iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41586

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

Review Paper on Network Data Verification Using Machine Learning Classifiers Based On Reduced Feature Dimensions

Miss. Maithili Deshmukh[1], Dr. M. A. Pund[2]
[1] *PG Scholar,* [2] *Professor, Computer Science & Engineering, Prof Ram Meghe Institute of Research and Technology, Amravati, Maharashtra, INDIA*

Abstract: ***With the rapid development of network-based applications, new risks arise and additional security mechanisms require additional attention to improve speed and accuracy. Although many new security tools have been developed, the rapid rise of malicious activity is a serious problem and the ever-evolving attacks pose serious threats to network security. Network administrators rely heavily on intrusion detection systems to detect such network intrusion activity. A major approach is machine learning methods for intrusion detection, where we learn models from data to differentiate between abnormal and normal traffic. Although machine learning methods are often used, there are some shortcomings in the in-depth analysis of machine learning algorithms in terms of intrusion detection. In this work, we present a comprehensive analysis of some existing machine learning classifiers with respect to known intrusions into network traffic. Specifically, we analyze classification with different dimensions, that is, feature selection, sensitivity to hyper-parameter selection, and class imbalance problems that are involved in intrusion detection. We evaluate several classifications using the NSL-KDD dataset and summarize their effectiveness using detailed experimental evaluation.***

Keywords: IDS, Machine Learning, Classification Algorithms, NSL-KDD Dataset, Network Intrusion Detection, Data Mining, Feature Selection, WEKA, Hyperparameters, Hyperparameter Optimization.

I. INTRODUCTION

To extract important features from these higher dimensions of variables and data. Statistical techniques were used to reduce noise and redundant data. However, we do not use all the features to train the model. We can improve our model with both correlated and non-correlated features, so the choice of features plays an important role.

Furthermore, it not only supports our model to be trained faster but also reduces the complexity of the model, makes it easier to understand and improves the accuracy, precision or metric performance in memory. There are four main reasons why the choice of facilities is necessary. First, discard the model to reduce the number of parameters. Further reduce training time, increase normalization to reduce overfilling, and avoid the curse of dimensionality.

In the field of data processing and analysis, datasets can be a large number of variables or characteristics that determine the applicability and usefulness of the data.

Also the challenge of classification is to pay attention to balanced and unbalanced data. Another motivation is to get the best model with high predictions and small errors.

High-dimensional means that there can be hundreds, thousands, or even millions of input variables. Low input dimensions usually have fewer parameters or a simple structure in machine learning models, called degrees of freedom. Models with too much freedom are likely to over-fit the training dataset and therefore do not perform well on new data. Instead it is desirable to have a simple model with good generalization and input data with few input variables. This is especially true for linear models where the number of inputs and the degree of freedom of the model are often closely related.

II. LITERATURE SURVEY

1) In this Paper their analysis results for the performance of the six different classifiers on the NSLKDD dataset shows that J48 and IBK are the best two classifiers in terms of accuracy detection but IBK is much better when applying feature selection techniques. For future work, we propose to carry out an exploration on how to employ optimization techniques to develop an intrusion detection model with a better accuracy rate

2) The conclusion of the present study can be summarized into the following three points:

a) A social media GIS which integrated a Web-GIS, an SNS and Twitter and which included a function for classifying submitted information was designed and developed. A system which supports utilization of information in order to reduce the effects of natural disasters which anticipates use not only in normal times but also use during disasters when there is an information glut was proposed. The system supports utilization of information by depicting submitted information based on location information and content using color-coded semitransparent circles, and by displaying information based on information about present location. Mitaka City in the Metropolis of Tokyo was selected as the region for operation of the system. After a survey of existing conditions was conducted, the system was developed in detail.

b) Since full-scale operation was to be conducted for ten weeks, a one-week operation test was conducted in advance and an area for improvement of the system was identified. After that, the system was reconfigured. People targeted as users of the system were those residing in, commuting to, or attending school in Mitaka City aged eighteen years of age or over. The number of users was fifty in total. Users in their forties and fifties made up the greatest proportion of users, at 32% and 30%, respectively, while the proportion of users in their twenties, thirties, and sixties and above was 14%, 12%, and 10%, respectively, so the system was used by a wide range of age groups. During the period of operation, users accessed the system from PCs and mobile information terminals, and submitted and viewed information.

c) An evaluation of the operation was conducted based on access log analysis and submitted information. The former showed that the system was continually accessed throughout the period of operation, and the later showed that 260 pieces of disaster information were submitted, dispersed throughout Mitaka City. Based on the results of the evaluation of the operation, measures for using the system even more effectively were summarized into the following two points: (1) Notification of information to passive users, and (2) Operation of the system using cloud computing. Future topics for research include expanding the stages of use of the system to times of post-disaster restoration and redevelopment, cooperating with firefighters and police, and operating the system with the participation of a wider user base, and increasing the track record of use of the system by operating it in other regions as well, and further increasing the significance of using the social media GIS developed in the present study

3) The dimensionality reduction problem is responsible for the process of reducing the dimension of a large feature set into a combined reduced-feature set that makes up a large sphere in the n-dimensional space. Hence the dimensionality reduction problem presents advantages like computational efficiency and redundancy removing and other disadvantages such as data-losing and feature-losing in datasets. We have worked in the fields of dimensionality reduction in large datasets. So, our machine learning method for reducing dimensional space in large datasets merges all datasets as a huge one in the first stage. Then, we used PCA for dimensionality reduction, that solves the problems encountered in previous proposals, after applying the ETL process. Hence, our novel method highlights the use of the Python framework where standard representation for numerical data and implementations for mathematical calculations at a high level of programming are performed in an efficient environment. Besides, we have applied our research in the real environment Epileptic Seizure Recognition Data Set provided by UCI Machine Learning Repository. The experimental results show that Random Forest outcomes better than the rest of the algorithms with this complex dataset obtaining an accuracy of 70.3%. We can also appreciate that MLP behaves very stable in general terms and together with Random Forest are the best algorithms to be optimized. In this regard, as shown in figure 2, we can realize how important is the correct selection of a particular scalar or another for the algorithms due mainly to the taxonomy of the data. The intrinsic complexity of the dataset tested in this manuscript suggests excellent conditions for adaptation to other health care scenarios, where the complexity of biological systems will also be adapted to our generic methodology. Although the results obtained are very encouraging, the greatest achievement in the authors' opinion is the possibility of future lines. At this point and due to the complexity of the data currently generated, with characteristics of variability and other aspects besides the volume, such as velocity, veracity of the big data, where a new world has opened up to continue in this fascinating area of data science research.

4) In this paper, the state of the art on ensemble methodologies to deal with class imbalance problem has been reviewed. This issue hinders the performance of standard classifier learning algorithms that assume relatively balanced class distributions, and classic ensemble learning algorithms are not an exception. In recent years, several methodologies integrating solutions to enhance the induced classifiers in the presence of classim balance by the usage of ensemble learning algorithms have been presented. However, there was a lack of framework where each one of them could be classified; for this reason, a taxonomy where they can be placed has been presented. We divided these methods into four families depending on their base ensemble learning algorithm and the way in which they address the class imbalance problem. Once that the new taxonomy has been

presented, thorough study of the performance of these methods in a large number of real-world imbalanced problems has been performed, and these approaches with classic ensemble approaches and non ensemble approaches have been compared. We have performed this study developing a hierarchical analysis over the taxonomy proposed, which was guided by nonparametric statistical tests. Finally, we have concluded that ensemble-based algorithms are worthwhile, improving the results that are obtained by the usage of data preprocessing techniques and training a single classifier. The use of more classifiers makes them more complex, but this growth is justified by the better results that can be assessed. We have to remark the good performance of approaches such as RUSBoost or UnderBagging, which despite being simple approaches, achieve higher performances than many other more complex algorithms. Moreover, we have shown the positive synergy between sampling techniques (e.g., undersampling or SMOTE) and Bagging ensemble learning algorithm. Particularly noteworthy is the performance of RUSBoost, which is the computationally least complex among the best performers

III. SYSTEM DESIGNS

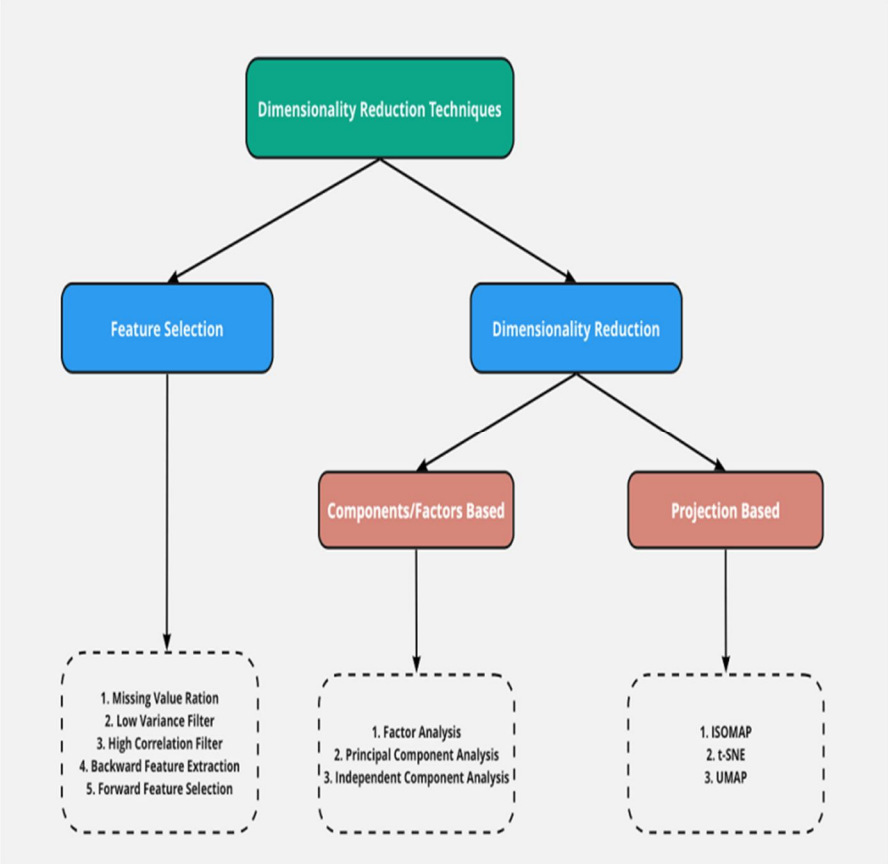


Fig : Network Data Verification Using Machine Learning

IV. CONCLUSION

Hence we discuss the above research paper & concluded the system which proposed Network Data Verification Using Machine Learning Classifiers Based On Reduced Feature Dimensions. The further improvisation of system is possible as the classifier and datasets are trained.

V. ACKNOWLEDGEMENT

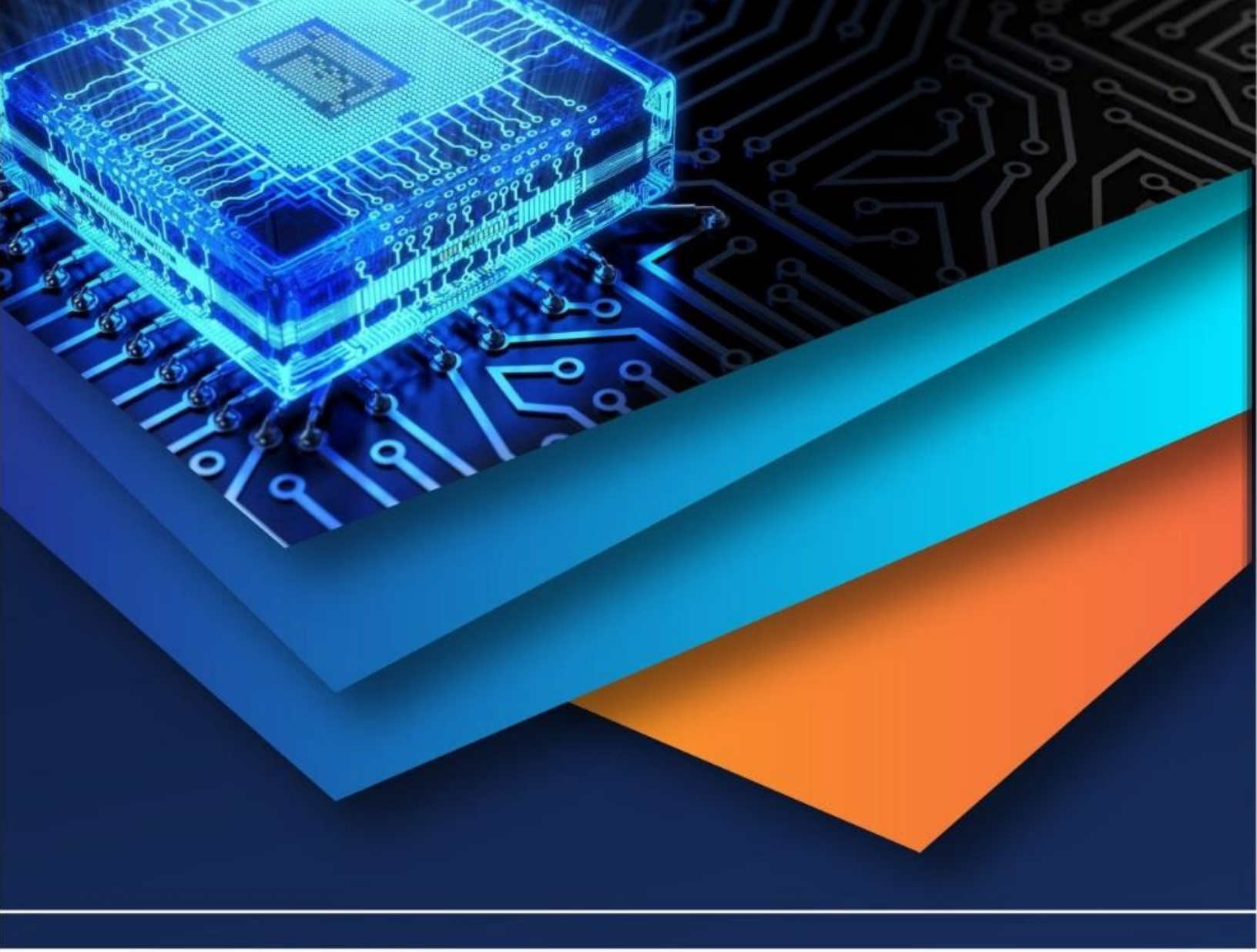
First and foremost, I would like to express my sincere gratitude to my Dr M. A. Pund who has in the literal sense, guided and supervised me. I am indebted with a deep sense of gratitude for the sconstant inspiration and valuable guidance throughout the work.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] Ahmed Mahfouz, Deepak Venugopal, Sajan G Shiva "Comparative Analysis of ML Classifiers for Network Intrusion Detection" August 2019 Conference: International Congress on Information and Communication TechnologyAt: London, UK
[2] D. B. Roy, R. Chaki, State of the art analysis of network traffic anomaly detection, in Ap-plications and Innovations in Mobile Computing (AIMoC), 2014, IEEE, 2014, pp. 186–192.
[3] RAFAEL MUÑOZ TEROL 1 , ALEJANDRO REINA REINA2 , SABER ZIAEI 3 , AND DAVID GIL 4 1 A Machine Learning Approach to Reduce Dimensional Space in Large Datasets
[4] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, ''A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,'' IEEE Trans. Syst., Man, Cybern.C, Appl. Rev., vol. 42, no. 4, pp. 463–484, Jul. 2012.
[5] R. Polikar, ''Ensemble based systems in decision making,'' IEEE Circuits Syst. Mag., vol. 6, no. 3, pp. 21–44, 3rd Quart., 2006.
[6] T. G. Dietterich, ''Ensemble methods in machine learning,'' in Proc. Int. Workshop Multiple Classifier Syst.. Berlin, Germany: Springer, 2000.
[7] R. M. O. Cruz, R. Sabourin, and G. D. C. Cavalcanti, ''Dynamic classifier selection: Recent advances and perspectives,'' Inf. Fusion, vol. 41,pp. 195–216, May 2018.
[8] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau, ''Dynamic selection of exemplar-SVMs for watch-list screening through domain adaptation,'' in Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods, 2017, pp. 738–745.
[9] P. R. L. de Almeida, E. J. da Silva Junior, T. M. Celinski, A. de Souza Britto, L. E. S. de Oliveira, and A. L. Koerich, ''Music genre classification using dynamic selection of ensemble of classifiers,'' in Proc. IEEE Int. Conf. Syst., Man, Cybern., Oct. 2012, pp. 2700–2705.
[10] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, ''Benchmarking state-of-the-art classification algorithms for credit scoring,'' J. Oper. Res. Soc., vol. 54, no. 6, pp. 627–635, Jun. 2003.
[11] H. Xiao, Z. Xiao, and Y. Wang, ''Ensemble classification based on supervised clustering for credit scoring,'' Appl. Soft Comput., vol. 43, pp. 73–86,Jun. 2016.
[12] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, ''A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,'' IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 42, no. 4, pp. 463–484, Jul. 2012.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)