sv(M)kmeans - A Hybrid Feature Selection Technique for Reducing False Positives in Network Anomaly Detection

Shubham Saini*
Undergraduate Student
shubham.saini2010@vit.ac.in

Shraey Bhatia*
Undergraduate Student
shraey.bhatia2010@vit.ac.in

I. Sumaiya Thaseen
Advisor
isumaiyathaseen@vit.ac.in

School of Computing Science and Engineering
Vellore Institute of Technology, India

ABSTRACT

Feature Selection in large multi-dimensional data sets is becoming increasingly important for several real world applications. One such application, used by network administrators, is Network Intrusion Detection. The major problem with anomaly based intrusion detection systems is high number of false positives. Motivated by such a requirement, we propose sv(M)kmeans: a two step hybrid feature selection technique. The proposed technique applies classification on false-positives and true positives; and on false-positives and true-negatives after an initial round of clustering. Specifically, SVM-RFE is applied on the results obtains from MK-Means. sv(M)kmeans is evaluated for its real world applicability using the benchmark NSL-KDD data set. We show the feature subset to significantly reduce the false positive rate and increase the accuracy in network anomaly detection.

Categories and Subject Descriptors

I.5.2 [**PATTERN RECOGNITION**]: Design Methodology—*Feature evaluation and selection*

Keywords

feature selection, false-positive, intrusion detection

1. INTRODUCTION

Feature selection is an active area of research in the domains that require the use of data-sets with tens or hundreds of features. In his work "Statistical Modeling: The Two Cultures" [1], Breiman refers to the Rashomon Effect derived from a Japanese movie in which four different witnesses, from different vantage points, report the same facts in court but with different stories of what happened. The same is true for features in a data-set. The objective of feature selection is to extract a faster and more relevant feature subset that gives a better understanding of the data generation process, with little or no reduction in the overall predictive performance.

With the use of advanced data mining techniques, anomaly based intrusion detection systems (IDS) are able to prevent intrusions with little human intervention. However, high number of false positives wastes a lot of time and effort of network administrators.

False Positives or False Alarms are legitimate activities that IDS detects as malicious. The primary role of IDS is to detect a substantial percentage of intrusions, with false positive rate at an acceptable level. In real environments IDS throw an abundance of alerts, most of them being false positives. High number of false positives wastes a lot of time and effort of network administrators.

Most clustering and classification algorithms suffer from problems such as computational complexity, over-fitting etc. Ensemble learning is a popular approach to reduce the effect of these common problems. Ensemble approaches work on a simple intuitive idea of a group of experts working together on a problem. A group of experts with diverse experience in solving a problem will have a higher probability of arriving at an acceptable solution than a single expert. We used the same idea to experiment with a hybrid approach for feature selection using two distinct approaches.

To address these problems, we present sv(M)kmeans: a two step hybrid feature selection technique. The proposed technique applies classification on false-positives and true positives; and on false-positives and true-negatives after an initial round of clustering.

The key contributions of this paper are:

- sv(M)kmeans helps in reducing the false positive rate in anomaly based network intrusion detection systems.
- The reduced feature subset obtained through sv(M)kmeans helps in faster execution of MK-Means algorithm.
- Validation of the proposed technique using the benchmark NSL-KDD data set.

2. BACKGROUND AND RELATED WORK

Several studies involving hybrid clustering and classification approaches have been conducted previously. Vibha et al.[8] proposed a combined classification based on clustering for multispectral LANDSAT images for soil mining. In the

*This work was completed while the authors were full-time undergraduate students.

The 20th International Conference on Management of Data (COMAD),
17th-19st Dec, 2014 at Hyderabad, India.

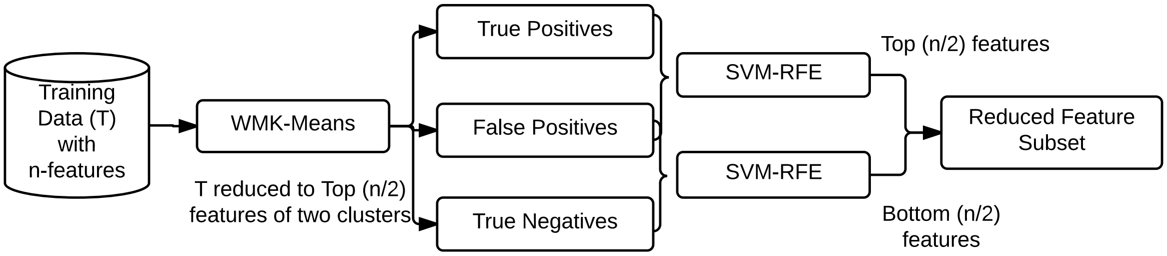

Figure 1: sv(M)kmeans: A two step hybrid feature selection technique

context of intrusion detection, Singh et al.[6] used a hybrid approach with a collective goal of reducing false positive rate and detection of previously unknown attacks. More recently, Li et al. [4] presented a clustering, ant-colony and SVM hybrid to incrementally remove features for intrusion detection. Ravale [5] proposed a similar K-Means - SVM hybrid feature selection by training a SVM classifier on the results of clustering.

We propose a two step hybrid methodology making use of MK-Means[2] and SVM-RFE[3] feature selection algorithms that helps to reduce the false positive rate in an IDS. Our work is significantly different from previous studies in several ways. The primary objective of our work is substantial reduction in the false positive rate of intrusion detection systems. Algorithms used to reduce false positive rate generally increase the false negative rate, there by having little improvement in the accuracy. sv(M)kmeans is experimentally shown to have little effect on the false negative rate, hence improving the overall accuracy.

2.1 MK-Means

K-Means is a machine learning algorithm used to partition a data-set into K clusters. The objects assigned to the resulting clusters have the smallest distance with respect to the centroid of their clusters. The Euclidean distance is the most popular measure used with K-Means. A more general measure is the Minkowski distance which can be considered as a generalization of both the Euclidean distance and the Manhattan distance.

$$d(x,y) = (\sum_{v=1}^{V} +|x_v - y_v|^p)^{1/p} \tag{1}$$

Amorim et al. makes use of the Minkowski measure to calculate the feature weights for each cluster. In order to make use of the calculated feature weights, the K-Means criteria is modified as:

$$W_p(S,C,w) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v=1}^{V} w_{kv}^{p} |y_{iv} - c_{kv}|^p \tag{2}$$

Feature weights are calculated using EQ (3). This follows from the idea that features with a small relative dispersion within a cluster are given a higher weight.

$$w_{kv} = \frac{1}{\sum_{u \in V} [D_{kv}/D_{Ku}]^{1/(p-1)}} \tag{3}$$

2.2 SVM-RFE

Support Vector Machines - Recursive Feature Elimination gives a ranking of features for a problem trained by SVM using a linear kernel function. After each iteration the feature with lowest ranking criteria is removed. SVM-RFE uses the weight magnitude as the ranking criterion.

1. Train an SVM on the training set
2. Order features using weights of the resulting classifier
3. Eliminate features with the smallest weight
4. Repeat the process with the training set restricted to the remaining features

3. SV(M)KMEANS

We now present sv(M)kmeans - our two step hybrid feature selection technique. Figure 1 gives a summary of the feature selection process. In order to reduce occurrence of false positives, we use a two step hybrid methodology making use of MK-Means and SVM-RFE feature selection algorithms. Algorithm 1 illustrated below is used to obtain a feature subset that we show to be better predictors and help reducing the false positive rate.

Algorithm 1 Feature selection using MK-Means and SVM-RFE

```
 1: GIVEN Dataset $D = \{(x_i, c_i)|x_i \in R^n, c_i \in \{0,1\}\}$
 2: Features F = 1,2,...n
 3: Take Training Data $T^f$ where $T \subset D$ and $f = F$
 4: $[U, W] \leftarrow MKMeans(T^f, 2, Beta, p)$
 5: $F_2 \leftarrow$ Top (n/2) features of $W_1 \cup$ Top (n/2) features of $W_2$
 6: Compare U and $c_i$
 7: $T_{FP}^{F_2} \leftarrow \{(xi, ci)|xi \in false - positives\}$
 8: $T_{TP}^{F_2} \leftarrow \{(xi, ci)|xi \in true - positives\}$
 9: $T_{TN}^{F_2} \leftarrow \{(xi, ci)|xi \in true - negatives\}$
10: $F_{FP-TP} \leftarrow SVM - RFE(T_{FP}^{F_2}, T_{TP}^{F_2})$
11: $F_{FP-TN} \leftarrow SVM - RFE(T_{FP}^{F_2}, T_{TN}^{F_2})$
12: $F_{final} \leftarrow$ Top (n/2) features of $F_{FP-TP} \cup$ Bottom (n/2) features of $F_{FP-TN}$
```

The algorithm requires a labeled data-set D, with class labels 0 and 1 for normal and anomaly records respectively. Each record in the data-set is a n-dimensional real vector. In line 3 we take a subset of data set as training data. We apply MK-Means function as given in Line 4 on the n-dimensional

Table 1: Summary of the Performance Metrics with Generic Anomalies

Stage	#Features	Avg FPR(%)	Avg Accuracy(%)	Min FPR(%)	Max Accuracy(%)	Execution Time (s)
Initial	41	15.40	83.29	12.90	83.59	5
After MK-Means	39	16.29	85.19	12.32	83.92	5
After SVM-RFE	31	4.10	85.90	3.92	86.99	3

Table 2: Summary of the Performance Metrics with Neptune Anomaly

Stage	#Features	Avg FPR(%)	Avg Accuracy(%)	Min FPR(%)	Max Accuracy(%)	Execution Time (s)
Initial	41	5.88	96.10	4.01	97.99	4
After MK-Means	40	5.76	96.17	4.03	97.78	4
After SVM-RFE	36	3.36	97.46	2.24	98.67	3

data T, specifying the number of clusters (K) the data has to be divided into, the weight exponent (Beta) and the distance exponent (p). The result is the clustering detail (U), and weights of all features in each cluster (W). In Line 5, we find out top (n/2) features of each cluster and their union is stored as reduced feature subset (F_2). The training data is reduced to the dimensionality of F_2. In line 6, we find out the records which are detected as false positives, false negatives, true positives and true negatives from U. In lines 7-9, we store the records detected as false positives, true positives and true negatives respectively. In lines 10-11, we run SVM-RFE to find the strongest features that distinguish false positive from true positives and false positive from true negative respectively. Finally, we take the union of top (n/2) from the result of SVM-RFE on false positive and true positive and bottom (n/2) from the result of SVM-RFE on false positive and true negative. The rationale behind this being that we obtain the features that strongly differentiate false positives from true positives, and at the same time remove the features that differentiate false positives from true negatives.

4. EXPERIMENTAL SETUP

4.1 Data Set

KDDCUP'99 is the mostly widely used data set for anomaly detection. To solve some of its inherent problems Tavallaee et al.[7] suggested the NSL-KDD data-set. The NSL-KDD data set has several advantages over the original KDD data set such as no redundant records in the train and test sets, the number of selected records from each difficulty level group inversely proportional to the percentage of records in the original KDD data set. Table 3 gives a summary of the data set. The NSL-KDD data includes 41 features and 5 classes; normal and 4 types of attacks: Dos, Probe, R2L and U2R. We measure the performance of our proposed technique using NSL-KDD data-set.

4.2 Analytical Implementation

Experiments were done using R language for statistical computing version 3.0 on Linux 64-bit platform running on Intel Core i3 processor 2.4Ghz with 3GB RAM. Amorim has provided MATLAB implementation of MK-Means algorithm [1] which was ported to R language for experimentation.

Class	Attack Type	Number of Instances	
		Training	Test
Normal	-	67343	9711
DOS	Back, Land, Neptune, Pod, Smurf, Teardrop, Mailbomb, Processtable, Udpstorm, Apache2, Worm	45927	7456
Probe	Satan, IPsweep, Nmap, Portsweep, Mscan, Saint	11656	2421
R2L	Guess_password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named	995	2756
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps	52	200

Table 3: Overview of the NSL-KDD data set

R implementation of SVM-RFE is available at the Data Mining group webpage [2] of Universidad Catlica de Crdoba. A hybrid MK-Means SVM-RFE algorithm was implemented by making use of the aforementioned pieces of codes.

Before applying the proposed algorithm to the dataset all attributes were normalized to the range 0-1.

5. RESULTS

Table 1 summarizes the performance metrics for data containing 25000 normal and 25000 generic anomaly class records. Further, Table 2 summarizes the test performed using 25000 records of specific anomaly type (Neptune class) with 25000 normal class records. Initially we start with 41 features in

[1] http://renatocamorim.com/2012/11/21/minkowski-weighted-k-means/

[2] http://www.uccor.edu.ar/paginas/seminarios/Software/SVM_RFE_R_implementation.pdf

the data-set, which are reduced to F_2 after running MK-Means. Finally, we have the feature subset F_{final} obtained using SVM-RFE. After each stage, we find the performance metrics by running MK-Means clustering algorithm on the reduced feature subset.

The results of MK-Means depend on the initial cluster centroids and number of centroids. We therefore ran each test 20 times. The average and best results are presented. It is evident from the experimental results that applying just MK-Means for feature selection did not yield a significant improvement. It was after the application of SVM-RFE that we notice reduction in false positive rate. It may also be noted that the decrease in false positives was accompanied by an increase in the overall prediction accuracy; thereby having little ill-effect on the false negative rate.

The computation time at each iteration is proportional to the product of the number of entities and the dimension of the feature space. The execution time for running the MK-Means algorithm is presented (rounded to the nearest integer). As evident from the table, the reduced feature space helps in reduction of the execution of the MK-Means algorithm, hence an added advantage.

6. CONCLUSION

In this paper, we presented sv(M)kmeans: a two step hybrid methodology for feature selection that allows for reduction in false positive rate in anomaly based network intrusion detection systems. It is evident from our experiments that the sv(M)kmeans is effective in reducing the false positive rate; from 12.9% to 3.92% in case of generic anomaly data and from 4.01% to 2.34% in case of specific anomaly data for the best case scenario. Interestingly, the technique reduced the false positive rate along with an increase in the overall prediction accuracy, thus having little effect on the false negative rate. This aspect needs further analysis to establish if it is true across diverse data sets. Apart from increasing the prediction accuracy, the reduced feature subset also helps reducing the execution time of MK-Means algorithm.

7. REFERENCES

[1] L. Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[2] R. Cordeiro de Amorim and B. Mirkin. Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *Pattern Recognition*, 45(3):1061–1075, 2012.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[4] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, and K. Dai. An efficient intrusion detection system based on support vector machines and gradually feature removal method. *Expert Systems with Applications*, 39(1):424 – 430, 2012.

[5] U. Ravale, N. Marathe, and P. Padiya. Attribute reduction based hybrid anomaly intrusion detection using k-means and svm classifier. *International Journal of Computer Applications*, 82(15):32–35, November 2013.

[6] V. Singh and P. Kaur. Adaptive distributed intrusion detection using hybrid k-means svm algorithm.

[7] M. Tavallaee, E. Bagheri, W. Lu, and A.-A. Ghorbani. A detailed analysis of the kdd cup 99 data set. In *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009*, 2009.

[8] L. Vibha, G. HarshaVardhan, S. Prashanth, P. Deepa Shenoy, K. Venugopal, and L. Patnaik. A hybrid clustering and classification technique for soil data mining. In *Information and Communication Technology in Electrical Sciences (ICTES 2007), 2007. ICTES. IET-UK International Conference on*, pages 1090–1095. IET, 2007.