE 6, JUNE 2013.

[13]. Gupta, K., & Jiwani, N. (2021). A systematic Overview of Fundamentals and Methods of Business Intelligence. International Journal of Sustainable Development in Computing Science, 3(3), 31-46. Retrieved from https://www.ijsdcs.com/index.php/ijsdcs/article/view/118-

[14]. Kaberi Das, Prem Pujari Pati, Debahuti Mishra, Lipismita Panigrahi "Empirical Comparison of Sampling Strategies for Classification" ICMOC-2012, Elsevier science direct.

[15]. Ligang Zhou," Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods." Contents lists available at SciVerse ScienceDirect Knowledge-Based Systems journal homepage: www.elsevier.com/locate/knosys online 3 January 2013

[16]. Nitesh V. Chawla "Data mining for imbalanced datasets: an overview" sprimger.

[17]. KDD CUP 1999. Availabe on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html October 2007

[18]. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", 2009 IEEE

[19]. Arun K Pujari "Data mining techniques" Universities Press.

[20]. Subaira A. S., Anitha P. "An Efficient Classification Mechanism For Network Intrusion Detection System Based on Data Mining Techniques:A Survey" International Journal of Computer Science and Business Informatics 2013.

[21]. Sebastiaan Tesink," Improving Intrusion Detection Systems through Machine Learning"

[22]. Weka, University of Waikato, Hamilton, New Zealand.

[23]. 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

[24]. Kazi Abu Taher, Billal Mohammed Yasin Jisan, Md. Mahbubur Rahman" Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection" 978-1-5386-8014-8/19/$31.00 ©2019 IEEE

[25]. Anish Halimaa A, Dr. K.Sundarakantham "machine learning based intrusion detection system" 9785386-9439-8/19/$31.00 ©2019 IEEE

[26]. Anushka Srivastava, Avishka Agarwal, Gagandeep Kaur "Novel Machine Learning Technique for Intrusion Detection in Recent Network-based Attacks" 978-1-7281-3651-6/19/$31.00 ©2019 IEEE

[27]. Koushal Kumar, Jaspreet Singh Batth Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms International Journal of Computer Applications (0975 – 8887) Volume 150 – No.12, September 2016