Journal of Applied Sciences

ISSN 1812-5654

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (12): 2178-2181, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.2178.2181

Hadoop-based Multi-classification Fusion for Intrusion Detection

Xun-Yi Ren and Yu-Zhu Qi
Department of Information Security, College of Computer,
Nanjing University of Posts and Telecommunication,
210003, Nanjing, China

Abstract: Intrusion detection system is the most important security technology in computer network, currently clustering and classification of data mining technology are often used to build detection model. However, different classification and clustering device has its own advantages and disadvantages and the testing result of detection model is not ideal. Cloud Computing, which can integrate multiple inexpensive computing nodes into a distributed system with a strong computing power, can quickly process massive data. Hadoop is the most widely used cloud computing platform. Currently, cloud-based intrusion detection system has become the new direction, but the study of effective IDS based hadoop is still scarce. This paper presents an intrusion detection system model with feature multi-classification fusion based on hadoop, which combined with K-means clustering and 1V1-SVM multi-classification method, using Map to form a new <Key, value>according to the classification center, to form a new classification, then to re-form a new detection model by removing the repeated value. The testing results of large amounts of data for the MIT Laboratory KDDCUP99 Experimental show that the fused classifier has more accuracy than mere classifier.

Key words: Cloud computing, hadoop, K-means clustering, 1V1-SVMmulti-classification, fusion technology

INTRODUCTION

Intrusion detection Yang *et al*. (2004) is an information technology for detecting an unauthorized user who attempts to undermine the security of computer systems, it is mainly through the monitoring of computer network system logs, audit information, status, behavior and system usage, to detect the user's unauthorized use and misuse behavior, these abnormal behaviors undermine the integrity and confidentiality of computer systems and the availability of resources. Intrusion detection can help system administrators investigate illegal intruders to prevent further attacks by hackers. Intrusion Detection Systems (IDS) is considered to be the second security gate after firewall.

Currently, IDS has many methods, especially the main clustering and classification using heuristic mining algorithms, to build detection model. But in the current high-speed network traffic environment, for massive and complex network data, the speed and accuracy of training samples by different clustering and classification algorithm cannot meet the real-time requirements and because of their advantages and disadvantages, inevitably result in the locality of feature extraction, making the IDS false positives and false negatives increase.

Cloud computing (Zhang and Cao, 2009) is a computing model which provides a dynamic and scalable virtualized. Resources via the Internet, it has a strong computing power and storage capacity and it is the essence of virtualization (Zhao and Hu, 2010). Hadoop Xie *et al*. (2010) is the mainstreaming framework for the cloud computing platform ,which is developed by Apache top-level project, it forms clusters (Assuncao *et al*., 2009) by a lot of basic hardware facilities, deals with the massive amounts of data in distributed parallel programming calculation at a faster speed. The most central part of hadoop are the distributed programming Mapreduce and Hadoop Distributed File System (HDFS). MapReduce framework which is running on<key,value> key-value pairs and its input is a set of key-value pairs, usesMap function to process data, then combinesand shufflesthe processed data, at last, distributes the intermediate results to the Reduce nodes, also it outputs a set of key-value pairs as <key,value>. Since, the program which is under Map Reduceruns in parallel, the complex data can be forwarded to cloud computing environments for processing to improve computing speed.

With the increasing complexity of invasive behavior, IDS need to constantly acquire new intrusion samples to obtain more accurate characterization of intrusion. As the diversity, complexity and bias of classification, feature

Corresponding Author: Xun-yi Ren, Department of Information Security, College of Computer,
Nanjing University of Posts and Telecommunication, 210003, Nanjing, China Tel: 136-1158-6255

fusion and classifier fusion (Han *et al*., 2006) technology is particularly important. Fusion technology leverages high-dimensional information to help detect more accurateinformation. Since, fusion is anoptimal process for multi-classification techniques, we can effectively utilize mapreduce computing model to fuss. This study proposes the hadoop-based feature multi-classification fusion technology, which combineswith K-means clustering (Luo *et al*., 2003; Xiang *et al*., 2003) and 1V1-SVM Support Vector Machine (SVM) (Chen *et al*., 2005) classification method in parallel and fusses (Li *et al*., 2001) the classified result in hadoop environment, it effectively improves the efficiency and accuracy of classification, shortening the fusiontime and improves the existing intrusion detection system performances with a greater help.

INTRUSION DETECTION MODEL BASED ON MULTI-CLASSIFICATION FUSION

We propose a cloud-based feature multi-classification fusion model for intrusion detection system ,which is divided into two phases. The first phase, use K-means clustering and 1V1-SVM multi-classification methods to classify in parallel, the second phase use the classified results to fuss in hadoop by distance-based fusion method. This method helps obtain a more comprehensive and effective classification, the model shown in Fig. 1.

Generally, the performance of detection system which constituted by a single classifier is limited, single classifier is typical for the specified application at a high detection rate, but it will expose its bias and limitations in face of massive data. The fusion of multiple classifiers is not simply reflected in the features fusion, in many cases in a complementary way, is to improve the decision-making capacity of classifier, the generalization error of fused classifier system will be much smaller than the weightedaverage generalization error ofthe single classifiers.

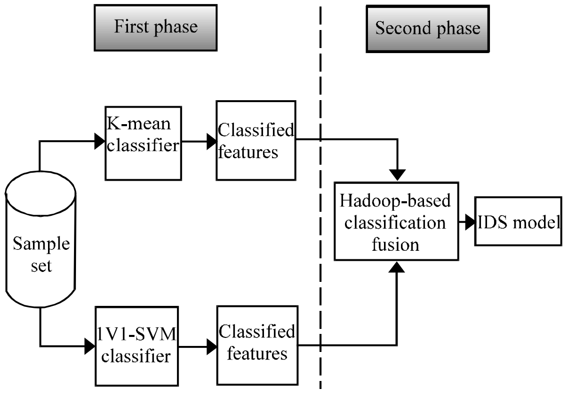

Fig. 1: Multi-classification fusion model for intrusion detection system

HADOOP-BASEDMULTI-CLASSIFICATION FUSION ALGORITHM

- For the proposed model, supposed that sample set S = {X_i}, i = 1...n object X_i= (X_{i1} X_{i2} X_{i3}..., X)$_{id}$ represents a d-dimensional vector space, X_i is called a data sample of the d-dimensional space, attribute set A = A_1, A_2, A_3,...A_d, $A_1 \times A_2 \times A_3 \ldots \times A_d$ represents a d-dimensional space A_1, A_2, ..., A_d, is the spatial dimension. By the K-means clustering, set cluster number as k_1, get the classification results C_1 = {C_1, C_2, C_3, ...C_{k1}}

  K-means clustering classification algorithm is described as follows: Clustering is a process of clustering data into different parts, it gathers close datas into a cluster by determining the similarity of datasto classify data. Based on K-means clustering classification algorithm, the sample set is classified into k categories. Firstly, give the number of clusters k (k>1), set I = 1, randomly selectk samples as initial cluster center Z_j (I) = {ω_1, ω_2, ω_3, ...ω_k}, j = 1, 2, 3,..., k; Secondly calculate the Euclidean distance D (X_i, Z_j(I)), i = 1, 2, 3...n of each data sample X_i to initial cluster center, European Eq.:

$$d(x_1,x_2)=\sqrt{\sum_{j=0}^{d}(x_{1j}-x_{2j})^2}$$

  (d for the vector dimension). if D (X_i, Z_j(I)) = min {D (X_i, Z_j(I)), i = 1, 2, 3, ..., n, then, $X_i \in \omega_j$, set I′ = I+1, calculate the new cluster center:

$$Z_j(I')=\frac{1}{n}\sum_{i=1}^{n_j}X_i^j,n_j$$

  represents the sample numbers of the jth cluster, that is Z_j (I′) = {ω_1′, ω_2′, ω_3′,...,ω′}; calculate squared error and criterion function:

$$R_c:R_c(I)=\sum_{j=1}^{k}\sum_{k=1}^{nj}\|X_k^j-Z_j(I')\|^2,\text{if}\,\|R_c(I')-R_c(I)\|<(\xi,(\xi=0.1)$$

  Then, terminate algorithm. Otherwise, set I = I+1 and re-coomputed D (X_i, Z_j(I)) for cycle the second step. Finally, output the cluster set C_1 = (C_1, C_2, C_3, ..., C_{k2})
- For 1V1-SVM classification, suppose the sample set S = {X_i}, set classification number as k_2 and the classification results C_2 = (C_1, C_2, C_3, ..., C_{k2})

SVM algorithm focus on the junctions of two types sample points, using a linear division principle and search optimal linear hyperplane,divide the sample points into two categories, it has a higher detection rate and hasthe maximum and optimal inter valcl as sification. However, objective things we need to face are far more than two categories, multi-classification SVM classification quickly become an important technology. One to one support vector machine (1V1-SVM) for multi-classification method, treat training samples classification as two categoriesof data classification, namely typical SVM, then iterative loop until the number reaches thek

- 1V1-SVM classification algorithm is described as follows: Firstly, using the standard SVM divide the sample set S $\{X_i\}$into two categories C_1', C_2'. Calculatethe new sample set $S' = \{(x_1y_1), (x_2y_2),..., (x_n, y_n))\}$, $x_i \in R^d$, $y_i \in \{-1, 1\}$, i = 1...n. Function ϕ (x_i) is the transformation from the input space R^d to the Hilbert space H. Kernel function is the inner product of two vectors in Hspace, that is K (u,v)= ϕ(u), ϕ(v), u, $v \in R^d$, set K (u,v) = $[u, v+1]^d$. For linear samples, classified plane is expressed as follows (w.x)+b = 0 , the iinverse of the class interval is $½\|w\|^2$, to get the results is to solve quadratic programming problems:

$$\min_{(w,b)} \frac{1}{2}\|w\|^2$$

constraints y_i $(((w.x_i)+b)+1)\geq 1$, i = 1, ..., 1. For nonlinear sample, we need the introduction of linear soft interval slack variable ξ_I based on nonlinear hardinterval, mappings $T = \{(x_1', y_1), ..., (x_1', y_1)\}$, $x_i' = \phi(x_i)$, classified plane (w'.x')+b = 0 to get the result is to solve quadratic programming:

$$\text{Problem}\min_{(w,b)} \frac{1}{2}\|w'\|^2 + C\sum_{i=1}^{1}\xi_i,\ \text{Constraints}\, y_i\left(\left(\left(w'.x_i^{'}\right)+b'\right)+1\right)$$
$$\geq 1-\xi_i, i=1,...,1$$

Then, we get the first SVM classification results C_1', C_2' after quadratic programming. Secondly, respectively C_1', C_2' continue the first step, loop until the classification number is k_2; Finally, output the cluster set $C2 = \{C_1, C_2, C_3, ..., C_{k2}\}$

- Set $k=k_2+k_2$, $C' = C1 \cup C2 = \{\omega_1, \omega_2, \omega_3,...,\omega_k\}$, with i = 0, 1, 2, ..., k for the key value, the corresponding value ω_1 for the Value, namely $<I, \omega_i>$ key-value pairs(number is k),as mapreduce frame input.
- Map phase: According to the principle that the shorter the distance between sample centers, the greater the similarity is, calculate:

$$O_i = \frac{1}{n}\sum_{j=1}^{ni} X_i^j,\ O_{i,j} = d(O_i, O_j)$$

for each ω_i sample center and calculate any twoof distance of k-centers as the Key value, $(\omega_m \cup \omega_n)$ as the Value value, namely $<O_{i,j}, (\omega_m \cup \omega_n)>$, at last, sort the key-value pairs

- Reduce phase: Given the value of threshold ξ, at the sorted key-value pairs, if $O_{i,j} < \xi$, then remove repeated value and re-merge for Value, output the new $<O_i', \omega_i'>$, ω_i' is fused classification. Otherwise retain the original classification. Finally, the merger of classification in both cases will be final classification results.

EXPERIMENT

To validate the effectiveness of proposed hadoop-based multi-classification fusion technology, we use authoritative sample data set by MIT Laboratory named KDDCUP99. This data set is collected in a simulated invasive network environment. It mainly contains intrusion behavior including denial of service attacks (Dos), remote attacks (R2L), unlawfullyelevate local users privileges attacks (U2R), illegal monitoring and probing (Probing). Every record of KDD99 has 41 attributes, in this experiment, we randomly select 1% of data sets as training data, which is about 50,000 records and select about 500,000 records as testing data.The testing data is made up with 64.6% normal data, 18.8% known intrusions and 16.6% unknown behavior.

Set $k_1 = k_2 = 5$, d = 41, fusion threshold $\xi = 0.5$, based on the model proposed in the study. Table 1 lists thenoemal and intrusive behavior tested by three classifiers. Table 2 lists the corresponding detection rate.

From the tables above, hadoop-based multi-classification fusion detection technology has more accuracy and a higher recognition rate compared to the K-means clustering and 1V1-SVM multi-classification. Hadoop-basedmulti-classificationfusiontechnology makes up the lower detection rate of K-means clustering and 1V1-SVM for Probing, R2L and U2R, correspondingly

Table 1: Normal and intrusive behavior

Classification	Normal	Dos	Probing	R2L	U2R
Real samples	319540	58803	21564	10035	528
K-means	301757	56202	19934	8934	445
1V1-SVM	304195	5713	19796	9272	429
Hadoop fuson	312109	58638	21116	9870	486

Table 2: Detection rate of normal and intrustive behavior

Classification	Normal (%)	Dos (%)	Probing (%)	R2L (%)	U2R (%)
K-means	95.33	95.57	92.44	89.03	84.28
1V1-SVM	96.1	97.17	91.8	92.3	81.25
Hadoop fuson	98.6	99.7	97.9	98.35	92.05

reduces false positives and false negatives. Experimental results show that the proposed hadoop-based classification fusion detection technology combines the advantages of the two classification methods, makes up the deficiencies of a single classification and has a better overall performance. In addition, on the classifier fusion detection systems, the greater the complementarity of each subsystem, the integrated effect is the better.

CONCLUSION

With the further development of network technology, network intrusion technology becomes increasingly complex, it bears increasing pressure immediately to follow the changes of network applications and network intrusions, intrusion detection technology urgently need to study more effective classification methods and more strong information processing technology. Efficient classification method help improve the performance of intrusion detection systems, feature fusion and multiple classifiers fusion can reduce the performance requirements of single classifier to optimize and also relatively easy to form high recognition systems. In this paper, K-means clustering and 1V1-SVM classifiers fusion reduces the uncertainty of target and reduces the ambiguity of information, enhances target or events identification, improves system detection accuracy.The next steps will further study the real-time and detection efficiency.

ACKNOWLEDGMENTS

This study is supported by the National Grand Fundamental Research 973 Program of China under (2011CB302903), National Natural Science Foundation of China (61073188, 61073115), Key University Science Research Project of Jiangsu Province (11KJA520002) and the Priority Academic Program Development of Jiang Su Higher Education Institutions (yx002001).

REFERENCES

Assuncao, M., A. Costanzo and R. Buyya, 2009. Evaluating the cost benefit of using cloud computing to extend the capacity of clusters. In: Proceedings of the 18th ACM International Symposium on High Performance Distributed Computing, June 11-13, 2009, Munich, Germany, pp: 141-150.

Chen, W.H., S.H. Hsu and H.P. Shen, 2005. Application of SVM and ANN for intrusion detection. Comput. Oper. Res., 32: 2617-2634.

Han, C.Z., H.Y. Zhu and Z.S. Duan, 2006. Multi-Source Information Fusion. Vol. 5, Tsinghua Univarsity Press, Beijing, China, pp: 76-82.

Li, X.L., J.M. Liu and Z.Z. Shi, 2001. The Chinese web page classifier based on support vector machine and unsupervised clustering. J. Comput., 24: 62-68.

Luo, M., L.N. Wang and H.G. Zhang, 2003. An unsupervised clustering-based intrusion detection method. Acta Electronica Sinica, 31: 1713-1716.

Xiang, J., N. Gao and J.W. Jin, 2003. Application of clustering algorithm in network intrusion detection. Comput. E, 16: 48-49.

Xie, J., S. Yin, X. Ruan, Z. Ding and Y. Tian, 2010. Improving mapreduce performance through data placement in heterogeneous hadoop clusters. In: Proceedings of the IEEE International Symposium on Parallel and Distributed Processing, Workshops and Phd Forum, April 19-23, 2010, Atlanta, GA., USA., pp: 1-9.

Yang, W., B.X. Fang and X.C. Yun, 2004. A high-performance distributed intrusion detection system research and implementation. J. Beijing Univ. Posts Telecom., 4: 83-86.

Zhang, J. and J.G. Cao, 2009. Research on internet cloud computing technology. Telecom. Network Tech., 10: 10-15.

Zhao, Y.J. and R. Hu, 2010. Virtualization-based green cloud computing. J. Hunan Univ. Sci. Tech. Nat. Sci., 25: 86-88.