iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** X **Month of publication:** October 2019

DOI: http://doi.org/10.22214/ijraset.2019.10044

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue X, Oct 2019- Available at www.ijraset.com

An Approach to Improve the Detection rate using Sampling of Imbalanced Data

Priyanka Tripathi[1], Rajni Ranjan Singh Makwana[2]
[1, 2]Department of computer science and Information Technology, MITS Gwalior (M.P.), 474005, India

Abstract: Synthetic Minority Over-sampling Technique (SMOTE) works by creating synthetic observations based upon the existing minority observations. In this research KDD Cup99 dataset is used. Through SMOTE we tried to increase the rare classes (U2R and R2L).The random forest was used to create the model in the Cost Sensitive Classifier. The tests were performed on many percentage ratios of rare classes. Results were better than the existing one.
Keywords: Data Mining, Intrusion Detection System (IDS), WEKA, SMOTE, Imbalanced Data, KDD Cup 1999.

I. INTRODUCTION

When data is collected from network of Intrusion Detection System it provides data with highly imbalance distribution of classes. To remove this problem of imbalance distribution we need to perform under sampling and oversampling of data. This kind of distribution causes mainly two types of classes majority and minority. Under sampling of majority classes is done so as to remove redundancy, duplicity of instances while oversampling is done so as to increase number of instances in minority classes or rare classes.

Figure 1:IDS

In section 2 KDD Cup 1999 refers related functions on the dataset and class imbalance. In Section 3, we recommend using a sample SMOTE ratio to create a numerical model and a new method. In Section 4, we discuss our experimental environment, processes and results. Last is conclusion of the paper in the Section 5.

II. LITERATURE SURVEY

Al ebachew Chiche and Million Meshesha (2017) proposed an intelligent intrusion detection system which can predict attacks in the network and suggest the proper corrective actions for predicted attacks. The system is developed by integrating data mining model and knowledge based system for detecting intrusion types. A model is constructed to predict the intrusion detection is proposed that uses four classifiers MLP, Naive Bayes, Decision tree using J48 and JRip algorithm using rule induction. Dataset used are samples from MIT Lincoln laboratory. Further, the knowledge for prevention techniques is acquired from domain experts and document analysis. The proposed system achieves 91.34 and 85 percent on system performance testing and user acceptance testing respectively. The result is promising to design an intelligent NIDP system by integrating data mining with knowledge based system. Evaluation results show that the proposed system registers 91.43% accuracy in network intrusion detection and 85% accuracy in user acceptance testing. This indicates proposed system performance is promising for plan intelligent network IDS that can effectively predict and provide a prevention mechanism.

Bing Hao Yan et al. (2017) to settle data imbalanced attributes in interruption recognition as of information point & afterward newer district versatile SMOTE calculation has projected to an answer. In the meantime, consecutive backward selecting method was utilized for accelerate recognition procedure by evacuating unnecessary features. Exploratory outcomes demonstrated RA-SMOTE calculation could adequately enhanced rare sample recognition rate, for example u2l & r2l using NSL-KDD dataset also outflanks additional ID techniques. It has as well revealed RA-SMOT algorithm receives greatest performance in compare to past algorithm to deal among the unbalanced setback [7].

III.PROPOSED METHODOLOGY

A. Feature Selection

In Feature Selection we select only relevant attributes and discard unwanted attributes from the data set. There are three types filter approach in it, wrapper approach and embedded approach. In filter approach, it selects features regardless of the model. Wrapper method evaluates subsets to detect the possible interactions between variables.

B. Random Forest

A random forest is a classifier consists of a collection of tree structured classifiers {h(x,Ok),k=1,….} where the {Ok} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

RF is a predictor that includes set of random base regression trees. $\{r_n(x,\Phi_m,D_n), m \geq 1\}$, here $\Phi_1,\Phi_2,...$ defined as randomized factor (Φ) outcomes. Sum of regression estimation is calculated by joining of the random trees

$\bar{r}_n(X,D_n) = E_\Phi\,[r_n(X,\Phi,D_n)]$,

Where, E_Φ representing the expectation for random factor, conditionally probable on the X & dataset D_n.[16]

C. Proposed Algorithm

Step:1 Input original dataset.
Step:2 Separate classes (dos,normal,probe,u2r,r2l) of data
Step:3 This will remove one class, similarly remove four classes and save one. Repeat till all the classes are separated.
Step:4 Remove redundancy from DOS and Normal class.
Step:5 Then combine all data of files.
Step:6 Open combine File.
Step:7 Choose attributes as in previous paper.
Step:8 Discretize the data
Step:9 Now apply Smote for 50% ,100%,…1000% of rare classes R2L and U2R.
Step:10 Repeat step 9.
Step:11 Apply cost sensitive classifier and classify using Random forest.
Step:12 Classified instances.
Step:13 Results.

IV.RESULT ANAYSIS

WEKA tool has been used in this research. The dataset used in this experiment is KDDCup 1999.
The DoS, Normal, Probe, R2L and U2R are four types of attacks categorized from dataset.

Table 2 Comparison between initial set and under sampled set

CLASSES	TRAINED SET (TRNS)	UNDER-SAMPLING TRAINED SET(TRNS_US)
Normal	97,276	87,830
Probe	4,107	4107
U2r	52	75
Dos	391,458	54,570
R2l	1,126	1681
Sum	494,020	148,277

In Table 2 initial number of instances were 494,020 but after under sampling the number of instances reduced to 148,277 as redundant data is removed through under sampling.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue X, Oct 2019- Available at www.ijraset.com

Table 3. Explanation attribute set

Attributes	Explanation
Period	time taken to connection (sec.)
Services	At receiver end defined n/w service types
root_shel	Achieved root shell otherwise else
FLAG	connection situation (it is normal or there is any error)
SRC_BYTES	Tot up data bytes which are sends from sender to the receiver
numb_files_creation	entire creating file operation
LOGGED_IN	successfully login or any others
NUM_FAILED_LOGINS	Entire attempt to login in failure
dest_host_reror_rates	Connection rates including ``REJ" errors
dest_host_dif_srs_rates	For different types services for connections rate

In Table 3 the list of selected attributes is obtained from huge amount of data from dataset.

Table 4. Detection rates (in %) of different ratios of rare classes

normal	probe	u2r	Dos	r2l	Observations
10.0	97.2	86.2	10.0	97.6	RC +50%
98.1	97.3	90.3	10.0	98.1	RC +100%
98.2	97.6	92.1	10.0	98.2	RC +150%
10.0	97.1	93.9	10.0	98.4	RC +200%
10.0	98.5	96.9	10.0	99.6	RC +400%
98.9	97.3	96.8	10.0	98.7	RC +600%
98.9	97.3	97.3	10.0	98.8	RC +800%
98.9	97.4	97.6	10.0	98.9	RC +1000%

In Table 5 RC is the rare classes and the result shows the detection rates of each class with various smote ratios.

Table 6. Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.001	0.999	1.000	0.999	0.999	1.000	0.999	DoS
0.984	0.000	0.998	0.984	0.991	0.991	1.000	0.991	Probe
0.983	0.000	0.983	0.983	0.983	0.985	1.000	0.981	U2R
0.999	0.001	0.999	0.999	0.999	0.999	1.000	1.000	Normal
0.998	0.000	0.998	0.998	0.998	0.998	1.000	1.000	R2L
0.999	0.001	0.999	0.999	0.999	0.998	1.000	0.999	Weighted Avg.

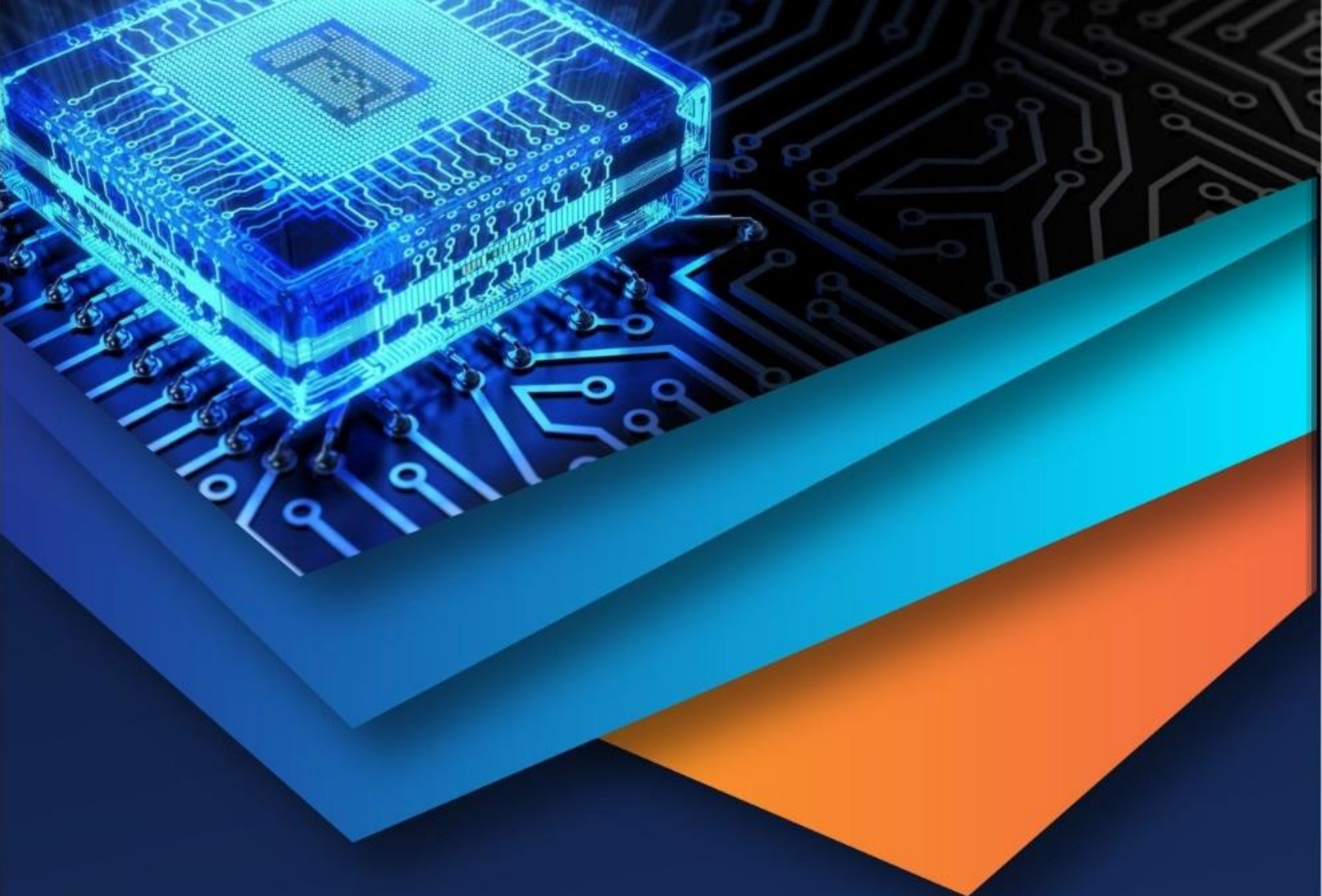
V. CONCLUSION

IDS network provides us dataset with imbalanced rare classes. Random Forest used in this research gives better result than ID3 as used in previous research. Cost sensitive classifier is also used in this research.

Various SMOTE ratios applied on rare classes (remote to local and user to root). It increases detection rate and lowers false negative rate.

REFFERENCES

[1] Shamir Hafez Amur & J Hamilton. 2010. Intrusion detection systems (IDS) taxonomy-a short review. Defense Cyber Security 13, 2 (2010), 23–30.
[2] Glozed Karats, Oder Demur, Ogre Kory Sahingoz, “Deep Learning in Intrusion Detection Systems” , International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism ,978-1-7281-0472-0/18/$31.00 ©2018 IEEE.
[3] Vokorokos, L. and A. Ballad. Host-based IDS in Intelligent Engineering Systems (INES), 2010 14th International Conference on. 2010. IEEE.
[4] Jays Surinam, Jagatai Sharma, Ischia Seraph, Nicoma Pori, Brava Navix, “A Survey On Intrusion Detection System”, International Journal of Engineering Development and Research, © 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939.
[5] Azhagusundari, and A.S. Thanamani, “Feature Selection based on Information Gain”, International Journal of Innovative Technologyand Exploring Engineering (IJITEE), Jan. 2013, pp 18- 21.[2]
[6] Breiman, “Random Forests”, Statistics Department University ofCalifornia, Berkeley, 2001.[3]
[7] Bridges and R. B. Vaughn, “Fuzzy data mining and Genetic algorithms applied to Intrusion detection”, Proc. 23rd
[8] National Information Systems Security Conference, Baltimore, MD, USA,2000.[4]
[9] Chawla, K. W. Bowyer, L. O. Hall, and W. P Kegelmeyer, “SMOTE : Synthetic Minority Over-sampling Technique”, Journal of Artificial Intelligence Research, Vol. 16, 2002, pp. 321–357.[5]
[10] Chen, A. Liaw, and L. Breiman, “Using Random Forest to Learn Imbalanced Data”, University of California at Berkeley, Berkeley, California, 2004



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY