The application of k-means algorithm based on 2-SVM in Intrusion System

Shizhen ZHAO, Qiang YU, Yu FU, Wei SONG

College of computer and software engineering, Xihua University, Chengdu, 610039, China

zhangyujianyj@sina.com

Keywords: Intrusion detection; local optimum; Detection rate

Abstract. After analyzing the superiority and the deficiency of K-means clustering algorithm, the author put forwards the IDS based on the 2-SVM, according to the research achievement on SVM in the field of IDS. The result shows that this algorithm can avoid the problem of decreasing detection rate by local optimum while using the K-means clustering algorithm alone.

Introduction

With the rapid development and wide use of WSN (Wireless Sensor Network), people get benefit from it in every aspects, such as traffic control, mechanical monitoring, military affairs, smart home and so on. At the same time, the safety protection problem of computer and network system has been becoming more and more important. Since it's much more common for the attacks on the Internet network, the tackles of attacks, with the technical essence of detection and control and the defending form, has become complex actively and dynamically. The IDT (Intrusion Detection Technology) is of high value in the application of WSN[1]. Due to its features, node energy and limited network computing power, the WSN has difficulty in designing IDS. How to devise an IDS with low energy consumption for one node, low false alarm rate, high detection accuracy and wide application has become a significant research subject[2-3].

IDS collects the tracks after being attacked from Internet network and system and other key points, and analysis whether attacks from outside or inside exists. There are mainly two ways of IDT, misuse detection and anomaly detection. The former one basically relies on studying the marked data sample from training, and once it encounters unknown detection, it needs use new to retrain the detection system, so the limitation of real-time cost when application is huge. On the contrary, the latter one can automatically identify unknown attacks without the need of relying on samples. Cluster analysis is widely used in IDS for its ability of non-relying on data setting, relatively accurate records and identifying anomaly action. The K-means cluster algorithm[4-5], which was proposed by Mac Queen, can highly analyses data automatically and develops potential mode. With the fast speed and strong ability for manipulating massive data, it's truly a cluster algorithm without supervision. By improving the algorithm of K-means, the author will detect the attack real-time. Nowadays there are various ways of detection on research of IDS home and abroad. Li Yang[6] firstly applied K-means algorithm with the IDS from the WSN, and it showed that it was convenient to do the application, and the demand for training DS was low, and it can continually develop potential mode. However, the performance on the hyper plane was bad, and it was sensitive for isolated point. Zhao Peng[7], et al designed a K-means cluster algorithm based on the weighted complex network, in which they introduced weighted degree of complex network, weighted aggregation degree and weighted aggregation index into cluster algorithm. In this way, the problem of initialization sensitivity could be avoided. But in the light of massive data set, the consumption of overall network energy by computation complexity cannot get better performance.

Nowadays the application of K-means cluster algorithm into IDS is of great research value, but with the effect of the noise and isolated point, this thesis improves the application of cluster algorithm in the literature No.6, and proposes that the use of 2-SVM, in which it solves the problem of local optimum when using K-means algorithm alone, with the principle of structural risk minimization. The result showed that, compared with using the K-means algorithm alone, it has

lower false alarm rate and higher detection rate, and has increased the quality of detection.

Correlation Theory

A brief introduction to K-means

K-means, which is a clustering algorithm, is one of the heuristic clustering partition methods. And it is the popular and promising calculating method these years because it can not only be used in technology that detects abnormal conditions without specific guide and also it can keep optimizing and updating model base of the monitoring system. The general idea of K-means is to randomly create k numbered clusters based on K as a parameter, and then to classify the N targets or the targets left in the tuple to their nearest K clusters, and it has to satisfy (1) K≤N and there should be at least one target or tuple in each cluster; (2) each target or tuple can be only belong to one single cluster. Then we can calculate the mean value of each cluster until the final expectation of the set criterion function shows up. After repeated and hardworking and enumeration of clusters, we finally find the the most optimized overall situation as well as the final clustering results. The criterion function used for judgments are generally used for the sum of squared errors, and it is defined as:

$$E = \sum_{i=1}^{k} \sum_{p \subset C_i} \left| p - m_i \right|^2 \tag{1}$$

E is the sum of squared errors of all targets or tuples, p is the given data object, m_i is the weighted average of the target C_i in the cluster. Both p and m_i are multidimensional data, and the numbers of C_i is decided by the classification. The distance between each target and the center of the clusters are generally used with Euclidean distance. And it can she showed as:

$$i=\left(x_{i_1}, x_{i_2}, \wedge, x_{i_p}\right) \text{ and } j=\left(x_{j_1}, x_{j_2}, \wedge, x_{j_p}\right) \tag{2}$$

After exhaustive iteration, N data objects will be classified into the nearest K clusters of minimal mean square error (mse) and converges to expectations.Accurate as far as possible do the same as close as possible to the subject matter of the same kind of objects as far as possible away from, to detect the intrusion behavior clearly, K - means algorithm is applied to intrusion detection system to achieve the goal.

A brief introduction to SVM

SVM (support vector machine) is a kind of machine learning method which is based on the theories of statistical learning, and the support of SVM is exactly a good way of detecting intrusion behavior. Its most prominent characteristic is based on structural risk minimization principle to improve the learning machine's generalization ability, namely by the limited sample training set small error will still be able to guarantee the independent test set to keep the minor errors. Because SVM is a convex optimization problem, therefore, the partial optimization is precisely the overall optimization, which can sufficiently solve the possible problem of the application of K-means clusters calculating to the detection of intrusion. For a better discussion, this paper will utilize the standard SVM algorithm, which is introduced as follows: Known categories of training data set {(xi, yj) | I = 1,...L}, which is suitable for two kinds, of which one type is xi, tag, as sample category, structure, the optimal separating hyperplane in the feature space the sample is divided into two classes, make it meet the sample on one side of the hyperplane is positive, the other side is negative, and away from the plane heterogeneous samples classification interval is the largest, express to the classification hyperplane

$$(w \cdot x_j)+b=0 \tag{3}$$

The following judgments should be used to distinguish the types of the unknown sample xi

$$(w \cdot x_j)+b \geqq 1 \rightarrow y$$

$$(w \cdot x_j)+b \leqq -1 \rightarrow y \tag{4}$$

To find the optimal hyperplane is the process of solving quadratic optimization problem

$$\min\left(\frac{1}{2}\left\|\mathrm{W}\right\|^2 + \mathrm{c}\bullet\sum_{i=1}^{l}\xi_i\right)$$
$$\text{st} \qquad y_i\left[\left(w\bullet x_i\right)+b\right]\geq 1-\xi_i \tag{5}$$
$$\xi_i \geq 0, \mathrm{i=1,\ 2\ldots l}$$

ξ represents the sample about the hyperplane deviation (slack variable), c means the penalty coefficient.Type (5) the second is penalty terms, a larger c means for error item specifies the larger penalty term, thereby reducing the misclassification of data points; Smaller c means that it ignores some tiny-little mistake classification points, thus it can get the classification of the large interval. By introducing the multiplier of Lagrange of $i=1,2,\wedge,l$, we can obtain its dual form:

$$\max\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K\left(x_i, x_j\right)$$

$$s.t.\quad 0\leq\alpha_i\leq c,\ \ i=1,2,\wedge,l \tag{6}$$
$$\sum_{i=1}^{l}\alpha_i y_i = 0$$

Hence, we can get the following resulting algorithm decision function:

$$f(x) = sign\left[\left(\sum_{i=1}^{l}\alpha_i y_i K\left(\chi, x_i\right)\right)+b\right] \tag{7}$$

The optimal hyperplane is decided by the two samples which are in the closest distance to the surface of classification.

The K-means cluster algorithm based on 2-SVM

To solve the problem of local optimum, the author put forward the cluster algorithm based on 2-SVM, in which it assumed that N objects of data set be divided into K clusters, and each cluster be divided into two groups by 2-SVM, and then 2K clusters into the new K_1 clusters. Such is the procedure, and after several times comes the result. The main procedures of cluster algorithm are as follows:

Step1: To avoid the problem caused by a large number annihilating a small number, the objects should be treated standardized to reduce the dependence of the units and mark the data sets. ..

Calculate the mean absolute error: $s_f = \frac{1}{p}\left(\left|y_{f_1}-m_f\right|+\left|y_{f_2}-m_f\right|+\wedge+\left|y_{f_p}-m_f\right|\right)$

In which $y_{f_1}\wedge y_{f_p}$ are p measurement values for the variable f ; m_f is the mean value for f , and $m_f = \frac{1}{p}\left(y_{f_1}+y_{f_2}+\wedge+y_{f_p}\right)$;

Calculate the measurement value $x_{f_i} = \frac{y_{f_i}-m_f}{s_f}, i=1,2,\wedge,p$

Step2: Choose K objects as average value(initial clustering center), and calculate the average value, then value N objects to the most similar clusters.

Step3: Map at least 1 to N-K objects in each cluster to two different high dimensional spaces, according to decision function, and at last separate each object with each other in K clusters, making sure that the range between them is as big as possible.

Step4: According to the mean value, rearrange K1 cluster by the data base distance, and value K1 to K.

Step5: After the procedure as above, recalculate the average value in clusters, and value every object to the most similar cluster.

Step6: Repeat Step3 and Step4 as above, and get the new cluster.

Setp7: After continuous partitions, the data objects will be no longer changing, and the optimal clustering result will be obtained.

The result and the analysis of the experiment

The resources of the data

The Intrusion detection data comes from the network environment of the US air force local network established by MIT Lincoln lab, After 9 weeks of network link and system auditing data,

In the following 1999, Prof. Sal Stolfo from Columbia University and Prof. Wenke Lee from South Carolina University analyzed and pre-processed the data above, and established the famous KDD Cup data set, which is the generally accepted and practical network auditing data set. This data set consists of the whole data sets, corrected.gz and 10% data set. This thesis used training data set from 10% data set (kddcup.data_10_percent) as analysis objects. There are 494021 pieces of data records, among which 97278 pieces are normal data records, 396473 pieces are abnormal data records, containing 4 major aggressive behaviors: Dos(Denial of Service), U2R(User to Root), R2L(Remote to User) and Probe, as shown in the chart 2 as follows. The 22 kinds of attack behaviors are consisted of 41 characteristic fields, and altogether 42 adding the last track. Using unmarked data sets to test intrusion detection algorithm, and using marked data sets to test algorithm performance.

Tab.1. Distribution of the Attack Data

Types of the Attack	Number	Percentage (%)
Normal	97278	19.691066
Probe	4107	0.831341
DOS	391458	79.239142
U2R	52	0.010526
R2L	1126	0.227926

Data Selection

When the cluster algorithm is used for anomaly detection, it is demanded that the data of anomaly behavior account for 3% to 10&. When the record of anomaly behavior in the cluster is bigger than that in the training set, this set is called the anomaly behavior set, and others are normal behavior sets. The anomaly detection establishes type mode by training process, and compares detection data with type mode. Generally detection rate and false positive rate are used to evaluate detection result.

The 41 characteristic fields are comprised of TCP connecting basic features, TCP connecting content features, statistical characteristics of network based on time and statistical characteristics of network based on the host. According to the demand of the research, 7 of them——duration, protocol type, times of logon failure, times of access to controlling files, number of connecting different hosts, number of SYN connection failure, number of connecting same source port——are chosen as the input value feature vector. According to the distribution of the attack data above, 500 sets of data are chosen, among which 20 are test data, the rest 480 are training data. Since only 52 pieces of data are left in U2R with the least ratio, so 480 sets are divided into four major sets, in which the top 3 only consists of Probe, Dos and R2L, and the last one consists of 4 kinds of data as above. The experimental environment is CPU3.00GHz, the memory as 2.0G, the operating system as Windows XP, and the development platform as VC++6.0.

Comparison of algorithm performance

Make a comparison between the detection rate in each kind of attack by K-means cluster algorithm before improvement, and that of attack after adding 2 kinds of vector algorithm, and the results are shown as follows:

Tab.2. Contradistinction of Detection rate

Detection rate	Probe	DOS	R2L	Mixing of four kinds of data
K-means algorithm	95.17	91.67	53.8	83.6
2-VSM Hybrid cluster algorithm	98.69	94.2	61.74	89.44

Tab.3. Contradistinction of False alarm rate

False alarm rate	Probe	DOS	R2L	Mixing of four kinds of data
K-means algorithm	7.24	6.83	17.43	15.33
2-VSM Hybrid cluster algorithm	3.62	2.39	9.76	11.26

It's obvious that, from the 2 tables above, under the same degree of attack and environment, the K-means algorithm that introduces 2-VSM algorithm have the higher detection rate and lower false alarm rate than the initial algorithm. When encountering with different types of attacks, it's normally seen that the false alarm rate has been decreased after introducing the 2-SVM algorithm. The reason is that the introduction of 2-SVM algorithm has improved the defect that the K-means algorithm may fall into the problem of local optimum.

Conclusion

In the field of IDT for WSN, cluster algorithm has many advantages on the application of anomaly detection. It can realize the automatic detection for IDS without the need to mark the information ahead. During the process of detection, it can continuously improve the existing data models, explore potential modes, in the meantime it can increase detection rate and decrease false alarm rate. Easy and unsupervised as it is, K-means algorithm is susceptible to noise or isolated point, thus lead to local optimum or other results that cannot get better cluster result in the process of algorithm application. In the light of the deficiencies for the algorithm its own, this thesis applies 2-SVM algorithm to K-means algorithm to sort and merge internally, which optimized the initial algorithm. The result shows that the algorithm which introduces support vector machine can get higher detection rate, lower false alarm rate and better performance when applied to IDS than the initial algorithm. Further research will be concentrated to the way to save the system's average energy consumption.

Acknowledgement

In this paper, the research was sponsored by the Chuihui Plan Project of Ministry of Education (Project No. 13226651), Application Foundation Research Project of Sichuan Provincial Education Department(Project No. 11226016), Research on Instruction Detection Key Technology for IOT(Project No. szjj2013-018) and Sichuan Provincial Science and Technology Program Project (Technology Innovation Project Special Fund, Project No. 2014ZZ0026).

References

[1] Murad A, RassamM A, MaarofAnazida Zainal. A Survey of Intrusion Detection Schemes in Wireless Sensor Networks [J]. American Journal of Applied Sciences, 2013, 9(10) 1-9.

[2] Shilpa, S, Patil, P, S, Khanagoudar. Intrusion Detection Based Security Solution for Cluster Based WSN [J]. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2013, 1(4) 123-132.

[3] Pooja, GuptaDr, Naveen, Hemrajani. Security Issues in Wireless Sensor Network: A Review [J]. International Journal of Engineering Sciences & Research Technology, 2013, 2(5) 342-350.

[4] Ordonez C,Omicecinski E. Efficient disk -based K-means clustering for relational databases[J]. IEEE Trans. Knowledge and Data Engineering, 2004, 16(8) 909-921.

[5] YAO Xin. Studies on Complex networks and its Clustering Degree [D]. Beijing: Tsinghua University, 2005.

[6] LI Yang. Application of K-means Clustering Algorithm in Intrusion Detection [J]. Computer Engineering, 2007, 33(14) 154-156.

[7] ZHAO Peng, GENG Huan-tong, CAI Qing-sheng, et al. A Novel K - means Clustering Algorithm Based on WeightedComplex Networks Feature [J]. Computer Technology and Development, 2007, 17(9) 35-37, 40.

[8] Burgers J.C., A tutorial on support vector machines for patten recognition. Data Mining and Knowledge Discovery, 1998, 2(2) 121-167.

[9] RAO Xian. Application of support vector machine in intrusion detection [J]. Computer Engineering and Design, 2007, 28(10):2297-2299

[10] Corts C., Vapnik V., Support vector networks. Machine Learning, 1995, 20(3) 273-297.

[11] LIU Fengzhu, GONG Xun. A Clustering Method for Anomaly Intrusion Detection [J]. Computer Security, 2013, (8) 13-16.

[12] ZHANG Xin-you, ZENG Hua-shen, JIA Lei. Research of intrusion detection system dataset-KDD CUP99 [J]. Computer Engineering and Design, 2010, 31(22) 4809-4812,4816.