Correlation based Feature Selection for Network Inconsistency Exploration

Miss. Snehal S Mandpe[1], Dr. R. R. Keole[2]

M.E. Scholar, Computer Science and Engineering[1]
Dr. Rajendra Gode Institute of Technology and Research, Amravati[1]

Associate Professor and Head, Department of Information Technology [2]
Hanuman Vyayam Prasarak Mandal's College of Engineering and Technology, Amravati, India[2]

---***---

Abstract - Feature Selection is one of the preprocessing steps in machine learning tasks. Feature Selection is effective in reducing the dimensionality, removing irrelevant and redundant feature. Malicious activities can harm the security of the system. These activities must be avoided. Network traffic data can be monitored and analyzed by using intrusion detection system. Different data mining classification techniques are used to detect network attacks. Dimensionality reduction performs key role in the Intrusion Detection System, since detecting anomalies is time-consuming. Recently a lot of work has been done in feature selection. But, most of the authors have modified the KDD99 test dataset. Modification of training dataset is valid but modifying test dataset is against the machine learning ethics. This work comprises some of the recently proposed feature selection algorithm such as Information gain, Gain Ratio and Correlation based feature selection with the objective of determining the reduced feature set. The performance is evaluated using a combination of any two feature selection technique.

***Key Words*:** KDD99 Dataset, Anomaly Intrusion Detection, Feature Selection

1. INTRODUCTION

We live in the information-age—collecting data is easy and storing it inexpensive. In 1991 it was alleged that the amount of stored information doubles every twenty months . Unfortunately, as the amount of machine readable information increases, the ability to understand and make use of it does not keep pace with its growth. Machine learning provides tools by which large quantities of data can be automatically analyzed.

Fundamental to machine learning is *feature selection*. Feature selection, by identifying the most remarkable features for learning, focuses a learning algorithm on those aspects of the data most useful for analysis and future prediction. The hypothesis explored in this paper is that feature selection for supervised classification tasks can be accomplished on the basis of correlation between features, and that such a feature selection process can be beneficial to a variety of common machine learning algorithms. A technique for correlation-based feature selection, based on ideas from test theory, is developed and evaluated using common machine learning algorithms on a variety of natural and artificial problems. The feature selector is simple and fast to execute. It eliminates irrelevant and redundant data also in many cases, improves the performance of learning algorithms. The technique also produces results comparable with a state of the art feature selector from the literature, but requires much less computation.

Feature Selection is one of the prominent preprocessing steps in many of the machine learning applications. It is the process of reducing the feature set by choosing the relevant features from the original feature set according to an evaluation criterion and also removing the redundant features from the whole feature set.

Different feature selection methods can be broadly categorized into the wrapper model [2], the filter model [3] and the hybrid model [4]. The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets. These methods are computationally expensive for data with a large number of features. The filter model separates feature selection from classifier learning and selects feature subsets that are independent of any learning algorithm. It relies on various measures of the general characteristics of the training data such as distance, information, dependency, and consistency.

The main measure is Correlation based method. Pearson's Correlation method is used for finding the association between the continuous features and the class feature. In [5], Symmetric uncertainty measures are used for finding the association between the discrete feature and the class feature. Also, it is used for finding the feature correlation to remove redundant feature.

TABLE I
DDoS Attack Type and corresponding Reasons

DDoS Attack Type	Reasons
Extortion	Financial Profit
Hacktivism	Stealing Customer data
Political Motivation	Revenge
Security Feints	Business Competitors
Freely available tools	LOIC
Just For Fun	Fame and name amongst attacker community

Correlation method is used for finding the association between the features for sample data. If the entire population is taken, (as there is a rapid growth in data size), Correlation Coefficient may not yield goodness of result. This motivated us to use the t-test of Correlation Coefficient to examine whether the association between the features is statistically significant.

2. LITERATURE REVIEW

2.1 Multivariate Correlation Anomaly Detection

Multivariate correlation anomaly detection techniques measure the correlations between features and based on deviations of these correlations are indicative for anomalous behavior. These correlations can be measured using either linear or non-linear correlation measures. In early works, such as by Jin and Yeung in [9, 10], simple covariance matrices were used to classify correlation changes of normal and anomalous behavior. Analysis of Feature Selection Techniques . The current state-of-the-art methods use transformation functions to transform features into correlation features using e.g. the Euclidean Distance [11], geometric correlation measures based on triangle areas [4, 12] or latest, addition-based functions in [13]. This allows, compared to methods based on covariance matrices, one to flag individual instances as anomalous rather than just a group of events [4]. A comprehensive survey of more non-linear entropy methods for network anomaly detection can be found in [14].

2.2 Feature Selection Techniques

Feature selection techniques can primarily be divided into wrapper-based, filter-based and hybrid combinations of the two [5]. Wrapper-based feature selection techniques commonly use a cross-validation approach in combination with statistical classifiers like Decision Trees or Bayesian Networks to obtain the most relevant features [5]. Filter-based methods employ statistical correlation measures, where a subset of features is selected based on their ability to describe the target class best. This can be as simple as measuring the correlation between a feature and the target class using linear correlation measures like the Pearson Correlation Coefficient [15] or non-linear measures like information gain [16].
Another very frequently applied filter-based selection technique is summarizing the feature set using Principal Component Analysis (PCA). To reduce redundant data and select the features that represent most of the information (high variance), PCA eliminates linear correlations between features by transforming them into higher order components.

3. METHODOLOGY

3.1 Data processing

A tremendous challenge for the development and evaluation of network intrusion detection techniques is the lack of publicly available dataset. To date, the now 19-year-old KDD99 dataset is still the most used dataset to evaluate network intrusion detection techniques even though it has been heavily criticized over the last few years for many reasons such as:

• Outdated network traffic patterns due to evolving technologies and applications [18]
• Lack of modern, sophisticated low-footprint attacks [19, 20]
• Large number of redundant records [21].
An improved version, the NSL-KDD dataset has been published by Tavalee et al. in
[21], however, as it is based on the same KDD99 dataset records, it still lacks a modern representation of network traffic and state-of-the-art attacks.

To address these shortcoming, Nour and Slay have proposed a new dataset, termed
UNSW-NB-15 [19, 20]. This dataset is based on a modern, simulated representation of
real world traffic and furthermore includes several, modern low footprint attacks.
With that, it has addressed the biggest shortcomings of the KDD99 and the NSLKDD
datasets, yet it still has weaknesses:
• It is not based on actual real-world traffic
• It has potential simulation artefacts.
While we do acknowledge that it is very hard to create and publish a dataset based on real-world network traffic due to privacy and traffic labelling challenges, simulated

datasets very often contain simulation artefacts, which affect the meaningfulness of the results. We intend to analyze this dataset for such artefacts in future work.
Regardless of that, it is the most state-of-the-art available dataset and due to that, we are using it for our evaluations. As the dataset contains categorical and numerical features, we first had to map the categorical features to numerical values by transforming them into ordered numbers,
e.g. the protocol types *TCP*, *UDP*, *ICMP* to *TCP* = 1, *UDP* = 2, *ICMP* = 3. The dataset contains a very large number of instances and we have decided to use the first 50 k, 100 k, 150 k and 200 k instances in a 10-fold cross-validation. For the features generated by PCA as well as by the Pearson correlation coefficient, we have selected the top 20 features based on the highest variance or the highest correlation with the class label respectively. Furthermore, as conducted in [13], we have normalized the data using a statistical normalization, where the data is normalized around the mean of a feature.

3.2 PERFORMANCE EVALUATION

- **Algorithm**

Algorithm of the proposed framework

Input:Set of 41 attributes from KDD'99 Cup dataset
Output: Best selected attributes subset

Step-1: Select the attributes having a value greater than or equal to 0.7 and less than or equal to 0.7 and apply correlation-based attribute selection.
Step-2: Calculate Pearson's Correlation Coefficient using equation-1.
Step-3: Select the subset of the attributes which satisfy the threshold.
Step-4: Repeatedly apply the attributes selection from the KNN and PCA on the features obtained from step-3

For feature selection, different techniques have been used namely Correlation-based feature selection using Principle Component Analysis (PCA). All the feature selection techniques are also used as a feature selection. Initially, the Correlation-based feature selection technique is applied to the dataset. The availability of a Correlated feature in a dataset can decrease the performance of the model, and also it can affect the accuracy of the model, so these features need to be dropped from the dataset. To make a correlation matrix *Pearson's Correlation Coefficient (PCC)* is used and it can be given as follows:

$$\boldsymbol{Coefficient} = \frac{\boldsymbol{covariance}(\boldsymbol{x},\boldsymbol{y})}{\left(\boldsymbol{stdv}(\boldsymbol{x}) * \boldsymbol{stdv}(\boldsymbol{y})\right)}$$

Where x is data and y is a random variable
By using these coefficients, the relationship between the features can be understood. The value of the coefficients always ranges from -1 to 1. In the KDD dataset, the values between range -0.5 to +0.5 have shown a significant correlation. By measuring this relationship and putting those values in a matrix between each pair of values available in the dataset will form a symmetric matrix. Based on this observation, correlating features are dropped. Features having their value greater than or equal to 0.7 and less than or equal to -0.7 are dropped. The number of features before dropping the correlated features was 41, and after dropping, 28 features are left.

K-Nearest Neighbor:
This algorithm finds the nearest neighbors and differentiates them in a class. It comes under the category of the supervised algorithm. It identifies the closest neighbor by using the Euclidean distance formula. The implementation of this technique is simple to understand. Initially, separate the dataset in training and testing set, and choose important features that are required to select from the training data. Then, find the distance between all the points by using the Euclidean distance formula and store it in a list.Further, sort that list and select the first n values (number of features needed) from the dataset and then allocate a class to test the points based upon the majority of classes available in the points that have been chosen. Following are the various distance measuring techniques possible in KNN along with their standard formulas:

Euclidean Distance Function -

$$\sqrt{\sum_{i=1}^{f}(\boldsymbol{X}_i - \boldsymbol{Y}_i)^2}$$

Manhattan Distance Function-

$$\sum_{i=1}^{f}|\boldsymbol{X}_i\boldsymbol{Y}_i|$$

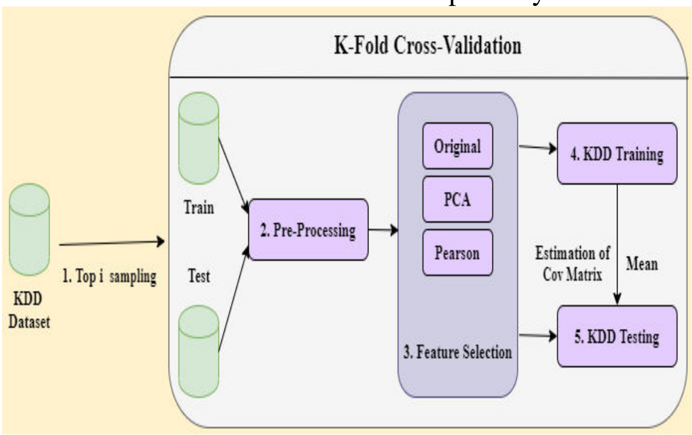
Minkowski Distance Function -

$$\left(\sum_{i=1}^{f}(|\boldsymbol{X}_i -|)^q\right)^{\frac{1}{q}}$$

Where X and Y are two distinct points and f is the number of instance points.

4 . IMPLEMENTATION

To test our theory, in comparison with the state-of-the-art anomaly identification MC system provided, I tested both PCA and Pearson Correlation empirically.



The lack of freely accessible resources is an incredible obstacle to build and test channel intrusion prevention strategies. Today, the still 19-year-old KDD 99 dataset is by far the most commonly used to test channel intrusion detection strategies, but for so many causes such as the evolution of technology and applications is stagnant channel traffic habits [18], failure of new, complex threats on low footprints [19, 20], the most obsolete reports [21], this is not based on actual traffic, and this has possible objects for simulation.

I picked the 20 leading attributes based on the highest variability or the highest association of the class label, for Principal Component Analysis generated characteristics and for Pearson correlation coefficient. Besides, I have normalized the data using the mathematical standardization as done in [13], which normalizes the data around the mean of a function. The data collection contains an integer, descriptive, and categorical characteristics.

Redundant characteristics were eliminated to guarantee that no redundant characteristics affected the rate, and false-positive rate.

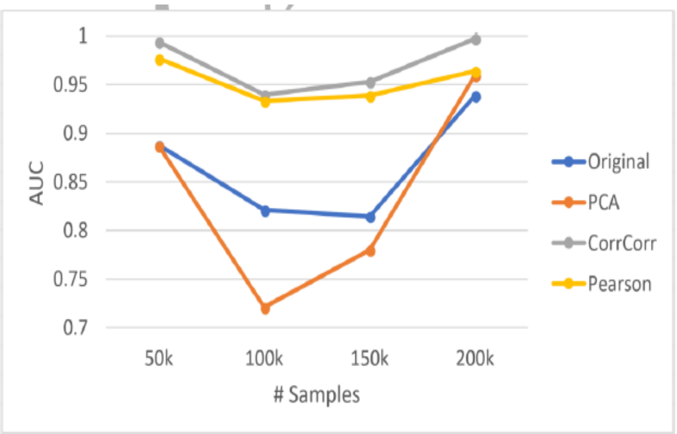
$$Accuracy = \frac{(True\ Positives + True\ Negatives)}{(True\ Positives + True\ Negatives + False\ Positives + False\ Negatives)}$$

$$Detection\ Rate\ (DR) = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positives}{(False\ Positives + True\ Negatives)}$$

5. RESULT AND DISCUSSION

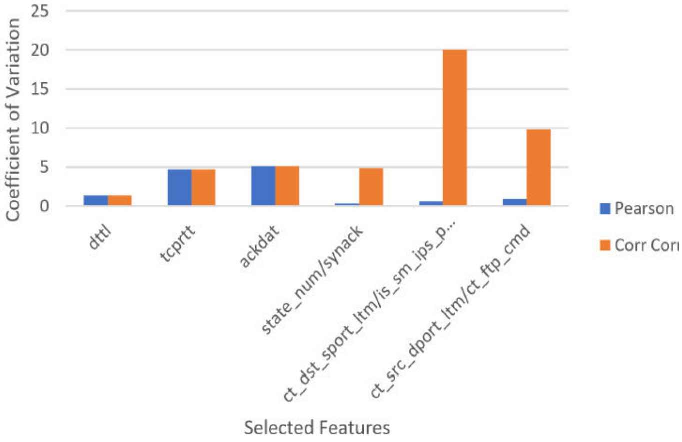
The principal component attributes and the Pearson correlation attributes are to assess CorrCorr's results in comparison to their original attributes. As seen from Figure 6, the chosen CorrCorr implementation obtained slightly improved performance, with an AUC of below or above 0.95 during the 10-fold cross-validation of all sample sizes.CorrCorr 's performance has been continuously reduced even with 100 thousand and 150 thousand samples when the performance of the main identity and PCA have reduced dramatically. CorrCorr 's output was close to the findings obtained with the Pearson correlation coefficient, but CorrCorr continuously did much better with all sample sizes.



As can be seen, the features selected by the Pearson Correlation consistently generated

the best results among all different sample sizes with an AUC as high as 0.975 for 50 k samples. On average the Pearson correlation features resulted in a ~9% performance enhancement over the original features, while the PCA components resulted in a ~3% loss of performance compared to the original features.

While the PCA results confirm the assumption that neglecting the temporal correlation changes in the selection process does not, this could also be dependent on the dataset and would need to be evaluated on different datasets. The Pearson correlation results show that a significant performance enhancement is possible if feature selection techniques are used in combination with MC-based anomaly detection methods.



In terms of protection, SYN-ACK reflects time between SYN and SYN-ACK packets for the identification of channel-based DoS/DDoS threats and the attributes chosen exclusively for Corr-Corr.

When compared with the outcomes of the KDD-data-set tests, Table IV reveals that our findings are close but marginally lower than the better results[35], and obtained the best results with a 98.65 percent accuracy. Although satisfactory results have been obtained in several other studies using approaches focused on multivariate correlation analysis, such as [9]. In these works, many datasets were used and the key emphasis was on the Distributed Denial of Service attacks (DDoS), which create a massive network imprint and thus a more readily detectable path.

6. CONCLUSION

Developments like Smart Cities will result in increasingly complex and dynamic IT infrastructure, which will lead to tremendous IT-security challenges. Multivariate correlation anomaly detection techniques are one possible method to address these challenges.

Despite it being common practice to use feature selection techniques for anomaly detection, they have very rarely been applied to MC-based techniques.

We have presented theoretical reasons why this is the case and have empirically evaluated several feature selection techniques on the state-of-the-art MC-based technique.

We have found that there is large performance improvement potential, especially using correlation-based feature selection techniques. We believe the performance can be further enhanced if the nature of MC-based techniques, temporal correlation changes among features, is considered in the feature selection process. This is what we are going to address in future work, to develop a specific feature selection algorithm for multivariate correlation-based anomaly detection techniques.

REFERENCES

[1]https://www.theguardian.com/technology/2017/nov/21/uberdata-hack-cyber-attack.

[2]Yahoo data breach 2013. Available from: https://www.reuters.com/article/us-yahoocyber/yahoo-says-all-three-billion-accounts-hacked-in-2013-data-theft-idUSKCN1C82O1.

[3]OPM data breach 2015. Available from: https://www.washingtonpost.com /news/theswitch/wp/2015/09/23/opm-now-says-more-than-five-millionfingerprints-compromised-in-breaches/.

[4] Ahmed M, Mahmood AN, Hu J. A survey of network anomaly detection techniques. Journal of Network and Computer Applications. 2016;60:19–31.

[5] Ahmed M, Mahmood AN. Network traffic analysis based on collective anomaly detection. In: Industrial electronics and applications (ICIEA), 2014 IEEE 9th Conference on. IEEE; 2014. p. 1141–1146.

[6] Sommer R, Paxson V. Outside the closed world: On using machine learning for network intrusion detection. In: Security and Privacy (SP), 2010 IEEE Symposium on. IEEE; 2010. p. 305–316.

[7] Gottwalt F, Karduck AP. SIM in light of big data. In: Innovations in Information Technology (IIT), 2015 11th International Conference on. IEEE; 2015. p. 326–331.

[8] Ye N, Emran SM, Chen Q, Vilbert S. Multivariate statistical analysis of audit trails for host-based

intrusion detection [Journal Article]. IEEE Transactions on computers. 2002;51(7):810–820.

[9] Jin, S., Yeung, D.S.: A covariance analysis model for DDoS attack detection. In: 2004 IEEE International Conference on Communications. IEEE (2004)

[10] Jin, S., Yeung, D.S., Wang, X.: Network intrusion detection in covariance feature space. Pattern Recogn. **40**(8), 2185–2197 (2007) Analysis of Feature Selection Techniques 15

[11] Tan, Z., et al.: Denial-of-service attack detection based on multivariate correlation analysis. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) Proceedings of Neural Information Processing: 18th International Conference, ICONIP 2011, Shanghai, China, November 13–17, 2011, Part III, pp. 756–765. Springer, Berlin (2011)

[12] Tan, Z., et al.: Triangle-area-based multivariate correlation analysis for effective denial-ofservice attack detection. In: 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE (2012)

[13] Li, Q., et al.: An intrusion detection system based on polynomial feature correlation analysis. In: Trustcom/BigDataSE/ICESS 2017. IEEE (2017)

[14] Nychis, G., et al.: An empirical evaluation of entropy-based traffic anomaly detection. In: Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement.

[18] Gottwalt, F., Karduck, A.P.: SIM in light of big data. In: 2015 11th International Conference on Innovations in Information Technology (IIT). IEEE (2015)

[19] Moustafa, N., Slay, J.: UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: Military Communications and Information Systems Conference (MilCIS). IEEE (2015)

[20] Moustafa, N., Slay, J.: The evaluation of network anomaly detection systems: statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. Inf. Secur. J. Glob. Perspect.

[21] Tavallaee, M., et al.: A detailed analysis of the KDD CUP 99 data set. In: Computational Intelligence for Security and Defense Applications, CISDA 2009. IEEE (2009)