DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

Order N°: .....

UNIVERSITY OF M'SILA

FACULTY OF MATHEMATICS AND INFORMATICS

Department of Computer Science

A memoir submitted in partial fulfilment
of the Requirements for the Degree of Master in Computer Science
Option:Networks

By: Youssouf CHADOULI

Subject

A New Feature Selection Approach For Network Intrusion Detection Systems

Committee members:

.....

.....

.....

supervised by:
Saoudi lalia

Promotion: 2013 /2014

Content Table

1.1 Context:

With the expanding and increasing use of networks, and accumulating number of internet users, network throughput has become massive and threats are more diverse and sophisticated. Network and information security are of high importance, and research is continuous in these fields to keep up with the development of attacks.

Intrusion Detection is a major research area that aims to identify suspicious activities in a monitored system, from authorized and unauthorized users.

1.2 Statement of the Problem:

The amount of network data to be examined by an IDS is normally huge as it contains information about various activities in computer network. The network data must be transformed into a format that is manageable and it needs to contain intrinsic or derived features so as to identify intrusions effectively. In addition, large number of features involved in the data complicates the intrusion identification task. Identifying intrusions solely based on human eyes is therefore extremely difficult. To alleviate the problem, network security experts utilize existing data mining and artificial intelligence techniques in search of possible intrusions. However, if the number of features involved in network data increases, identifying intrusions can become difficult because of the complex relationships between features. Complex relationships can be seen as well between the features and intrusion classes. This contributes to high computational costs in processing tasks and subsequently leads to delays in identifying intrusions. Seeing the limitations of both humans and computers, feature selection is thus important to be conducted so that the load in processing data and time consumed in detecting intrusions can be reduced.

1.3 Objectives:

1.3.1. General Objectives:

Features selection, is an important issue in IDSs. A reduced features set improves system accuracy and speeds up the training and testing computation process considerably.

Our objective is to propose a new method for features selection utilizing the bayesian network classifier and Backward Feature selection algorithm, to achieve high accuracy of intrusion detection while keeping the number of features low.

1.3.2. Specific Objectives:

The specific objectives of this research are the following:

- To review different literatures on the concept of intrusion detection in the area of data mining particularly feature selection approaches.
- To experiment effectiveness of classification based wrapper feature selection models.
- To construct models using classification machine learning algorithm on optimal selected features generated by different feature selection techniques.
- To compare the performance of the models built with other recent works.

1.4 Research Approach and Methodology:

This project is comparative study of two feature selection approaches, filter based feature selection approach and wrapper based feature selection approach, using three commonly used filter techniques: *ReliefF*, *Gain Ration* and *Info Gain* attributes selection techniques, and a an effective wrapper technique: *feature sequential search strategy* selection technique (FSSS). In addition, a new wrapper based features selection technique is proposed in this project which mainly focus on performing the highest detection rate and the minimum learning time cost.

Our proposed method is mainly based on building features selection models that pre-process the given input dataset and select the minimal number of features that can perform the best accurate results. Our research methodology consists of 5 main phases as follows:

1.4.1 Research and survey

Include reviewing the recent researches of feature selection for anomaly detection that is closely related to the thesis problem statement. Then analyzing the existing methods, and identifying the drawbacks and disadvantages of each method in order to be overcome in our research.

1.4.2 Data set collection and preprocessing

The dataset to be used in our experiment is the NSL-KDD dataset [21] which is a new dataset for the evaluation of researches in network intrusion detection field. It consists of

selected records of the complete KDD 99 dataset. NSL-KDD dataset solve the issues of KDD 99 benchmark.

The NSL-KDD dataset is available in text format; so to be read by WEKA tool it has to be changed into ARFF format.

1.4.3 Implement features selection algorithms (filter and wrapper approach)

The aim of this research to implement a new features selection algorithm to solve IDS's accuracy and computation problem.

1.4.4 Apply our feature selection algorithm

Using weka program, we will apply our feature selection algorithm with Bayesian network classifier

1.4.5 Evaluate the obtained results

In this stage we will analyze the obtained results and justify the effectiveness of our algorithm by comparing it with other algorithms.

1.5 Thesis Outline:

To meet our objective this paper is structured as follows:

The First chapter consider the context of our work. First of all provides an overview of intrusion detection system and their classification, their detection techniques and in which security strategy they are applied.

The Second chapter will focus on data mining techniques and their application in features extraction methods and highlight their differences.

The third chapter present the conceptual aspect of the study, which is the knowledge discovery process, and the different steps to follow during this study, input datasets preparation, the feature selection models generation and results collection in addition to an explanation of the used evaluation techniques.

In the fourth chapter we will discuss the results obtained from running the feature selection and validation models, revealing the improvements resulting from the proposed method.

We conclude this thesis by a general conclusion and perspectives.

General Conclusion

Nowadays, Threats from the Internet have become more and more sophisticated and are able to bypass the basic security solutions such as firewalls and antivirus scanners. Additional protection is therefore needed to enhance the overall security of the network. One possible solution to improve the security is to add an intrusion detection system (IDS) as an additional layer in the security solutions.

Even though the domain of traffic classification is relatively well explored, our primary goal is to enrich existing research efforts by our own contributions. The issues considered in this work were inspired by common problems existing in real - operational networks.

In this Work, we have proposed an effective wrapper feature selection approach based on Bayesian Network classifier and applied it for network intrusion detection. In order to evaluate the performance of the selected features, a detailed comparison between the proposed approach and the other four feature selection methods (IG, GR, ReliefF and FSSS) is conducted on the NSL-KDD dataset. Experimental results illustrate that the features extracted by our approach make the BN classifier achieve higher classification accuracy than the other methods. In addition, it is conducive to cut down computing consumption for IDSs. And with feature selection, the performance of IDS is improved or at least maintained at a high level. It can detect Probe, DoS and R2L attacks with high TPR and low FPR.

Bibliography

[1] P.Wood , G. Egan, K. Haley, T. K.Tran,O. Cox, C. Wueest, and et al., "Symantec Internet Security Threat Report: Trends for 2011 (Volume 17)," Symantec Corp, April 2012. Availiable: http://bit.ly/JQNde6.

[2] Sundaram, A. An introduction to intrusion detection, Crossroads, Volume.2, Issue 4, pp. 3-7, April 1996.

[3] NSA, National Security Agency. Defence in Depth. [PDF]. [Cited 2010-11-16]. Available at: http://www.nsa.gov/ia/_files/support/defenseindepth.pdf

[4] Fogla, P., Lee, W. Evading network anomaly detection systems: formal reasoning and practical techniques, Proceedings of the 13th ACM conference on Computer and communications security, pp. 59-68, Alexandria, Virginia, USA, 2006

[5] Gates, C., Taylor, C., Challenging the anomaly detection paradigm: a provocative discussion. In Proc. of ACM Workshop on New Security Paradigms 2006, Schloss Dagstuhl, Germany, September 2006.

[6] Denning, D. E., An intrusion-detection model, IEEE Transactions on Software Engineering, Volume 13, Issue 2, pp. 222-232, February 1987

[7] Javitz, H.S., Valdes, A. The SRI IDES Statistical Anomaly Detector, In Proceedings of the IEEE Symposium on Security and Privacy, pp. 316-326, May 1991

[8] Chan, P., Mahoney, M., Arshad, M. A Machine Learning Approach to Anomaly Detection, Department of Computer Sciences, Florida Institute of Technology, Melbourne, 2003

[9] Fontugne, R., Hirotsu, T., Fukuda, K. An image processing approach to traffic anomaly detection, Proceedings of the 4th Asian Conference on Internet Engineering, pp. 17-26, November 2008, Pratunam, Bangkok, Thailand

[10] Wang, K., Stolfo, S. J. Anomalous Payload-based Intrusion Detection, Computer Science Department, Columbia University, New York, 2004

[11]. Witten I. and Frank E. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Third editions, Morgan Kaufmann, Massachusetts.

[12]. Quinlan J. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA, Massachusetts.

[13]. Chang-Tien L., Arnold P. and Prajwal M. (2005). Exploiting Efficient Data Mining Techniques to Enhance Intrusion Detection Systems. Department of Computer Science Virginia Polytechnic Institute and State University. IEEE, PP. 512-

[14]. Witten I. and Frank E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Second editions, Morgan Kaufmann, Massachusetts

[15]. Jose F. (2009). Data clustering for anomaly detection in Network Intrusion detection. Research Alliance in Math and Science, PP. 1-12.

[16]. Kayacik H., Zincir-Heywood A. and Heywood M. (2003). On the Capability of an SOM Based Intrusion Detection System. In Proc. Int. Joint Conf. Neural Network Jul, Vol. 3, PP. 1808-1813.

[17]. Hipp U. and Nakhaeizadeh G. (2000). Algorithms for Association Rule Mining: A General Survey and Comparison. Journal machine learning, Vol. 2, No.1, PP. 58-64.

[18]. Meera G., Gandhi and Srivatsa S. (2010). Adaptive Machine Learning Algorithm (AMLA) Using J48 Classifier for an NIDS Environment. Advances in Computational Sciences and Technology, Vol. 3, PP. 291–304.

[19]. Hendrickx I. and Vanden B. (2005). Hybrid Algorithms with Instance-based Classification. ILK, Tilburg University, PP. 158-169.

[20]. Quinlan J. (1986). Induction of Decision Trees. Journal of Machine Learning, Vol. 1, PP. 81-106.

[21]. Breiman L., Freidman R. Olshen and Stone C. (1984). Classification and Regression Trees. New edition , Hapman and Hall/CRC, Wadsworth Belmont.

[22]. Pfleeger C. and Pfleeger S. (2003). Security in computing. Prentice Hall.

[23]. Pearl J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, Massachusetts.

[24]. Guyon I. and Elisseeff A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research Vol. 3, PP. 1157-1182.

[25]. Liu H. and Yu L. (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering, Vol.17, No. 4, PP. 491-502.

[26]. Vidal-Naquet M. and Ullman S. (2003). Object recognition with informative features and linear classification. IEEE Conference on Computer Vision and Pattern Recognition, PP. 112-145.

[27]. Hild E., Erdogmus J. (2001). Principe, Blind Source Separation Using Renyi's Mutual Information. IEEE Signal Processing Letters, Vol. 8, PP. 174-176.

[28]. Hall M. (1999). Correlation-based Feature Subset Selection for Machine Learning. PhD, Dissertation, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

[29]. Gulshan K., Krishan K. and Monika S. (2010). An Empirical Comparative Analysis of Feature Reduction Methods for Intrusion Detection. International Journal of Information and Telecommunication Technology, Vol.1, PP. 44-51.

[30]. Alexander H. (2004). Feature Selection for Intrusion Detection: An Evolutionary Wrapper Approach. Institute for Computer Architectures, IEEE, PP. 1563-1568.

[31]. Kohavi R. and John G. (1997). Wrappers for Feature Subset Selection.Artificial Intelligence, Vol. 1, PP. 273–324.

[32]. Jain A. and Zongker D. (1997). Feature Selection: Evaluation, Application, and Small Sample Performance. IEEE Trans Pattern Analysis and Machine Intelligence, Vol. 19, PP. 153-158.

[33]. Jain A., Duin R. and Mao J. (2000). Statistical Pattern Recognition: A Review, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.22, PP. 4-37.

[34]. Das S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. In Proceedings of the Eighteenth International Conference on Machine Learning, PP. 74 -81.

[35]. Meera G., Gandhi and Srivatsa S. (2010). Adaptive Machine Learning Algorithm (AMLA) Using J48 Classifier for an NIDS Environment. Advances in Computational Sciences and Technology, Vol. 3, PP. 291–304.

[36]. Chang-Tien L., Arnold P. and Prajwal M. (2005). Exploiting Efficient Data Mining Techniques to Enhance Intrusion Detection Systems. Department of Computer Science Virginia Polytechnic Institute and State University. IEEE, PP. 512-

[37]. Witten I. and Frank E. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Third editions, Morgan Kaufmann, Massachusetts.

[38]. Jose F. (2009). Data clustering for anomaly detection in Network Intrusion detection. Research Alliance in Math and Science, PP. 1-12.

[39]. Hipp U. and Nakhaeizadeh G. (2000). Algorithms for Association Rule Mining: A General Survey and Comparison. Journal machine learning, Vol. 2,No.1, PP. 58-64.

[40]. Almuallim H. and Dietterich T. (1994). Learning Boolean Concepts in the Presence of Many Irrelevant Features. Artificial Intelligence, Vol. 2, PP. 279-305.

[41]. Blum A. and Langley P. (1997). Selection of Relevant Features and Examples in Machine Learning. Artificial Intelligence, Vol. 97, No. 1-2, PP. 245- 271.

[42]. Miller A. (1990). Subset Selection in Regression. Chapman and Hall, New York.

[43]. Pearl J. (1984). Heuristics: Intelligent Search Strategies for Computer Problem Solving. Addison-Wesley, USA.

[44]. Nigel W., Sebastian Z. and Grenville A.(2006). Evaluating Machine Learning Algorithms for Automated Network Application Identification. Centre for Advanced Internet Architectures (CAIA).Technical Report Swinburne University of Technology Melbourne .Available at: citeseer.ist.psu.edu/viewdoc/[accessed: march 15, 2011].

[45]. Yang J. and Honavar V. (1998). Feature Subset Selection Using a Genetic Algorithm. IEEE Intelligent Systems (Special Issue on Feature Transformation and Subset Selection), Vol. 13, PP. 44-49.

[46]. Dash M. and Liu H. (2003). Consistency-based Search in Feature Selection. Artificial Intelligence Vol. 151, PP. 155-176.

[47]. Fayyad U., Piatetsky-Shapiro G., and Smyth P. (1996). The KDD process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, Vol. 39, PP. 27-34.

[48]. NSL-KDD Data set for Network-based Intrusion Detection Systems. Available at: http://nsl.cs.unb.ca/NSL-KDD/ [accessed: December 15, 2010].

Abstract:

Processing huge amounts of network data is one of the largest challenges for network-based intrusion detection system (IDS). Usually these data contain lots of irrelevant or redundant features. To improve the efficiency of IDS, relevant features are necessary to be extracted from original data via feature selection approaches. In this work, an effective feature selection approach based on Bayesian Network classifier is proposed. And with the same intrusion detection benchmark dataset (NSL-KDD), the performance of the proposed approach is evaluated and compared with other commonly used feature selection methods. It is shown by empirical results that features selected by our approach have decreased the time to detect attacks and increased the classification accuracy as well as the true positive rates significantly.

Key words: Networks, intrusion detection systems, feature selection, data mining, classification, Bayesian networks, true positive.

Résumé:

Prétraiter les volumineuses données d'un trafic d'un réseau est l'un des taches les plus fastidieux pour un système de détection d'intrusions (IDS). Ces données, généralement, contiennent un nombre important d'information inutiles. Afin d'améliorer l'efficacité des IDS, seule les attributs importants doit être extraites a partir des données brutes en utilisant les approche de sélection des (attributs). Dans ce travail, une méthode efficace de sélection, basée sur l'algorithme de classification « Réseau Bayésien », a été proposée. Et, en utilisant le même ensemble de donnée référentiel (NSL-KDD), les résultats de cette méthode ont été évalués et comparés avec celles des méthodes de sélection fréquemment utilisées. Il était démontré par des résultats empiriques que la sélection des attributs par la méthode proposée a réduits le temps de réponse, a augmenté le taux de précision de classification et même le taux de vrais positifs de façon significatif.

Mots clé : réseaux ; system de détection d'intrusions, sélection des attributs, classification, réseaux Bayésien, vrais positif.

ملخص:

إن معالجة الكم الهائل من بيانات الشبكة يعد عملية مجهدة بالنسبة لأنظمة إستشعار الأختراقات الرقمية, غالبا ما تحتوي هذه البيانات على حجم كبير من المعلومات الغير مفيدة. ولذلك, و بهدف تحسين فعالية أنظمة الإستشعار الرقمية, المعلومات ذات الجودة فقط هي التي يتوجب إستخراجها من خام المعطايات بإستعمال طرق الإستخراج. في هذا العمل, تم عرض طريقة إستخراج معلومات بناء على خوارزمية الشبكات البايزيانية لتصنيف, وبإستعمال بنك المعطيات (NSL-KDD). النتائج المتحصل عليها من طريقة الإستخراج, وبعد مقارنتها مع نتائج طرق أخرى متواترة الإستعمال, أظهرت أن إستخراج المعلومات بواسطة الطريقة المقترحة يزيد من دقة إشتشعار وتصنيف البيانات وفق مقارنة بسابقاته من الطرق.

كلمات مفتاحية: بيانات الشبكات, أنظمة إستشعار الإختراقات الرقمية, طرق الإستخراج, خوارزمية, الشبكات البايزيانية, تصنيف المعطيات.