Intrusion Detection System Based on Modified K-means and Multi-level Support Vector Machines

Wathiq Laftah Al-Yaseen[1,2], Zulaiha Ali Othman[1], and Mohd Zakree Ahmad Nazri[1]

[1] Data Mining and Optimization Research Group (DMO)
Centre for Artificial Intelligence Technology (CAIT)
School of Computer Science, Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia (UKM), 43600 Bandar Baru Bangi, Malaysia
[2] Al-Furat Al-Awsat Technical University
wathiqpro@gmail.com, {zao,zakree}@ukm.edu.my

Abstract. This paper proposed a multi-level model for intrusion detection that combines the two techniques of modified K-means and support vector machine (SVM). Modified K-means is used to reduce the number of instances in a training data set and to construct new training data sets with high-quality instances. The new, high-quality training data sets are then utilized to train SVM classifiers. Consequently, the multi-level SVMs are employed to classify the testing data sets with high performance. The well-known KDD Cup 1999 data set is used to evaluate the proposed system; 10% KDD is applied for training, and corrected KDD is utilized intesting. The experiments demonstrate that the proposed model effectively detects attacks in the DoS, R2L, and U2R categories. It also exhibits a maximum overall accuracy of 95.71%.

Keywords: intrusion detection system, network security, support vector machine, K-means, multi-level SVM.

1 Introduction

Intrusion detection systems (IDS) limit the serious influence of attacks on system resources. They are used as tools behind firewalls to identify suspicious patterns by monitoring and analyzing the events in a computer network. IDS is classified as either a signature or an anomaly detection system [1]. Signature detection systems (misuse detection systems) aim to determine the defined patterns or signatures of attacks in traffic networks. These systems identify known attacks efficiently but fail to detect new attacks (zero-day attacks) whose signatures have not been saved previously in the database. By contrast, anomaly detection systems identify new attacks by learning the normal behavior of the system and then generating an alarm in the event of a deviation from the normal behavior. This deviation is considered an intrusion [2].Therefore, anomaly detection systems report higher false alarm rates than signature detection systems do.

In many approaches, anomaly detection systems are implemented with different techniques to improve IDS accuracy, as discussed in the subsequent section. A popular technique used with IDS is the support vector machine (SVM). This technique has

M.W. Berry et al. (Eds.): SCDS 2015, CCIS 545, pp. 265–274, 2015.
DOI: 10.1007/978-981-287-936-3_25

satisfactorily classified data, particularly in conjunction with IDS. Nonetheless, these results rely heavily on the quality of the training data set used to train SVMs. If the training data set is large, then the training complexity of SVM is high. This occurrence may cause system failure because of the high consumption of memory [3]. Given that the majority of training data sets for IDS is large, including the KDD Cup 1999 data set, these defects must be addressed by reducing the number of instances in training data sets. Some researchers have removed redundant instances from data sets as a preprocessing step as in [4, 5], whereas others have used techniques to reduce the size of training data sets, as in [3].

In the present study, we propose a model that utilizes a modified K-means algorithm at the preprocessing stage to reduce the number of instances for training data sets. This model also uses SVM as a multi-level classifier to build an anomaly intrusion detection system that can detect unknown attacks. We select the K-means algorithm because of its capability to cluster instances into highly similar groups. We employ this algorithm to generate a new training data set that represents all instances in the original training data set by improving the method of selecting the initial centroids of clusters that represent all cases. First, a training data set is separated during preprocessing into five categories: Normal, DoS, Probe, R2L, and U2R. Then, the number of instances in each category is reduced through modified K-means while maintaining the high quality of the categories for the training data set. The resultant five categories of the data set are then employed to learn multi-level SVMs. The proposed model can reduce training time and achieve a favorable detection performance as a result of IDS. The remainder of this paper is organized as follows. Section 2 provides an overview of the K-means algorithm and the SVM classifier. Section 3 describes the proposed system. Section 4 presents the experimental results. Finally, Section 5 provides the conclusion.

2 Related Work

Many machine learning and data mining techniques have recently been proposed to design IDS models that can detect known and unknown attacks. However, the detection and false alarm rates of an anomaly intrusion detection system remain poor. Some of these models combine two or more techniques to improve accuracy. In the current study, we review previous studies related to the selection of the initial centroids of clusters for K-means and the studies that use multi-levels to implement classifiers for IDS. All of the following studies employ the KDD Cup 1999 data set to evaluate performance.

The K-means algorithm is highly sensitive to the initial centroids of clusters. In fact, many studies seek to improve the method of selecting the optimal initial centroids of clusters. These centroids effectively separate clusters and accelerate their convergence behavior. The initialization methods for K-means were investigated comparatively by Celebi et al. [6]. Gao and Wang [7] identified the initial center of clusters as instances with the least similar degree of information entropy. Sujatha and Sona [8] proposed the initial method to enhance K-means, in which the sensitivity of local minima and the randomness for K-means are reduced. However, this method has a long processing time. Kathiresan and Sumathi [9] utilized the Z-score ranking method to select improved initial centroids for K-means. Nonetheless, the complexity time is long given large data sets. Nazeer and Sebastian [10] proposed an iterative process to select initial centroids in which the distance of each data point from all other data points must be calculated. Therefore, a large set of data points requires much computation.

Multi-level models were successfully used to construct IDS and to improve detection accuracy. Pfahringer [11] presented the bagged boosting of C5 as a model for IDS. Xiang et al. [12] proposed multiple tree classifier models that employ the C4.5 technique at each level. The DoS, Probe, and Normal categories were classified at the first level, whereas R2L and U2R were classified at the second level. In 2008, Xiang et al. [13] presented another multi-level hybrid classifier that combines decision trees and Bayesian clustering. The C4.5 model was used to extract DoS and Probe attacks. Then, Bayesian clustering (AutoClass technique) was employed to cluster the R2L, U2R, and Normal categories. The largest cluster represents the Normal classes, whereas the other clusters denote R2L and U2R attacks. The AdaBoost algorithm with a single weak classifier was proposed by Natesan et al. [14] to build IDS. The classifiers used in this algorithm are Bayes Net, Naïve Bayes, and decision trees. Ambwani [15] presented the multi-class SVM that uses the one-versus-one method to classify each attack. Nonetheless, the proposed method is no better than the winner method established in [11]. Ambwani [16] also presented a model that uses neural network and fuzzy theory to reduce the high rate of false positive alarms. This study analyzes the advantages and disadvantages of neural network and fuzzy logic. It then generates a new model with enhanced generalization, learning, and mapping capability. Lu and Xu [17] proposed a three-level hybrid IDS that combines supervised classifiers such as C4.5 and Naïve Bayes with unsupervised clustering (i.e., Bayesian clustering) in different levels to classify various classes. In the first level, the C4.5 algorithm was used to separate the data set into three categories: DoS, Probe, and Others. Naïve Bayes was used to distinguish the U2R category from the other categories in the second level. In the third level, Bayesian clustering separated the category R2L from Normal with high detection. Finally, Gogoi et al. [18] proposed a multi-level hybrid intrusion detection system that combines supervised, unsupervised, and outlier methods to improve detection rate. The proposed method classified the DoS and Probe categories at the first level using the CatSub+ supervised classifier. In the second level, the unsupervised classifier K-point algorithm was applied to distinguish the Normal category from the rest of the test data set. In the final level, the remaining data were grouped into R2L and U2R using the outlier-based classifier GBBK.

In summary, all studies that employ K-means for IDS attempt to improve performance by enhancing the method of selecting the initial centroids of clusters. However, these methods are flawed in terms of the increased complexity of processing time and the fact that each resultant cluster retains many instances from different classes. Moreover, choosing the best sequence with which to classify classes with high accuracy remains difficult for the multi-level classifier model.

3 Proposed Modified K-means and Multi-level SVMs Model

The proposed model that combines modified K-means with multi-level SVMs is described in this section. We thus summarize its steps as follows:

- The training data set is examined (10% KDD data set).
- The symbolic attributes *protocol type*, *service,* and *flag* are converted into numeric types, as in [19].
- The training data set is normalized to [0, 1], as demonstrated in [19].
- The training data set is divided into five categories (Normal, DoS, Probe, R2L, and U2R).
- Modified K-means is applied to each category to generate five new training data sets.
- Each SVM is trained with one of the new training data sets.
- The first three steps are repeated to test the data set (corrected KDD data set).
- The multi-level SVMs in Fig. 1 are applied to classify the instances of testing the data set.
- The performance of the model is assessed in terms of accuracy, detection rate, and measures of false alarm rate.

Before training with SVM, the training data should be preprocessed, such as by converting and normalizing attributes. Then, the training data set is divided into the Normal, DoS, Probe, R2L, and U2R categories. Modified K-means is applied to each category to reduce the number of instances by clustering and by computing the average of each cluster as a new instance. For example, the result is a set of clusters with similar instances when modified K-means is implemented in the Normal category. Thus, the instances of the new Normal category are represented by computing the average of each cluster as a new instance. Table 1 shows the number of instances of the 10% KDD Cup 1999 data set before and after this stage. The quality of the resultant instances represents that of all of the instances in the original training data set.

Table 1. Number of instances in the 10% KDD Cup 1999 data set before and after categorization and applying modified K-means.

Category	# of instances (before)	# of instances (after)
Normal	97,278	639
DoS	391,458	140
Probe	4,107	134
R2L	1,126	51
U2R	52	25
Total	494,021	989

The K-means algorithm depends on two factors, namely, the number of clusters and the initial centroids of clusters, to optimize the clustering of instances [20]. The details and pseudo-code of standard K-means are shown in [21]. Our modified K-means must specify these two factors to identify a threshold value as the maximum distance between the centroid of clusters and the instances of the data set. Algorithm 1 shows the steps of the modified K-means algorithm. The number of clusters k is computed dynamically without requiring the user's input (steps 1 and 2), unlike in the standard algorithm. The modified algorithm computes the initial centroids of clusters by searching for all of the instances in a data set with distances that are larger than the threshold, as indicated in steps 1 and 2 of Algorithm 1, whereas the standard algorithm generates these instances randomly. Accordingly, the differences between the modified and standard K-means are presented in steps 1 and 2 of Algorithm 1.

Algorithm 1. Modified K-means algorithm

```
Input: Whole instances of category D
Output: High quality instances of category D′
Step 1. Set $k = 1$, $c_1$ = First instance $\omega_1 \in D$
Step 2. For every instance $\omega_I \in D$ and $i \neq 1$ Do
Step 2.1. If $\|\omega_i - c_s\| > threshold, s = 1, \ldots, k$ Then
Step 2.2. $k = k + 1$, $c_k = \omega_i$
Step 3. Assign every instance $\omega_I \in D$ to closest centroid in order to make k Clusters $\{C_1, C_2, \ldots, C_k\}$
Step 4.Calculate cluster centroids $\overline{\omega_\iota} = \frac{1}{k_i}\sum_{j=1}^{k_i} \omega_{ij}, i = 1, \ldots, k$
Step 5. For every instance $\omega_I \in D$   Do
Step 5.1. Reassign $\omega_i$ to closest cluster centroid; $\omega_i \in C_s$ is moved from $C_s$ to $C_t$
          If $\|\omega_i - \overline{\omega_t}\| \leq \|\omega_i - \overline{\omega_j}\|$ $for\ all\ j = 1, \ldots, k, j \neq s.$
Step 5.2. Recalculate centroids for clusters $C_s$ and $C_t$.
Step 6. If cluster instances are stabilized Then (D′ = centroids of clusters) Else go to
          Step 4.
```

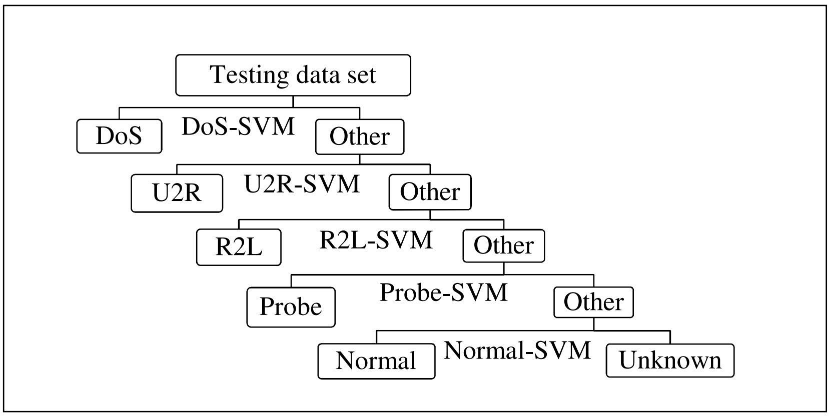

Fig. 1. Multi-level classification testing data set using SVM

Therefore, the proposed system generates five new training data sets from these categories. The first is the normal training data set, which considers normal instances as class 1 and the other instances of other categories are class 2. The same steps are repeated with the other categories. Finally, the system uses these new training data sets to learn five SVM classifiers that utilize different parameters to improve the performance of IDS. These SVMs are Normal-SVM, DoS-SVM, Probe-SVM, R2L-SVM, and U2R-SVM.

In the testing phase, the converted and normalized data are preprocessed to test the data set. Then, the multi-level classification depicted in Fig. 1 is applied. Previous studies [13, 17, 18] have proposed different multi-level methods to implement classifiers, as discussed in the section on Related Works. The ideal multi-level classification of the applied SVMs for testing a data set is derived from the results of several experiments, as exhibited in Fig. 1. DoS-SVM is implemented first because the classes of DoS have been less similarity to other categories given that this kind of intruder does not use the legitimate behavior of a user during attacks. In the subsequent levels, U2R, R2L, and Probe are classified according to the amount of instances for each category. U2R attacks involve the fewest instances in comparison with the other attacks. While the risk of these attacks is significant, U2R attacks are considered the most dangerous. In addition, the reason behind extracting these categories before normal is the similarity factor among their instances and those of the Normal category. The final SVM applied to the proposed system is Normal-SVM, which separates normal instances from the remaining instances. The remaining instances that are not classified under any category are considered unknown attacks.

4 Experimental Results

To ensure experimental persuasiveness and convenience, the proposed system uses the KDD Cup 1999 data sets as benchmarks to evaluate the experiments. These data sets originated from the Lincoln Laboratory of the Massachusetts Institute of Technology. They were developed by DARPA and are considered standard benchmarks for the evaluation of intrusion detection systems. The training and testing data sets of KDD Cup 1999 contain 4,898,431 and 311,029 instances, respectively. All instances of these data sets fall into the five main categories Normal, DoS, Probe, R2L, and U2R. The training data set contains 22 types of attacks in addition to those in the normal class, whereas the testing data set contains only an additional 17 types of attacks. Each instance in the data set displays41 continuous and discrete features [13, 22].

In this experiment, we use the 10% KDD data set for training. This data set contains 494,021 instances. The corrected KDD data set is utilized for testing and contains 311,029 instances. A computer that runs on an Intel Core i5 processor with 2.60 GHz and 12 GB RAM is employed. The freeware package LibSVM [23] is coded using Java to implement the proposed system. We apply nu-SVC and RBF kernels to run the LibSVM in this study, and the ideal values of the parameters nu and gamma are determined for each category as per the results of several experiments, as listed in Table 2. The threshold value to reduce the number of instances for all categories using modified K-means is 0.5, as indicated in Table 1. The popular measures of intrusion detection systems, such as accuracy, detection rate (recall), and false alarm rate are used to evaluate system performance.

Table 2. Parameter values of the nu-SVC classifier

Category	nu	gamma (γ)
Normal	0.06	0.09
DoS	0.004	0.5
Probe	0.1	0.3
R2L	0.05	0.008
U2R	0.05	0.008

The best performance of this system in terms of accuracy is 95.71%, that of detection rate is 95.02%, and that of false alarm rate is 1.45%. The details of the results are shown in the confusion matrix of Table 3.

The detection rates of R2L and U2R are minimal in comparison with those of DoS and Probe because the number of instances in these attacks is much less than that in DoS and Probe given the KDD Cup 1999 data set. Concurrently, two types of attacks are found in R2L:7,741 snmp get attack and 2,406 snmp guess. Their features are highly similar to those of Normal and may match these features100%. Hence, the predicate number of R2L as Normal is high. The proposed model with combined standard K-means is initially compared with multi-level SVMs to highlight the capability of the modified K-means to build a new training data set with high-quality instances. To compute the results of standard K-means, we must identify the best number of clusters. The ideal value of k is 90, at which the accuracy is high at 91.65%. Therefore, we can compare the results of the proposed model with those of the combined standard K-means and the multi-level SVMs when k is equal to 90. The accuracy, detection rate, and false alarm rate of the proposed method are generally enhanced under the combined standard K-means with multi-level SVM, with the exception of the detection rate of U2R.

Table 3. Confusion matrix for the proposed system with 10%KDD for training and corrected KDD for testing

		Predicate						Total	Re-call
		Normal	DoS	Probe	R2L	U2R	Unknown		
Actual	Normal	59714	84	116	255	7	417	60593	98.55
	DoS	722	223347	107	148	0	5529	229853	99.57
	Probe	598	193	2885	3	0	487	4166	80.94
	R2L	11060	1	1	1603	6	3518	16189	31.63
	U2R	101	0	74	16	26	11	228	16.23
	Total	72195	223625	3183	2025	39	9962	311029	

Table 4. Comparison ofthe proposed method and the combined standard K-means with multi-level SVMs

Method	Normal	DoS	Probe	R2L	U2R	Accuracy	FAR
Standard K-means with multi-level SVMs	88.93	96.57	69.14	6.26	**52.85**	91.65	11.07
Proposed method	**98.55**	**99.57**	**80.94**	**31.63**	16.23	**95.71**	**1.45**

Therefore, we compare the performance of the proposed model with that of other methods, such as the bagged boosted (Winner's) [11], multi-class SVM [15], neuro-fuzzy controller(NFC; artificial neural network and fuzzy) [16], adaptive importance sampling (AIS),multi-object genetic fuzzy IDS (MOGFIDS; GA and fuzzy) [24], balance iterative reducing and clustering using hierarchies (BIRCH), and SVMs [3], as depicted in Table 5.

The proposed method is the most accurate overall among the other methods and reports the best detection rates for DoS and R2Lattacks. The multi-class SVM has the best detection rate for the Normal class, whereas those for the other categories are worse than the rates obtained with other methods. Hierarchical BIRCH and SVM achieve high detection rates for attacks in Probe and U2R.Moreover, the accuracy and detection rates of Normal and DoS are moderate and are close to the results of the proposed model. Therefore, the detection rate of Probe decreases with the increase in the Normal or DoS detection rates. The reason of detection rate of Normal category is small compared with the other methods due to the level of Normal-SVM in multi-level model is the last one as depicted in Fig. 1.For instance, when change the level of Normal-SVM to the first level, then the detection rate of Normal will be increased, but this change will effect on the performance of the proposed model with the other categories like R2L and U2R. However, the detection rate of the proposed method for the attacks in Probe category is less than other methods because there is a type of new attack called MScan belong to Probe has a low detection rate with SVM classifier. Consequently, the overall detection rate of Probe with SVM is small. We believe that the proposed method generates the best results in relation to the balance state among all of the categories. As a result, its accuracy exceeds those of other methods. Several methods are employed to evaluate the proposed IDS, such as 10-fold cross validation or the application of the same data set for training and testing. Some methods also utilize data sets that are generated randomly from the original KDD Cup 1999 data set. Hence, performance is high. Consequently, any proposed method for IDS should be compared according to the same evaluation method. Thus, we use the methods in Table 5 only for comparison given that the best evaluation method for IDS involves training and testing the KDD Cup 1999 data sets.

Table 5. Comparison with other methods in terms ofdetection rate, accuracy, and false alarm rate

Method	Normal	DoS	Probe	R2L	U2R	Accuracy	FAR
Winner's (2000)	99.50	97.10	83.30	8.40	13.20	93.30	0.55
Multiclass SVM (2003)	**99.6**	96.8	75	4.2	5.3	92.46	**0.43**
MOGFIDS (2007)	98.36	97.20	88.6	11.01	15.79	93.20	1.6
BIRCH and SVM (2011)	99.3	99.5	**97.5**	28.8	**19.7**	95.7	0.7
NFC (2014)	98.2	99.5	84.1	31.5	14.1	N/A	1.9
Proposed method	98.55	**99.57**	80.94	**31.63**	16.23	**95.71**	1.45

5 Conclusion

In this paper, we proposed a model of modified K-means with multi-level SVMs to construct a high-performance intrusion detection system. Modified K-means was applied to reduce the number of training data sets and to obtain new, high-quality training data sets with which to learn SVMs. The nu-SVM and RBF kernel functions of LibSVM were employed to implement multi-level SVMs. The converted and normalized training and testing data sets were preprocessed to render them suitable for the SVM classifier. This model classified the attacks in DoS, R2L, and U2R effectively. In addition, its capability to classify other types of instances, such as Normal and Probe, is not worse than those of other models. In future studies, we attempt to improve performance in relation to the Normal and Probe categories and conduct comparisons with other studies that employ different evaluation methods.

References

1. Ghanem, T.F., Elkilani, W.S., Abdul-kader, H.M.: A hybrid approach for efficient anomaly detection using metaheuristic methods. J. Adv. Res. Article in Press (2014)
2. Om, H., Kundu, A.: A hybrid system for reducing the false alarm rate of anomaly intrusion detection system. In: 1st International Conference on Recent Advances in Information Technology (RAIT), pp. 131–136. IEEE (2012)
3. Horng, S.-J., Su, M.-Y., Chen, Y.-H., et al.: A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Syst. Appl. 38, 306–313 (2011)
4. Hasan, M., Nasser, M., Pal, B., Ahmad, S.: Intrusion Detection Using Combination of Various Kernels Based Support Vector Machine. International Journal of Scientific & Engineering Research 4, 1454–1463 (2013)
5. Yao, J., Zhao, S., Fan, L.: An enhanced support vector machine model for intrusion detection. In: Wang, G.-Y., Peters, J.F., Skowron, A., Yao, Y. (eds.) RSKT 2006. LNCS (LNAI), vol. 4062, pp. 538–543. Springer, Heidelberg (2006)
6. Celebi, M.E., Kingravi, H.A., Vela, P.A.: A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst. Appl. 40, 200–210 (2013)
7. Gao, M., Wang, N.: A Network Intrusion Detection Method Based on Improved K-means Algorithm. Adv. Sci. Technol. Lett. 53, 429–433 (2014)
8. Sujatha, M.S., Sona, M.A.S.: New fast k-means clustering algorithm using modified centroid selection method. International Journal of Engineering Research and Technology 2, 1–9 (2013)

9. Kathiresan, V., Sumathi, P.: An efficient clustering algorithm based on Z-Score ranking method. In: International Conference on Computer Communication and Informatics (ICCCI), pp. 1–4. IEEE (2012)
10. Nazeer, K.A., Sebastian, M.: Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. In: Proceedings of the World Congress on Engineering, vol. 1, pp. 1–3 (2009)
11. Pfahringer, B.: Winning the KDD99 classification cup: bagged boosting. ACM SIGKDD Explorations Newsletter 1, 65–66 (2000)
12. Xiang, C., Chong, M., Zhu, H.: Design of mnitiple-level tree classifiers for intrusion detection system. In: 2004 IEEE Conference on Cybernetics and Intelligent Systems, vol. 2, pp. 873–878. IEEE (2004)
13. Xiang, C., Yong, P.C., Meng, L.S.: Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees. Pattern Recognit. Lett. 29, 918–924 (2008)
14. Natesan, P., Balasubramanie, P., Gowrison, G.: Improving the Attack Detection Rate in Network Intrusion Detection using Adaboost Algorithm. Journal of Computer Science 8, 1041–1048 (2012)
15. Ambwani, T.: Multi class support vector machine implementation to intrusion detection. In: Proceedings of the International Joint Conference on Neural Networks, vol. 3, pp. 2300–2305. IEEE (2003)
16. He, L.: An Improved Intrusion Detection based on Neural Network and Fuzzy Algorithm. Journal of Networks 9, 1274–1280 (2014)
17. Lu, H., Xu, J.: Three-level Hybrid Intrusion detection system. In: International Conference on Information Engineering and Computer Science, ICIECS 2009, pp. 1–4. IEEE (2009)
18. Gogoi, P., Bhattacharyya, D., Borah, B., Kalita, J.K.: MLH-IDS: A Multi-Level Hybrid Intrusion Detection Method. The Computer Journal 57, 602–623 (2014)
19. Sabhnani, M., Serpen, G.: Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context. In: MLMTA, pp. 209–215 (2003)
20. Jianliang, M., Haikun, S., Ling, B.: The application on intrusion detection based on k-means cluster algorithm. In: International Forum on Information Technology and Applications, IFITA 2009, vol. 1, pp. 150–152. IEEE (2009)
21. Bhatia, M., Khurana, D.: Experimental study of Data clustering using k-Means and modified algorithms. International Journal of Data Mining & Knowledge Management Process (IJDKP) 3, 17–30 (2013)
22. KDD Cup 1999 Data set. `http://archive.ics.uci.edu/ml/machine-learning-databases/kddcup99-mld/`
23. LibSVM. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
24. Tsang, C.-H., Kwong, S., Wang, H.: Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. Pattern Recognit. 40, 2373–2391 (2007)