IRAQI Academic Scientific Journals

Samarra Journal of Pure and Applied Science

SJPAS

www.sjpas.com

ISSN:2663-7405

Toward Constructing a Balanced Intrusion Detection Dataset Based on CICIDS2017

Amer Abdulmajeed Abdulrahman [1*], Mahmood Khalel Ibrahem [2]

1- Informatics Institute for Post Graduate Studies (amer6567@gmail.com)

2- Al-Nahrain University-College of Information Engineering (mahmood_khalel@yahoo.com)

Article Information

Received: 14/05/2020

Accepted: 23/07/2020

Keywords:

Imbalanced dataset Classification, SMOTE, CICIDS2017 dataset, Random Forest, Naïve Bayesian, Multilayer Perceptron

Abstract

Several Intrusion Detection Systems (IDS) have been proposed in the current decade. Most datasets which associate with intrusion detection dataset suffer from an imbalance class problem. This problem limits the performance of classifier for minority classes. This paper has presented a novel class imbalance processing technology for large scale multiclass dataset, referred to as BMCD. Our algorithm is based on adapting the Synthetic Minority Over-Sampling Technique (SMOTE) with multiclass dataset to improve the detection rate of minority classes while ensuring efficiency. In this work we have been combined five individual CICIDS2017 dataset to create one multiclass dataset which contains several types of attacks. To prove the efficiency of our algorithm, several machine learning algorithms have been applied on combined dataset with and without using BMCD algorithm. The experimental results have concluded that BMCD provides an effective solution to imbalanced intrusion detection and outperforms the state-of-the-art intrusion detection methods.

1. Introduction.

Intrusion detection system (IDS) has played a pivotal role in defending the network by directing security officials to warn them about malignant behaviors such as attacks, malware, and intrusions. The presence of IDS is a compulsory line of defense to protect vital networks from these ever-increasing issues of intrusive activities. Therefore, research in the field of IDS has flourished over the years to suggest better IDS systems. However, many researchers are struggling to find valid and comprehensive datasets that able testing and evaluating their proposals, the major challenge in itself is having an appropriate dataset [1]. There are many available dataset which are used in IDS such as AWID-2015, Booters-2013, Botnet-2010\2014, CICDoS-2012\2017, CICIDS2017, CTU-13, DARPA-1998, DDoS2016, ISCX2012, ISOT-2010, KDD CUP 99-1998, Kyoto 2006+, LBNL-2004, NDSec-1-2016, NGIDS-DS-2016 and NSL-KDD-1998 etc. [2].

CICIDS2017 is a state of art dataset presented by Canadian Institute of Cybersecurity that contains the latest attacks and features [3]. This dataset draws attention of many researchers as it represents threats which were not addressed by the older datasets. While undertaking an experimental research on CICIDS2017, it has been found that the dataset has few major shortcomings. One of these issues is imbalanced dataset problem.

Imbalanced datasets are spread across many areas and sectors, such as financial services fraud to non-performing loans. The challenge arises when machine learning algorithms attempt to identify these rare states in fairly large datasets. The disparity of class's variables causes that an algorithm tending to classify into the class with more instances of majority class while at the simultaneously giving a false sense that the model is of high precision. The unpredictability of rare events, the misleading accuracy and the minority class detracts from predictive models that have been built. Figure 1 illustrates example of balanced and imbalanced datasets.

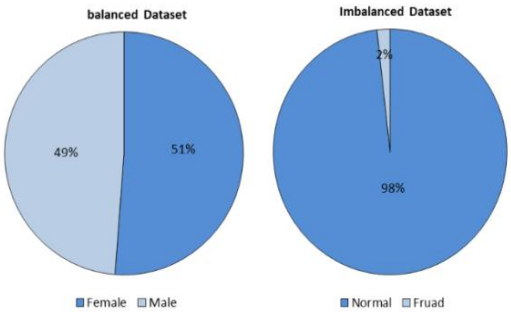

Figure 1: Balanced and Imbalanced datasets

The problem of class imbalance between the minority and the majority can be frustratingly, but expected. This problem has been comprehensively researched in literature. It occurs when the occurrences of one class outnumbers the occurrences of other classes. There are many methods have been developed which can be implemented during the preprocessing stage to overcome these challenges. Resampling is one commonly used strategy which it works by altering the distribution of training examples. This strategy includes under-sampling, oversampling and SMOTE techniques. Under-sampling can be achieved by removing the instance from the overrepresented to balance the dataset. Oversampling can be achieved by adding similar instances of minority class to balance the deflection class ratio. SMOTE technique uses a distance measure for some of selected observations to create a synthetically new instance with the same properties as the available features. SMOTE analyzes one feature at a time by taking the difference between an instance and its closest neighbor. It multiplies the difference with a random number between zero and one then, it identifies a new point by adding the random number to the feature. This way creates a new synthetic instance instead copy observations. Resampling could be done with or without replacement. [4, 5]

The contributions in this paper are twofold. Firstly, analyzes the CICIDS2017 dataset to reduce high class imbalance problem and collected these datasets in one dataset to get a dataset

which contain several types of attacks. Secondly, execute several common machine learning algorithms to evaluate this dataset.

2. Related Works.

A class imbalance processing technology which combines SMOTE and under sampling for clustering based on Gaussian Mixture Model (GMM) to improve the detection rate of minority classes. The proposed technique was implemented on RF and MLP classification algorithms using the UNSW-NB15 and CICIDS2017 datasets. The experimental results showed that for multiclass classification on the CICIDS2017 dataset, the detection rate of RF, MLP and CNN achieved 93.08%, 99.64% and 99.85% respectively [6] .

Used SMOTE to enhance the sensitivity of arrangement for minority classes in individual file of CICIDS2017 dataset which contains DDoS attack. The researchers increased number of minority class instances (DDoS) in training data from 29285 to 87855 by implementing SMOTE with a minority oversampling class of 200%. The performance metrics of training data were calculated by using AdaBoost classifier with several feature selection techniques such as PCA and EFS. The results proved that their proposed method outperforms the performance conducted by previous literature with the accuracy of 81.83%, precision of 81.83%, recall of 1, and F1 Score of 90.01% [7].

Developed a uniform distribution based balancing (UDBB) approach to handle the imbalanced distribution of the minority class instances in the CICIDS2017 dataset. The researchers merged all data files of Monday CICIDS2017 dataset together into a single combined file. Their study compares between the imbalanced case (with original distribution of CICIDS2017) and balanced class distribution (after applying the uniform distribution-based balancing approach). The performance metrics of training data were calculated by using Random Forest, Bayesian Network, LDA and QDA classifiers with 10 features. After applying UDBB, the accuracy of RF, NB, LDA and QDA achieved 98.8%, 97.6%, 95.7% and 98.9% respectively. Also the F measure of RF, NB, LDA and QDA achieved 98.8%, 97.7%, 95.7% and 99.0% respectively[8].

Resolved class imbalanced in Customer Churn dataset by utilizing three resampling techniques random oversampling, under sampling and SMOTE. These techniques have been executed and evaluated on random forest classification model. The results illustrate that random oversampling is a better to balanced dataset [9].

Also resolved the unbalanced class problem to predict atrial fibrillation in the obese patient by utilize the SMOTE. This technique has been performed and evaluated on Logit Boost and GLMBoost ensemble classification methods. The results of prediction illustrated that performance metrics have higher and accurate values with LogitBoost on the balanced dataset by SMOTE. [10]

3. CICIDS2017 Dataset Description.

CICIDS2017 dataset is generated by Canadian Institute for Cybersecurity. Each dataset contains benign and the most up-to-date common attacks such as DoS, DDoS, brute force SSH, brute force FTP, heartbleed, infiltration and botnet which make it the most up-to-date, compared to other dataset. They also include analyzing network traffic results using CICFlowMeter with

labeled flows based on the IP source and IP destination, source and port destination port, time stamp, protocols and attacks [11].

CICIDS2017 dataset is designed for intrusion detection and network security purposes. There are several attack profiles created based on the latest updated list of common attack families and implemented with related tools and codes.

The main types of attack profiles are:

• Distributed Denial of Service DDoS Attack: This usually occurs over victim resources, multiple systems or bandwidth overwhelms. This attack is often the result of multiple hacked systems (a botnet an example) flooding the target system by generating the massive network traffic [12].

• Port Scan attack: this attack sends client requests to a set of server port addresses on a host, with the goal of finding an active port and exploiting the known security sensitivity of that service. Surveying, as a way to discover exploitable communication channels, has existed throughout the ages. The idea is to check as many listeners as possible, and track down recipients or beneficiaries for your own need [13].

• Botnet: Number of Internet connected devices used by the owner of robots to perform different tasks. It can be used to send spam, steal data and allow an attacker to access and connect to the device [14].

• Web Attack: These types of attacks come out every day, because individuals and organizations are taking security seriously now. We use the SQL Injection, which an attacker can create a series of SQL commands, and then use it to force the database to respond the information, Cross-Site Scripting (XSS) that occurs when developers do not properly test their code to find the ability to inject script, and Brute Force over HTTP which can try the list of passwords to find the administrator password [3].

• Infiltration Attack: Internal network infiltration often exploits vulnerable software such as Adobe Acrobat Reader. After successful exploitation, the tailgate will be executed victim's computer and can perform various attacks on the victim's network such as full port scanning, IP sweep and number service using Nmap [3].

Unlike other IDS datasets that separate training from testing data, CICIDS2017 gathered all labelled records of each specified type attack into a unique CSV file format. Each CSV file is composed of a given number of labelled records, and 79 features that describe these records.

4. Experimental Results.

Similar to all datasets, CICIDS2017 dataset contain unwanted elements (missing, redundant or infinite values) that should be removed or transformed. It has been necessary to clean up this dataset from errors which could occur while flow data are been acquiring. The following steps include preprocessing work:

• First step redundant records and redundant attribute have been dropped from the whole dataset. Attribute (Fwd_Header_Length) appeared twice in the list of attributes. This attribute has been removed. All missing values have been replaced by zeros and infinite values have been replaced by the mean of their attribute value. CICIDS2017 dataset contain features that have

been recorded while acquiring data flow, those features are related to a specific network and don't have any impact on model results.

• Second step of preprocessing is removing features with low standard deviation [14]. Standard deviation denoted by sigma (σ) is the average of the squared root differences from the mean. As shown in the following formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(xi - \mu)} \quad (1)$$

Where, xi is an individual value, μ is the mean expected value and N is the total number of values. In this work, standard deviation criteria is used to remove all features with standard deviation value equal to zero, since removing those increases the model's accuracy. Also, those features are irrelevant in data and can decrease the performance of the model analysis. By applying the standard deviation criteria, twelve features have been eliminated from datasets which are: "Bwd PSH Flags, Fwd URG Flags, Bwd URG Flags, FIN Flag count, RST Flag count, CWE Flag Count, Avg Bytes/Bulk, Fwd Avg Packets/Bulk, Fwd Avg Bulk Rate, Bwd Avg Bytes/Bulk, Bwd Avg Packets/Bulk and Bwd Avg Bulk Rate". The rest number of features from above steps will be used in our experimental.

In this work all five individual files of CICIDS2017 dataset are combined to generate other intrusion detection dataset for experiment with multiclass classification. This dataset contains multiple types of attacks. For multiclass classification we either left the labels as they were (textual) or converted them into one-hot encoding depending on the framework requirements. Table (1) shows the instances distribution of multiclass dataset for each label after preprocessing this dataset.

Table 1: Instances distribution of combined dataset for each label.

Benign	DDoS	PortSacn	Botnet	Infiltration	Web
795725	128016	90819	1953	36	2143

In table (1), it can be seen that the prevalence of majority class (Benign) is 78.11% while for the all rest minority class are 12.56% (DDoS), 8.91% (PortSacn), 0.19% (Botnet), 0.0035% (Infiltration) and 0.21% (Web attack). In such a large difference in prevalence, the potential reagent may tend to be benign. This situation causes high-grade imbalance when the dataset is used to train of a classification or detection.

The DMwR package of R has been used to carry out SMOTE. It has used to balance binary dataset. Our proposed Balanced Multiclass Dataset algorithm (BMCD) has been used with SMOTE to solve the problem of imbalanced multiclass dataset.

Input: ID = Imbalanced Multiclass Dataset
Output: BD = Balanced Multiclass Dataset
m is the class number of ID
Pi is the probability of each class, i..m

Split ID to dateframe Ci..m such as Ci is dataframe of each class
D = Sort(Ci) base on Pi
MJ is majority class, MN is minority class
K= nearest neighbor
MJ = D1
For i .. m-1
Tem = MJ + Di+1
P = Pi / 2 - 1 { P is percentage over}
B = SMOTE (Tem, (P x 100) ,K)
Split B to MJ and MN
Append (MJ) to Newdataframe
End for
BD = Newdataframe

Figure 2 shows the class distribution before and after using proposed algorithm with SMOTE for combined dataset.

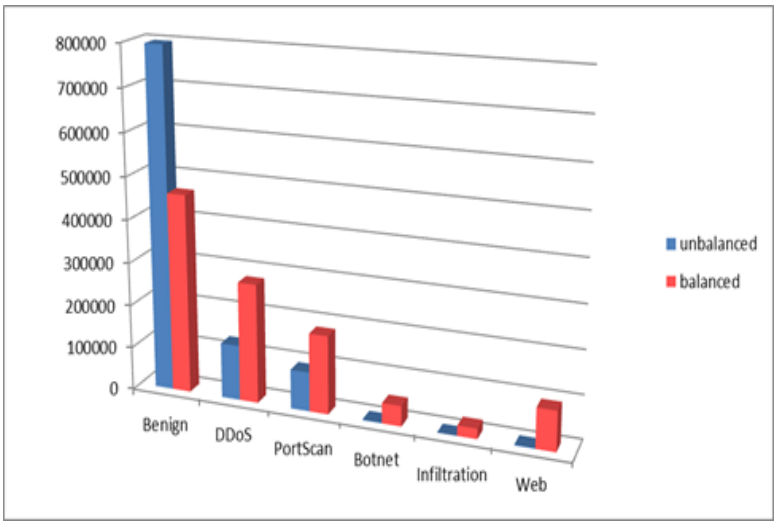

Fig 2. Graphical representation of class distribution when using proposed algorithm

In this experimental, 10-fold of validation technique has been used to evaluate models and obtain unbiased results from models. The three selected common classification algorithms, namely Naïve Bayes (NB), Random Forest (RF) and Multilayer Perceptron (MLP) were used to compare the performance of classifying balanced and imbalanced combined dataset. The Caret package of R has employed to implement all modelling and validation processes.

The common performance metrics Accuracy (Acc), Kappa, Area Under Curve (AUC) of Receiver Operating Characteristic (ROC), Precision, Recall and F-Measure have been used to evaluate selected models. The following equations are used to calculate these metrics. [15]

$$Acc = (TP+ TN)\ (TP + FP + TN + FN) \quad (2)$$

$$Kappa= (Po – Pe) \ (1 – Pe) \quad (3)$$

$$Precision=TP/ (FP+TP) \quad (4)$$

$$Recall=TP/(FN+TP) \quad (5)$$

$$F\text{-}Measure = (2 * Precision * Recall) / (Precision + Recall) \quad (6)$$

Where: The correct classifications are representing as True Positive (TP) and True Negative (TN). False positive (*FP*) represents the incorrectly predicted result as positive but it is negative actually. False negative (*FN*) represents the incorrectly predicted result as negative but it is positive actually. *Pe* is expected agreement and *Po* is observed agreement. All above values are derived from the confusion matrices.

Table 2 illiterates the performance metric results of the classification models of Naïve Bayes, Random Forest and Multilayer Perceptron which it applied on the balanced and imbalanced combined dataset. The performance metrics are calculated for accuracy, kappa, F-Measure and AUC. According to this table, all performance metrics results of balanced combined dataset were outperformed the results of imbalanced dataset at all classification models. These findings are clearly.

Table 2: The Performance examination results of the balanced and imbalanced combined dataset for each classification model.

	Imbalanced				Balanced			
	Acc	*Kappa*	*F-Measure*	*AUC*	*Acc*	*Kappa*	*F-Measure*	*AUC*
RF	0.9678	0.8407	0.8223	0.8943	**0.9932**	**0.9896**	**0.9836**	**0.9827**
MLP	0.8981	0.4988	0.2048	0.5879	**0.9482**	**0.9198**	**0.8063**	**0.9447**
NB	0.5866	0.33	0.8809	0.9307	**0.7535**	**0.6683**	**0.9082**	**0.9410**

In our work, ROC curves analysis has been used in order to assess the accuracy of classifier independent of any threshold. The measure of quality of a probabilistic classifier can be provided by the area (AUC) under ROC curve. The generated results are the value of area under the ROC curve which is useful in determining the best classification model. Hence, that AUC value in practice should be close to 1 for a good classifier. Figure 3 shows the comparison of the ROC curves of RF, NB and MLP classifiers. The utilization of proposed algorithm has been proven in enhancing the detection rate of the minority classes in imbalanced training data. Also this figure indicates that the classification power for the balanced dataset is more than the unbalanced dataset for each classification model. Moreover, the RF classifier is clearly higher-level to other classifiers used.

The graphical representation in figure 4 displays the performance metric results of the balanced and imbalanced dataset according to the classification models of Random Forest, Naïve Bayes and Multilayer Perceptron. In a similar way, the graphical representation in figure 5

displays the comparative of performance metrics results Naïve Bayes, Random Forest, and Multilayer Perceptron that it applied on the balanced dataset.

When Figure 4 is inspected, it clear that almost all results of performance metrics for all classification models used were increased in balanced dataset than in imbalanced dataset. As well when figure 5 is examined, we notice that all the performance metrics results of Random Forest model for balanced dataset outperform other models.

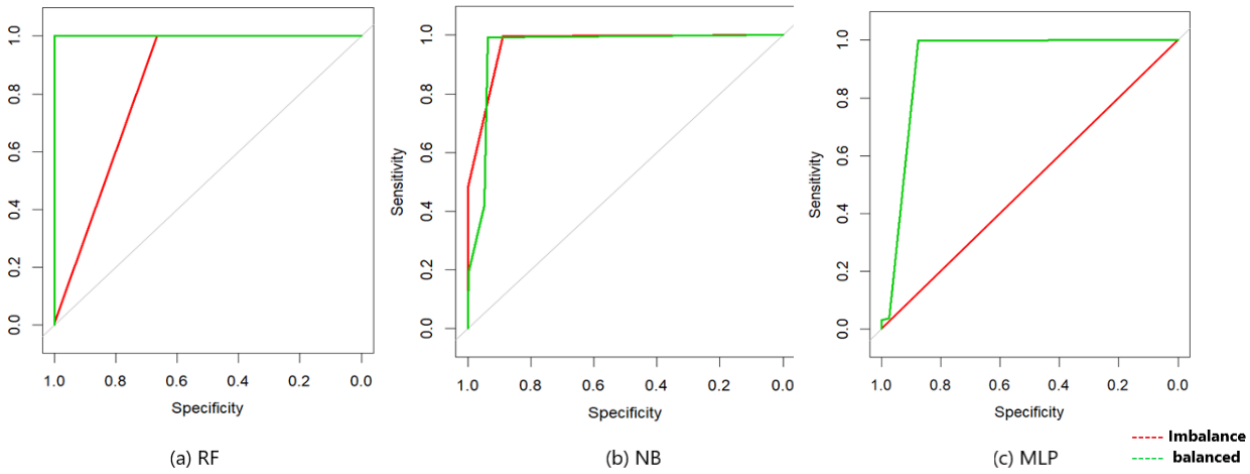

Fig 3. ROC curve for classification models.

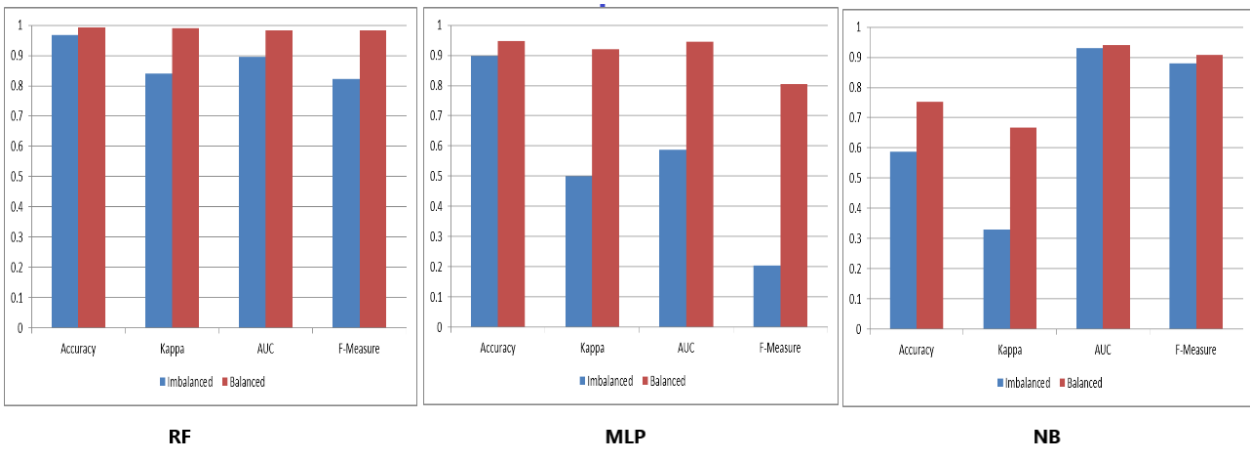

Fig 4. Graphical representation of performance examination results for the balanced and imbalanced combined dataset according to each classification model.

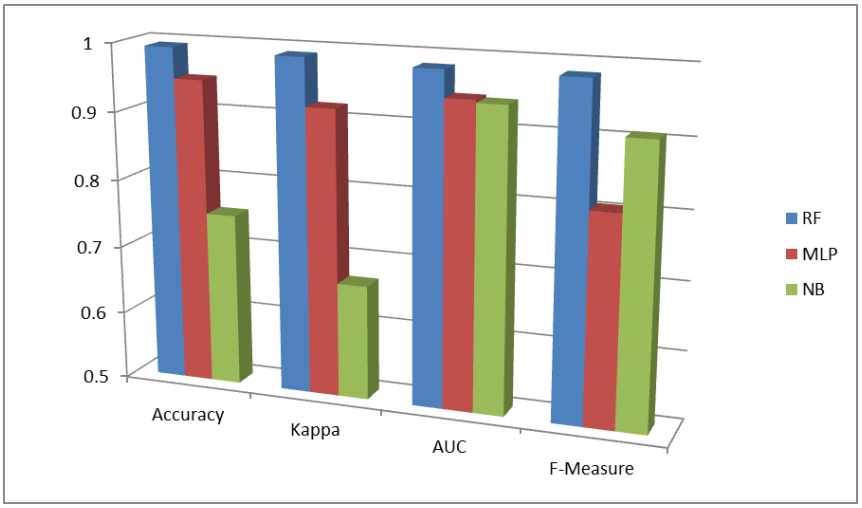

Fig 5. Graphical representation of performance examination results for balanced combined dataset according to the classification models.

Table 3 shows the difference between our study and other studies in the literature that resolved class imbalanced problem.

Table 3: A comparison of the proposed model and previous studies.

Works	Dataset	method	Classes	Classifier	Acc(%)	Kappa (%)	AUC (%)	F1(%)
Hongpo Zhang et al (2020)	CICIDS2017	SGM	multiclass	RF	93.08			94.67
				MLP	99.60			99.69
				CNN	99.85			99.86
Arif Yulianto et al (2019)	CICIDS2017	SMOTE	binary	AdaBoost	81.83			90.01
Razan Abdulhammed et al (2019)	CICIDS2017	UDBB	multiclass	RF	98.8			98.8
				NB	97.6			97.7
				LDA	95.7			95.7
				QDA	98.9			99.0
Aamer Hanif et al (2017)	Customer Churn	Undersampling Oversampling SMOTE	binary	RF			98.5	97
Cengiz Colak et al (2017)	Atrial fibrillation	SMOTE	binary	GLMBoost	82.47		82.59	
				LogitBoost	96.95		96.96	
Our work	CICIDS2017	BMCD	multiclass	RF	99.32	98.96	98.27	98.36
				MLP	94.82	91.98	94.47	80.63
				NB	75.35	66.83	94.10	90.82

4. Conclusion.

Intrusion Detection Systems (IDS) is still an area of primary concern for researchers and producers in this filed. This paper was described the latest intrusion detection dataset and it presented the solution of imbalanced dataset problem. The results of predicted indicated that random forest classifier when our proposed algorithm is applied to create balanced dataset has achieved the values highest accurate values in performance metrics. The results indicate that

SMOTE as well as other oversampling methods can be very useful to overflow the problems of class unbalance in intrusion detection systems.

References

1. Koch, R., Golling, M. G., and Rodosek, G. D. (2017), Towards comparability of intrusion detection systems: New data sets. In Proceedings of the TERENA Networking Conference.
2. Markus R, Sarah W, Deniz S, Dieter L and Andreas H. (2019), A Survey of Network-based Intrusion Detection Data Sets. arXiv:1903.02460v2 [cs.CR].
3. Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, (2018), Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization, 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal,
4. Verbiest N, Ramentol E, Cornelis C, Herrera F.,2014, Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. Appl S Comput; 22: 511-517.
5. Fernández A, Jesus MJ, Herrera F,(2015), Addressing overlapping in classification with imbalanced datasets: a first multiobjective approach for feature and instance selection, Intel Data Eng Autom Learn IDEAL; 36-44.
6. Hongpo Zhang, Lulu Huang, Chase Qishi and Zhanbo Li. (2020) .An Effective Convolutional Neural Network Based on SMOTE and Gaussian Mixture Model for Intrusion Detection in Imbalanced Dataset. DOI: 10.1016/j.comnet.2020.107315.
7. Arif Yulianto, Parman Sukarno and Novian Anggis Suwastika, (2019), Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset. In Journal of Physics: Conference Series (Vol. 1192, No. 1, p. 012018). IOP Publishing
8. Razan Abdulhammed, Hassan Musafer, Ali Alessa, Miad Faezipour and Abdelshakour Abuzneid., (2019, Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection. Electronics 8(3):322; DOI: 10.3390/electronics8030322.
9. Aamer Hanif and Noor Azhar (2017). Resolving Class Imbalance and Feature Selection in Customer Churn Dataset, International Conference on Frontiers of Information Technology, Yogyakarta, Indonesia.
10. Cengiz Colak, E.Karaaslan, C. Colak, A. K.Arslan, N. Erdil, (2017). Handling imbalanced class problem for the prediction of atrial fibrillation in obese patient, Biomedical Research India 2017 Volume 28 Issue 7.
11. Specht, S.M. and Lee, R.B. (2004), Distributed denial of service: taxonomies of attacks, tools, and countermeasure, ISCA PDCS ,.
12. R. Shirey, (2000), Internet Security Glossary, RFC Editor, United States.
13. Paul Sabanal, (2017), Thingbots: The Future of Botnets in the Internet of Things, Security Intelligence.
14. B. Venkatesh, J. Anuradha, (2019), A Review of Feature Selection and Its Methods, Cybernetics and Information Technology, Volume 19, No 1, ISSN: 1314-4081.
15. Kiranmai, B., and A. Damodaram., (2014), A review on evaluation measures for data mining tasks. International Journal of Engineering and Computer Science 3, no. 07.

IRAQI Academic Scientific Journals

Samarra Journal of Pure and Applied Science

SJPAS

www.sjpas.com

ISSN:2663-7405

بناء مجموعة بيانات متوازنة لكشف التسلل استنادا الى CICIDS2017

عامر عبد المجيد عبد الرحمن [1]*، محمود خليل ابراهيم [2]

1- معهد المعلوماتية للدراسات العليا (amer6567@gmail.com)

2- كلية هندسة المعلومات، جامعة النهرين (mahmood_khalel@yahoo.com)

البحث مستل من اطروحة دكتوراه الباحث الاول

معلومات البحث:

تأريخ الاستلام: 2020/5/14

تأريخ القبــول: 2020/07/23

الكلمات المفتاحية:

Imbalanced dataset Classification, SMOTE, CICIDS2017 dataset, Random Forest, Naïve Bayesian, Multilayer Perceptron

الخلاصة:

تم اقتراح العديد من أنظمة كشف التسلل (IDS) في العقد الحالي. تعاني معظم مجموعات البيانات التي ترتبط بمجموعة بيانات كشف التسلل من مشكلة الفئات الغير متوازنة. تحد هذه المشكلة من أداء المصنف للفئات الاقل. قدمت هذه الورقة تقنية جديدة لمعالجة الخلل في التوازن لمجموعة بيانات متعددة الفئات على نطاق واسع، واشير اليها باسم BMCD. تعتمد خوارزمياتنا على تكييف تقنية أخذ العينات الزائدة للأقليات الاصطناعية (SMOTE) مع مجموعة بيانات متعددة الفئات لتحسين معدل الكشف عن فئات الأقليات مع ضمان الكفاءة. في هذا العمل، تم دمج خمس مجموعات بيانات CICIDS2017 فردية لإنشاء مجموعة بيانات متعددة الفئات تحتوي على عدة أنواع من الهجمات. لإثبات كفاءة الخوارزمية الخاصة بنا، تم تطبيق العديد من خوارزميات التعلم الآلي على مجموعة البيانات المدمجة مع خوارزمية BMCD وبدونها. وقد خلصت النتائج التجريبية إلى أن BMCD يوفر حلاً فعالًا لاكتشاف الاختراق غير المتوازن ويتفوق على أساليب كشف الاختراق الحديثة