Enhanced Dependency-Based Feature Selection to Improve Anomaly Network Intrusion Detection

K. Bennaceur, Z. Sahraoui, M. A. Nacer

Enhanced dependency-based feature selection to improve anomaly network intrusion detection

K.Bennaceur[1], Z.Sahraoui[1], and M.A. Nacer[2]

[1] Ecole Militaire Polytechnique, Computer Scince Department,
Algiers, Algeria
zakaria.sahraoui@emp.mdn.dz
[2] Université des Sciences et de la Technologies Houari Boumedienne
Computer Science Department, Algiers, Algeria
anacer@mail.cerist.dz

Abstract. In daily live, online computer systems are becoming more pervasive and integrated. However, the access to the Internet can produce significant issues like cyber-attacks. The network intrusion detection system (NIDS) is a promising security solution that is used to detect attacks. It recently used Deep Learning in the detection process to obtain high performance. The performance of an NIDS depends on the used training dataset and the quality of features, where irrelevant features may decrease the detection performance, oppositely to relevant ones that are able to improve it. Feature selection is a good solution to select only relevant features to participate in the detection process. Chi-square is a supervised feature selection method that select only the most dependent features of the class feature. In this work, an Enhanced Chi-square (EChi2) method is proposed to select and weight features considering its degree of relevance. Experiments results, using the well-known NSLKDD dataset, shows that the proposed method outperforms the Chi-square.

Keywords: Feature selection · Chi-square · Enhanced Chi-square · Anomaly network intrusion detection.

1 Introduction

A network intrusion detection system (NIDS) is a security solution that becomes the most common components of every network security infrastructure [1]. It is able to detect all possible attacks. There are two main types of NIDS: 1) Misuse-based NIDS (MNIDS) that detect known attacks using the corresponding signatures, it research the existence of these latter in the traffic payload. 2) Anomaly-based NIDS (ANIDS) that is able to detect unknown or zero-day attacks. It is based on a behavioral approach by classifying traffic sessions into normal or attacks. This supervised classification is applied to a transformed traffic that is composed from a set of featured sessions named training dataset.

The anomaly network intrusion detection is a real-time process which needs to be accelerated by reducing the number of the dataset features. On the other

hand, classification performance depends strongly on the feature quality [2]. Whereas, a training dataset inevitably contains both relevant and noisy features. Rising these two challenges, feature selection becomes an essential dataset preprocess to improve classification performance, which is used for selecting the most relevant subset of features from the original high-dimensional dataset. It is widely researched in many different domains, such as choosing the causative genes in medical study [3], image segmentation in computer vision [4] and so on.

Authors of [10], have grouped feature selection methods in three groups, namely, wrapper, filter and hybrid. Wrapper methods are based on both a classifier and a classification evaluator (generally, the accuracy) to evaluate the importance of features. Using the classifier they learn different subsets datasets feature, then the classification evaluator indicates witch subset is the best. Whereas, filter methods do not use any classifier to evaluate features which are selected considering some dataset characteristic. The hybrid methods combine both of the wrapper and the filter methods to achieve better classification performance. The advantage of filter methods is that they are scalable and independent from any classifier, they are based on the description of the dataset distribution which is more suitable to study the behavior of the traffic network.

Chi-square is a supervised filter feature selection method that is able to calculate the dependency of each feature on the class feature, then remove any one that have a dependency value less than a threshold. The other ones, that are selected, are deemed to be equivalent in terms of relevance without considering the dependency value. For example, the most dependent feature to the class that is considered as the most relevant feature participate in the classification similarly like the less one feature. In this work, an enhanced Chi-square (EChi2) feature selection method is proposed to weight selected features, where weights express the relevant degree of features. It should improve classification performance as feature space will be compressed in the dimension of features with low relevance and expanded along the dimension of features with high relevance. The proposed method is implemented using useful datasets in anomaly network intrusion detection, namely NSLKDD. Then, a Deep Learning-based binary classification using a full-connected network is applied in order to classify the training dataset and detect attacks. The proposed method's efficiency assessment is based on the classification performance. Experimental results show that the proposed EChi2 outperforms the simple Chi-square.

The remainder of this paper is organized as follows: Section 2 reviews the related work. The proposed method is described in Section 3. Section 4 presents the experimental results and discussion. Finally, Section 5 presents the conclusion.

2 Related work

This section reviews all recent works that have been interested by weighted feature selection.

In [5], a combination of a Genetic Algorithm (GA) with K-Nearest Neighbor(KNN) classifier has been proposed for weighting features of each class separately from the others. But, it has been found that this separation is unable to match different class accuracies. This indicates that the weighting depends on the entire training dataset. To correct, weights have been averaged, and consequently classification accuracy has been reduced.

In [6], three different weighted feature selection methods are proposed basing on three different classifiers namely, Artificial Neural Network (ANN), Support Vector Machine (SVM), and Decision Tree (DT). The first one learns dataset using one-hidden layer ANN architecture. Since each node in the input layer represents one feature, each feature have been weighted by the sum of all weights of the corresponding node leading to the hidden layer. Selected feature are that having weight greater than a threshold. The second method is based on the Support Vector Machine Recursive Feature Elimination (SVM-RFE) [7]. In each iteration, after training the SVM classifier, features have been weighted and ranked using the guiding coefficient of the hyper-plan, The feature with the smallest weight has been eliminated. The third method has been considered as the strongest one, it has been based on the C4.5 model. A top-down tree has been constructed basing on recursive divide-and-conquer approach [8], where nodes represent features. This method has been began by selecting the top-three level nodes, then different test nodes have been generated in order to update selected feature list.

Authors of [9] have developed a new method to select and weight features, namely AGRM. The method is based on eliminating correlation, i.e. if two features are correlated, it would be better to keep only one and remove the other. Correlation between features, that should be minimized, has been modeled using a mathematical problem, the solution of the latter indicates the weight of each feature.

The weight here represents the level of influence from input feature to the first hidden layer [32]. If the value is small (nearly zero), it means that the feature is not a deciding fac- tor to pick whether a file is a malware or benign file. We will measure the average weight of all features and set a threshold value. We pick all features that have weight value higher than the threshold

3 Enhanced Chi-square feature selection method

The proposed method selects the most relevant features, then it weight them considering their degree of relevance. The more the feature is relevant the more the feature weight is greater. This way should improve classification performance because it gives more significance to the feature that supports more information. Feature weights are measured statically from the dataset distribution using the Chi-square coefficient. All weights are scaled in the range [0, 1], where the most relevant feature is weighted by 1 and all the other features 'f_i' are weighted by a scaled 'w_i'. There are several steps in this method.

Firstly, the degree of relevance of each feature 'f_i' is evaluated by calculating the dependency 'd_i' of this feature relative to the dataset class, using the Chi-square coefficient given in formula (1).

$$d^2 = \frac{N * (WZ - YX)^2}{(W + X)(W + Z)(Y + X)(Y + Z)}. \quad (1)$$

where,

W is the number of times of the co-occurrence of the feature *'f_i'* and the class *'c'*,

X is the number of times of the appearance of *'f_i'* without *'c'*,

Y is the number of times of the appearance of *'c'* without *'f_i'*,

Z is the number of times of the nonappearance neither *'c'* nor *'f_i'*,

N is the total number of dataset sessions.

This formula shows that the relevance degree measurement is based on the dataset distribution. It increases when the variation of the feature 'f_i' depends on the class *'c'*.

Then, all features are ranked based on these coefficient values, where the greatest value is associated with the most relevant feature or the most dependent on the class. Whereas, the less dependent ones that have not any influence on the class are considered as noises, because relevant features cannot be independent of the class [11]. In order to remove these features, an optimal threshold value is empirically set. If the coefficient value is lower than the threshold, the corresponding feature is removed. Features that are not removed are selected to participate in classification.

In the last, each selected features 'f_i' is weighted by a scaled 'w_i' given in formula (2). Weighting is performed by multiplying any selected feature values by the corresponding weight.

$$w_i = \frac{d_i}{d_{max}}. \quad (2)$$

Where, d_{max} is the maximum of all the d_i.

4 Results and discussion

In this section, the experimental dataset for the assessment of the proposed method is described, as well as its preparation. Afterward, it discusses the different performed experiments and the results.

4.1 Experimental dataset

Reliable and publicly available datasets is one of the fundamentals concerns of researchers and producers in intrusion detection [12]. In this work, tests are performed basing on the NSLKDD dataset [13]. It is the most used in ANID to train and check a lot of Deep Learning-based classifiers [14, 15]. It is a new version of

the old KDDcup99 dataset, it overcomes its statistical weakness that is duplicating samples in both training and testing datasets. Duplication has been removed from NSLKDD to get high actual accuracy. NSLKDD dataset size becomes reasonable, making it affordable to perform tests on the full dataset without selecting a small subset. One other advantage of NSLKDD is the separation of the testing dataset from the training one. Therefore, prediction steps are performed using the original dataset without any random dividing.

4.2 Dataset features Preparation

In the NSLKDD dataset, there are two types of features: numeric and nominal. The first type is ready for computation, but nominal features are not. To fix this problem, any nominal features are converted to numeric features by applying the method 1-to-N features proposed in [16]. This method converts each nominal feature, which varies among N values, into N binary features representing only one value. This conversion is performed using WEKA software. In fact, NSLKDD is initially featured by 42 features, but after the 1-to-N transformation, the number of features becomes 122. This is another reason to perform the feature selection.

4.3 Experimental Results and discussion

The proposed supervised weighted feature selection method is evaluated on a set of experiments. First, it is implemented using the NSLKDD dataset. Several sub-datasets result from varying the number of selected features (NSF). Then, they are classified using a Deep Learning-based full-connected network. Finally, the proposed method is compared with the Chi-square and Pearson feature selection methods considering several classification performance metrics.

Tables 1,2,3 and 4 present the comparison between the proposed method and the Chi-square method in terms of classification accuracy, precision, recall, F-score, respectively.

NSF	5	10	15	20	30	40	50	60
Pearson	91.2 %	91.5 %	92.0 %	92.8%	**92.9**%	92.3%	91.2%	89.7%
Chi-square	92.9 %	94.1 %	94.8 %	94.9 %	**95.0** %	95.0 %	94.8 %	93.2 %
EChi2	93.1 %	93.9 %	94.2 %	95.0 %	96.6 %	**97.7 %**	96.5 %	95.7 %

Table 1. Comparison between the proposed EChi2 and other methods in terms of accuracy

The classification performances without any feature selection are: 79.9% of accuracy, 88.2% of precision, 92.1% of recall, and 90.3% of F-score. So, it is noticed that all feature selection methods can boost the classification that is performed by introducing all features, including noisy ones.

NSF	5	10	15	20	30	40	50	60
Pearson	89.5 %	91.0 %	91.9%	92.2 %	**92.9**%	92.6%	92.5%	91.8%
Chi-square	90.5 %	92.6 %	91.4 %	92.5 %	**93.8 %**	92.3 %	92.1 %	91.5 %
EChi2	91.2 %	90.9 %	90.7 %	91.7 %	92.9 %	**93.7 %**	92.6 %	93.3 %

Table 2. Comparison between the proposed EChi2 and other methods in terms of precision

NSF	5	10	15	20	30	40	50	60
Pearson	96.6 %	97.3 %	97.6%	98.2%	98.5%	**98.6** %	98.1%	97.5%
Chi-square	96.9 %	96.7 %	99.8 %	98.7 %	98.9 %	**99.0 %**	98.8 %	98.1 %
EChi2	96.5 %	98.4 %	99.4 %	97.7 %	99.2 %	99.2 %	**99.5** %	99.2 %

Table 3. Comparison between the proposed EChi2 and other methods in terms of recall

NSF	5	10	15	20	30	40	50	60
Pearson	93.1 %	93.8 %	94.6%	95.3%	**95.8**%	95.1%	94.6%	94.3%
Chi-square	93.6 %	94.6 %	95.4 %	95.5 %	**96.3 %**	95.5 %	95.2 %	95.2 %
EChi2	93.7 %	94.5 %	94.8 %	94.6 %	95.9 %	**96.4 %**	95.9 %	95.2 %

Table 4. Comparison between the proposed EChi2 and other methods in terms of F-score

Concerning all classification metrics using Chi-square, Pearson and the proposed method, generally, there is firstly a slight improvement when only a few noisy features are removed (NSF$>$ 40). Then, when NSF gets the optimal value, a maximal enhancement is reached. When more features are removed (NSF$<$ 30), the classification performance starts to decrease because the dataset begins to lose relevant features that help distinguish classes.

Fig. 1 presents the comparison between the best results of both Chi-square and EChi2 in terms of the four classification metrics. Regarding the accuracy, the proposed method outperforms the simple Chi-square. This is the strong advantage of this proposed method because the accuracy metric measures the degree of closeness to the perfect classification that does not make any mistakes. Weighting selected features helps the classifier to distinguish more efficiently the classes. Concerning the recall, it is slightly improved to be near to the perfect value, and the precision is not missed.

5 Conclusion

This work proposes an enhancement of the Chi-square feature selection method to improve Deep Learning-based anomaly network intrusion detection. The proposed approach weights each selected feature considering the relevant degree. The critical point of our proposal is that it is based on the dataset distribution to improve behavioral-based intrusion detection. Experiments using the useful

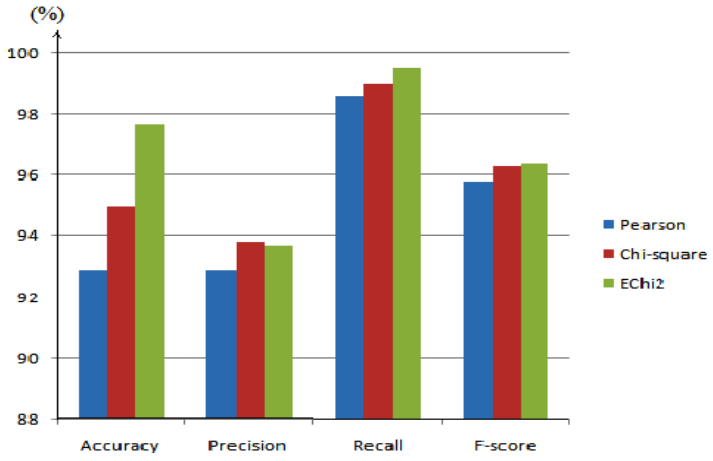

Fig. 1. Comparison between the best results of both Chi-square and the proposed EChi2.

NSLKDD dataset show that the proposed method outperforms the simple Chi-square and Pearson in terms of accuracy which is the most important classification metric. The other metrics are also considered and the classification performances are not missed. This promising method opens avenues to design a new ANID system based on new and real datasets for potential servers.

References

1. Constantinos Kolias, Georgios Kambourakis, and Manolis Maragoudakis, "Swarm intelligence in intrusion detection: A survey" in Journal of computers & security, vol. 30, n. 30, pp. 625–642, (2011).
2. Francesco Palmieri, Ugo Fiore, and Aniello Castiglione, "A distributed approach to network anomaly detection based on independent component analysis" in Concurrency and Computation: Practice and Experience, vol. 26, n. 5, pp. 1113–1129, (2014).
3. Mohua Banerjee, Sushmita Mitra, and Haider Banka, "Evolutionary rough feature selection in gene expression data" in IEEE Transactions on Systems, Man, and Cybernetics, vol. 37, n. 4, pp. 622–632, (2007).
4. Ming-Ming Cheng, Yun Liu, Qibin Hou, Jiawang Bian, Philip Torr, Shi-Min Hu, and Zhuowen Tu HFS, "Hierarchical feature selection for efficient image segmentation".
5. Melanie J Middlemiss, and Grant Dick, "Weighted feature extraction using a genetic algorithm for intrusion detection" in Congress on Evolutionary Computation, vol. 3, pp. 1669–1675, (2003).
6. Muhamad Erza Aminanto, Rakyong Choi, Harry Chandra Tanuwidjaja, Paul D Yoo, and Kwangjo Kim, "Deep abstraction and weighted feature selection for Wi-

Fi impersonation detection" in IEEE Transactions on Information Forensics and Security, vol. 13, n. 3, pp. 621–636, (2017).
7. Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik, "Gene selection for cancer classification using support vector machines" in Machine learning, vol. 46, n. 1, pp. 389–422, (2002).
8. Chotirat Ann Ratanamahatana, and Dimitrios Gunopulos, "Scaling up the naive Bayesian classifier: Using decision trees for feature selection" (2002).
9. Feiping Nie, Sheng Yang, Rui Zhang, and Xuelong Li, "A general framework for auto-weighted feature selection via global redundancy minimization" in IEEE Transactions on Image Processing, vol. 28, n. 5, pp. 2428–2438, (2018).
10. Veeran Ranganathan Balasaraswathi, Muthukumarasamy Sugumaran, and Yasir Hamid, "Feature selection techniques for intrusion detection using non-bio-inspired and bio-inspired optimization algorithms" in Journal of Communications and Information Networks, vol. 2, n. 4, pp. 107–119, (2017).
11. Girish Chandrashekar, and Ferat Sahin, "A survey on feature selection methods" in Computers & Electrical Engineering, vol. 40, n. 1, pp. 16–28, (2014).
12. Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, âĂIJToward Generating a New Intrusion Detection Dataset and Intrusion Traffic CharacterizationâĂİ, in 4th International Conference on Information Systems Security and Privacy, (2018).
13. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani, "A detailed analysis of the KDD CUP 99 data set" ,in IEEE symposium on computational intelligence for security and defense applications, pp. 1–6, (2009).
14. Mohamed Amine Ferrag, Leandros Maglaras, Sotiris Moschoyiannis, and Helge Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study" ,in Journal of Information Security and Applications, vol. 50, pp. 102419, (2020).
15. Hanan Hindy, David Brosset, Ethan Bayne, Amar Seeam, Christos Tachtatzis, Robert Atkinson, and Xavier Bellekens, "A taxonomy and survey of intrusion detection system design techniques, network threats and datasets" ,in arXiv preprint, (2018).
16. Leo Breiman, Jerome Friedman, Charles Stone, J Olshen and A Richard , "Classification and regression trees" in CRC press, (1984).