www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume – 5 Issue -04 April, 2016 Page No. 16156-16161

A Survey on Improving Classification Performance Using Data Pre processing And Machine Learning Methods on NSL-KDD Data

Mr.Shobhan Kumar[1], Mr.Naveen D.C[*2],
[#]Computer Science & Engg, NMAMIT NITTE
Udupi (D) Karnataka
[1]shobhank9@gmail.com
[2]chandavarkar@gmail.com

Abstract: This paper gives an indication of our study in building rare class prediction models for identifying known intrusions and their variations and anomaly detection schemes for detecting novel attacks whose nature is unknown. Data mining and machine learning have been subjected to general explore in intrusion detection with emphasis on improving the accuracy of detection classifier. The quality of the feature selection methods is one of the important factors that affect the effectiveness of Intrusion Detection scheme (IDS). This paper evaluates the performance of data mining classification algorithms namely C4.5, J48, Nave Bayes, NB-Tree and Random Forest using NSL KDD dataset and focuses on Correlation Feature Selection (CFS) assess. The results demonstrates that NB-Tree and Random Forest outperforms other two algorithms in terms of predictive accuracy and detection rate

I. INTRODUCTION

As network-based computer systems contribute vital roles in present-day society, they have become the target of intrusions by enemies and criminals. With the development of internet, network security becomes an indispensable factor of computer technology. Therefore, the role of Intrusion Detection Systems (IDS), as special-purpose devices to detect anomaly- lies and attacks in the network is becoming more important as it gathers and analyzes information from various areas within a computer or a network to identify possible security breaches.

The research in the intrusion detection field has been mostly focused on anomaly-based and misuse-based detection techniques. Data mining techniques are used to explore and analyse large dataset and find useful patterns. Classification is the category that consists of identification of class labels of records that are typically described by setoff features in dataset. The term Knowledge Discovery from data (KDD) refers to the automated process of knowledge discovery (data mining) from databases, it comprises of many steps namely data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation. The aim of this paper is to develop a system which uses various pre-processing methods such as Feature Selection and Descritization. With the help of Feature selection algorithm required features are selected and due to Descritization the data is discredited which can be applied to various classifier algorithms such as Naive Bayes, Hidden Naive Bayes and NB-Tree.

II. LITERATURE SURVEY

Datta H. Deshmukh et al. [1] developed a system which uses pre-processing methods like feature selection and descritization. With the help of feature selection algorithm required features are selected and due to descritization the data sets are discredited which is then applied to classifier algorithms like Naive Bayes, Hidden Naive Bayes [15], NB Tree. The proposed system explains the need to relate data mining methods to network events to classify network attacks and in order to increase the accuracy of the classifier they implemented pre-processing methods like Feature selection and descritization on NSL-KDD dataset [3]. By using Fast Correlation based Filter Algorithm [14] there was an attempt to overcome the problem of high dimensionality of dataset. Naive Bayes has drawback of conditional independence assumption to overcome this issue they implemented Hidden Naive Bayes classifier and NB Tree classifier. Thus this improved the accuracy of the classifier and decreased the error rate. The output of the proposed method are checked for true positive, true negative, false positive and false negative. Based on these

values accuracy and error rate of each classifier was computed.

Adetunmbi A. Olusola et al.[3] studied the relevance of each feature in KDD 99 intrusion detection dataset for detection of each class. Rough set degree of reliance and dependency proportion of each class were employed to decide the sharpest features for each class. Selecting the right features is challenging, but it must be performed to diminish the amount of features for the sake of efficient processing speed and to remove the irrelevant, redundant and noisy data for the sake of predictive accuracy. The training set employed for the analysis is the 10 percentile KDD-Dataset. Since the degree of dependency is calculated for discrete features, continuous features are discredited based on entropy. Prior to the descritization, superfluous records from the dataset were detached since rough set does not require spare instances to classify and categorize discriminating features. In this experiment, two approaches are adopted to detect how significant a feature is to a given class. The first approach is to compute degree of dependency for each class based on the available number of class instances in the data set. Thus, signifying how well the feature can discriminate the given class from other classes. Secondly, each class labels are mapped against others for each attribute. That is, generating a frequency table of a meticulous class label against others based on variations in each attribute and then a assessment made to generate the dependency ratio of principal classes inorder to detect all the relevant features distinctive one class from another. Graphical analysis is also employed in the analysis in order to detect the relevant features for each class.

Mahbod Tavallaee et al. [5] conducted a statistical analysis on KDD CUP 99 data set, they found some important issues which extremely affects the presentation of evaluated systems, and results in a very poor evaluation of anomaly detection approaches. To address these issues, they have anticipated a new data set, NSL-KDD [3] which has advantages over the original KDD data set

Sunitha Beniwal et al. [12] discussed a variety of Classification and Feature Selection technique in data mining. Classification is a procedure used for discovering classes of mysterious data. Feature selection is used to avoid over fitting and to improve model performance to provide quicker and further cost effectual models. Various methods for classification exist like Bayesian, decision trees, etc.

Hence before applying any mining technique irrelevant attributes needs to be filtered. Filtering is done using different feature selection techniques like wrapper, filter, and embedded technique. The data mining tasks can be broadly classified in two categories: the first one is descriptive and the second one is predictive. Descriptive mining tasks characterize the broad-spectrum properties of the data in the catalog. Predictive mining responsibilities perform inference on the current data in order to make predictions.

Lei Yu et al. [14] introduced a novel concept, Fast correlation based filter solution for the feature selection of high dimensional data which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. Feature selection has two categories they are filter model and wrapper model. The filter model depends on general attributes of the preparation information to select a few elements without including any learning algorithm. The wrapper form requires one predestined learning calculation in highlight determination and employments its execution to assess and figure out which features are chosen. When the number of features extracted becomes very large. A new feature selection algorithm FCBF (Fast Correlation Based Filter Method) is implemented and evaluated through extensive experiments comparing with associated feature collection algorithms. In these algorithms, first a goodness measure of feature subsets based on data distinctiveness is used to decide best subsets for the given cardinality and then cross validation is exploited to decide a best subset across different cardinalities. These algorithms mainly focus on combining filter and wrapper algorithms to attain best promising presentation. In this work, they focused on the filter model and aimed to develop a new feature collection algorithm, that is be capable of successfully remove both irrelevant and redundant features and is less expensive in calculation than the presently available algorithm. The efficiency and effectiveness of their method is demonstrated through general comparisons with other methods by means of real-world data of high dimensionality.

V. Bolon-Canedo et al. [9] proposed a new approach that consists of combination of descritization and filter methods designed for improving categorization performance in KDD CUP99 dataset. Study of KDD Cup 99 dataset suggests that there are some features which are irrelevant and correlated. In this work during feature selection process filter method is used which allows to diminish dimensionality of the dataset. This method was chosen because of large size of the KDD Cup'99 dataset. Filter method cannot deal directly with numerical attributes so there is a need to discredit the data before applying filter method. Descritizators chosen in this work are Entropy Minimization Discrization (EMD) and Equal width Descritization (EWD), Equal Frequency Descritization (EFD) in order to work well with large dataset. A new method is proposed so that we can use the whole training dataset, that is by combining a descritization and feature selection method followed by a filter method thus it was possible to obtain good performance results because of important reduction in number of input features Experimental analysis in 3 large datasets showed the improvement in performance of binary classification. Dokas et al. [2] did a research on building rare class prediction models for identifying known intrusions and their variations they also done a study on anomaly detection [11] schemes for detecting novel attacks whose nature was un-known. Using these techniques it is possible to automatically retrain intrusion detection models on different input data that includes new types of attacks but those needs to be labelled appropriately. In misuse detection [11] technique high degree of accuracy was seen in detecting known attacks and their variations. By using anomaly detection technique it was possible to identify new types of intrusions. Experimental results on the KDD Cup '99 dataset showed that rare class predictive model is much more efficient in the detection of intrusive behaviour than any other standard classification techniques. When performing experiments on DARPA 98 dataset, the unsupervised SVM were very promising in detecting new intrusions but they had extremely high counterfeit alarm rate. Therefore, future work was needed in order to keep high detection rate while lowering the false alarm rate.

Ozguru Depren et al. [11] proposed a novel intrusion detection system architecture utilizing both anomaly and misuse detection approaches. It contains anomaly detection module, misuse detection module and also decision support system which is built from the results obtained by these two modules. The proposed anomaly detection modules uses self organizing map (SOM) to model normal behaviour. Deviation from the normal behaviour is considered as an attack. Proposed misuse detection module uses J.48 decision tree algorithm to classify the different types of attacks. Their goal was to increase the presentation of the projected method with KDD Cup 99 data set when compared to data set used by IDS researchers. Using Decision Support System they could interpret the results of anomaly and misuse detection modules. By the experimental analysis they found that proposed hybrid approach gave better performance than the individual approaches. From the simulation results of both anomaly and misuse detection module based on KDD 99 dataset they obtained a detection rate of 98.96 percentile and false positive rate of 1.01 percentile for anomaly detection module and also a classification rate of 99.61 percentile and a very low false positive rate of 0.20 percentile for the misuse detection module.

Amudha P et al., Have done evaluation study on performance of data mining classification algorithms namely J48, Naive Bayes, NB Tree and Random Forest using KDD-Cup'99 dataset [7]. Data mining techniques are the new approach for intrusion detection. Since the quality of the feature selection methods is an important factor that affects the effectiveness of IDS it is necessary to evaluate the performance of data mining classification algorithms namely Naive Bayes, NB Tree, and Random Forest using KDD Cup'99 dataset. Here the main focus is on Correlation Feature Selection in order to obtain optimal set of features for classification but the detection performance of these methods closely relies on the huge amount and high quality of training samples. When data mining is introduced into the intrusion detection it mainly focus on two problems that is to establish the adaptive feature dataset and to improve the detection rate. In this paper they developed series of experiments on KDDCup'99 dataset for classifying the attack and to examine

the effectiveness of correlation feature selection measure the results indicate that Random Forest gives better accuracy, detection rate, false alarm rate and NB Tree gives better accuracy.

Lee Wenke et al. [6] presented data mining framework for adaptively building intrusion detection models. Their central thought was to make use of auditing programs to take out an extensive set of features that describe each network link or host session and relate data mining programs to study rules that accurately capture the behaviour of intrusions and normal activities. Those rules can be used by both the categories of IDS; data mining programs used has the strengths like classification, meta-learning, association rules. They developed set of automatic tools that can be applied to variety of audit information sources to produce intrusion discoverymodels. Their automatic approach eliminates the need of manual examine and encoding of intrusion patterns. Experiments show that the frequent patterns mined from audit data can be used as consistent anomaly discovery models and also as guidelines for selecting features to build effective classification models.

Mrutyunjaya Panda et al. [8] did a comparative study of data mining algorithms for network intrusion detection. In this paper, the presentation of three well recognized data mining classifier algorithms namely, ID3, J48 and Nave Bayes [15] were evaluated based on the 10-fold cross validation test. They compared the effectiveness of the classification algorithm Nave Bayes with the decision tree algorithms [4] namely, ID3 and J48. This helped to construct an effective network intrusion detection system. It was observed from the experimentation results that the Nave Bayes model is quite pleasing because of its ease, sophistication, robustness and effectiveness. On the other hand, decision trees had proven its efficiency in both generalization and detection of new attacks. The results showed that there is no single finest algorithm to smash others in all situations. In certain cases there might be dependence on the characteristics of the data. To choose a suitable algorithm, the results of the classification were required to make better decisions. Since, these do not deal with unknown attacks future investigation was directed towards handling new attacks.

Ron Kohavi et al.[13] proposed a new algorithm for Scaling up the accuracy of Naive Bayes classifier called Decision-Tree Hybrid. In bigger database accurateness of the Naive Bayes and decision trees does not scale up very well. So they proposed a algorithm called NB tree which is hybrid of decision tree classifier and Naive Bayes classifier. This approach made them to retain the inter operability of Naive Bayes as well as decision trees, and also worked well for the larger database. Naive Bayesian classifier are very vigorous to irrelevant attributes and to make the final prediction it should consider evidence from many attributes this does not have any main effect. In order to get efficient accuracy in larger data base they had to create strong self-rule assumption on Bayes classifier. Decision tree classifiers are fast but their current method includes recursive partitioning which faced the fragmentation problem. NB tree approach takes advantage of both of these classifier Decision tree built with uni variant splits at each node and Naive Bayes classifier at the leaves.

The final classifier resembles Utgoff's Perception trees, but the induction process is very different and geared toward larger datasets. This resulting classifier was as easy to interpret as Decision tree and Naive Bayes. NB tree hybrid approach was suitable for learning the things when many attributes are likely to be relevant for the classification task. After the experimental analysis they found that their approach to NB tree classifier works well for real world datasets that they tested and scales up well in terms of its accuracy.

Harry Zhang et al. [15] proposed a novel model called Hidden Naive Bayes (HNB). Learning about optimal structure of the Bayesian network was time consuming so there was a need to build Bayesian model without considering its structure learning. To overcome the limitations of Naive Bayes Hidden Naive Bayes was introduced by adding hidden parent for each attribute on Naive Bayes. Hidden parents were created using weighted one-dependence estimators. In this process weights were given more importance than the structural learning. They used conditional mutual information to estimate the weights directly from the data. HNB inherits the structural simplicity of Naive Bayes and it can be easily learned without the structure learning. They proposed an algorithm for learning HNB based on conditional mutual information and tested HNB in provisions of classification exactness using the 36 UCI data sets recommended by Weka then they compared it with the

naive Bayes. In their experimental results they found that HNB was better than Naive Bayes and any other method. A Bayesian network comprises a structural representation and a set of conditional probabilities. The structural model is a directed chart in which nodes symbolize attributes and arcs represent attribute dependencies. The attribute dependencies are quantified by restrictive probabilities for each and every node given its parents. Due to the simplicity and comprehensibility of HNB it was considered to be a promising model that could be used for many real world applications. Implementations of HNB were based on one dependence estimator which was later generalized to arbitrary dependence estimator.

Manish Kumar et al. [4] done a study on Misuse and Anomaly attack detection using Decision Tree algorithm .IDS are classified as anomaly-based or signature based. In misuse detection [11] abnormal system behaviour is defined first, and then other actions are defined as usual behaviour. The anomaly detection [11] utilizes reverse approach, defining normal system behaviour first and any other unknown behaviour is seen as abnormal. Signature-based systems are similar to virus scanners and appear for known, doubtful patterns in their input data. Anomaly-based systems watch for deviations of actual behaviour from reputable profiles and categorize all abnormal activities as malicious. The advantage of signature-based designs is, they can identify attacks with an acceptable accuracy and they tend to produce fewer false alarms than the anomaly based. Although an anomaly-based variant has advantage of being able to find prior unknown intrusions, the costs of dealing with a huge number of counterfeit alarms is often prohibitive. Decision trees use a pre-classified dataset to learn to categorize data based on current trends and patterns. After the tree is created, the logic from the decision hierarchy can be integrated into a numeral of diverse intrusion detection technologies including firewalls and IDS signatures.

Zheguo Chen et al. [16] introduced a new intrusion detection method based on Support Vector Machines improved by artificial immunization algorithm. In order to solve problems with nonlinearity had high dimensionalities, a powerful computational tool called Support Vector Machine novel classification technique was used and it had shown higher performance than traditional learning methods. The parameter of SVM can greatly affect the performance of classifier thus in order to improve the capability of the SVM classifier, the artificial immunization algorithm is applied so that it is possible to optimize and select the constraint. This paper tells about the simulation that showed intrusion detection based on support vector machine improved by artificial immunization algorithm to achieve better generalization performance than that without parameter selection. The experimental results showed that the above mentioned technique gave higher recognition accuracy than the general SVM.

Huy Anh Nguyen et al. [10] explained the application of data mining to network intrusion detection using classifier range model. They concluded that the performance of a comprehensive set of classifier algorithms using KDD 99 dataset. Based on evaluation outcome, the best algorithms for each assault category was chosen and two classifier algorithm selection models are proposed which promise for the performance improvement and real time systems application they are Parallel model for classifier selection and Parallel model for real time application classifier selection. The simulation result judgment indicates that clear performance improvement and real time intrusion detection can be achieved as they applied their proposed models to detect different kinds of network attacks.

III Conclusion.

A survey of feature collection, extraction and importance of the machine learning algorithms are projected. The objective of both method concerns the diminution of feature space in order to advance data analysis. This aspect becomes more important when real world datasets like NSL-KDD data-sets are considered, which can contain hundreds or thousands features. The main dissimilarity between feature selection and extraction is that the first performs the diminution by selecting a subset of features exclusive of transforming them, while feature extraction reduces dimensionality by computing a renovation of the novel features to create other features that ought to be more momentous. Hence with proper feature selection method and machine learning algorithms like SVM and KNN it is possible to achieve higher accuracy rate. The attribute selection

improves comprehension of the process under consideration, as it points out the features that mostly influence the considered observable fact. Moreover the calculation time of the adopted learning machine and its accuracy need to be considered as they are crucial in machine and data mining applications.

REFERENCES

[1] Datta H. Deshmukh, Tushar Ghorpade, Puja Padiya, Improving Classification using Preprocessing and Machine Learning Algorithms On NSL-KDD Dataset, 2015.

[2] Dokas, P., Ertoz, L. Kumar, V., Lazarevic, A., Srivastava, J. and Tan, P.N, (2002, Nov). Datamining for Network Intrusion Detection in proc.NSF workshop on Next Generation Datamining.(pp.21-30)

[3] Olusola, Adetunmbi A., Adeola S. Oladele, and Daramola, O. Abosede. "Analysis of NSL-KDD99 Intrusion Detection Dataset for Selection of Relevance features". Proceedings of the world congress on Engineering and Computer Science. Vol. 1. 2010.

[4] Kumar, Manish, M. Hanumanthappa, and TV Suresh Kumar. "Intru-sion Detection System using Decision Tree algorithm." Communi-cation Technology(ICCT), 2012 IEEE 14th International Conference on.IEEE,2012.

[5] Tavallaee, Mahbod, et al. A detailed analysis of the KDD CUP 99 data set. Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009.

[6] Lee,Wenke, Salvatore J. Stolfo, and KuiW. Mok. A data mining framework for building intrusion detection models. Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium on. IEEE, 1999.

[7] Amudha, P., and H. Abdul Rauf. Performance Analysis of Data Mining Approaches in Intrusion Detection. Process Automation, Control and Computing (PACC), 2011 International Conference on. IEEE, 2011

[8] Panda, Mrutyunjaya, and ManasRanjanPatra. A comparative study of data mining algorithms for network intrusion detection. Emerging Trends in Engineering and Technology,2008. ICETET08. First Inter-national Conference on. IEEE, 2008.

[9] Bolon-Canedo, Vernica, N. Sanchez-Maroo, and Amparo Alonso-Betanzos. A combination of discretization and filter methods for improving classification performance in KDD Cup 99 dataset, Neural Networks, 2009. IJCNN 2009. International Joint Conferenceon. IEEE, 2009.

[10] Nguyen, HuyAnh, and Deokjai Choi. "Application of Data Mining to Network Intrusion Detection:classifier selection model," Challenges for Next Generation Network Operations and Service Management. Springer Berlin Heidelberg,2008. 399-408

[11] Depren, O., Topallar, M., Anarim, E.,and Ciliz, M. K. (2005). An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. Expert systems with Applications, 29(4), 713-722.

[12] Beniwal, Sunita, and Jitender Arora. Classification and feature selec-tion techniques in data mining. International Journal of Engineering Research and Technology. Vol. 1. No. 6 (August-2012). ESRSA Publications, 2012.

[13] Kohavi, Ron. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In KDD, pp. 202-207. 1996.

[14] Yu, Lei, and Huan Liu. Feature selection for high-dimensional data: A fast correlation based filter solution. In ICML, vol. 3, pp. 856-863. 2003.

[15] Zhang, Harry, Liangxiao Jiang, and Jiang Su. Hidden naive bayes. Proceedings of the National Conference on Artificial Intelligence. Vol. 20. No. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999,2005.

[16] Zhenguo Chen, Guanghua Zhang. "Support Vector Machine Improved By Artificial Immunisation Algorithm for Intrusion Detection".