An Attack Threat Effect Analysis Method Based on K-Means Evaluation

Jindong He[1], Bo Liu[2] and Qian Guo[3]

[1]State Grid Fujian Electric Power Research Institute, 350000 Fuzhou, China
[2]State Grid Henan Electric Power Company, 450000 Zhengzhou, China
[3]Global Energy Interconnection Research Institute CO., LTD. 210000 Nanjing, China

Abstract. To take full advantage of the specified features of the attack dataset in network attack effect evaluation, maximize the efficiency of evaluation without losing its accuracy. This paper proposed a K-Means evaluation technique using dimensional entropy components, derived from changes in network entropy through attack period and the advantages of clustering algorithm in data mining. This method makes a pre-process of the attack dataset on the basis of network entropy, mapping it to a two-dimensional plane and utilize the output of pre-process as the input of clustering. Then establish a relation between the attack dataset and the effect category via an improved K-Means algorithm, thus achieving an explicit division of attack effect set and provide efficient evaluation result. The experimental results prove that the method can process attack dataset with high efficiency, as well as provide a visualized evaluation result by the form of cluster tree.

Keywords: Marketing Safe; Effect evaluation; K-Means algorithm.

1 Introduction

With the rapid development of information technology, Internet network structure is becoming more complex and evolved in the direction of diversification and integration. Followed by network attacks are increasing, according to incomplete statistics, by 2017 the number of attacks on the network to host up to 946 million times, therefore, it's necessary to carry out research on the different types of aggressive behaviour, and promote attack effect evaluation to improve host resistance to attack, and to improve the security of information systems.

At present the attack effect evaluation is mainly based on the analytic hierarchy process (AHP) the establishment of evaluation index system and the use of fuzzy comprehensive evaluation of the rough set theory. In China, Qianmu li and other combined with analytic hierarchy process and fuzzy comprehensive evaluation method to the quantitative evaluation index [1-2], Li Q based on set pair theory put forward a kind of effective processing assessment classification standard boundary ambiguity sets of evaluation methods, based on rough set attribute importance according to determine the index weight and the denial of service attack resistance to the network set to evaluate [3]. Compared with the traditional evaluation of directly using the original data calculation method, to measure the network performance before and after the attack on entropy difference change to unified standard will be collected different kinds of index normalized, simplified assessment in the process of data pre-processing step, and intuitively reflect dynamically against the effects on the system.

This paper combines the characteristics of network entropy and the clustering algorithm, this paper proposes an attack effect evaluation based on k-means clustering algorithm, this algorithm using target network entropy difference before and after the attack will collect data mapped to two dimensional vector space, and use the heuristic information to improve the k-means clustering algorithm of the index data, by quantitative calculation process to get the final evaluation results, at the same time of meet the requirements of batch operation efficiency greatly overcome the subjectivity of the evaluation process.

2 Evaluation Based on entropy difference attack effect

The father of information Theory C.E. Shannon in reference to the concept of entropy in thermodynamics, to remove the redundant information is put forward after the average amount of information as the entropy as a measure of the uncertainty of information, using discrete random event probability calculation formula of the information entropy is given:

$$H = -K\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

The PI said the probability of random events, K is a positive constant. From the information entropy calculation formula, we can see, the smaller the probability of a random event occur, the entropy value, the greater the show that the greater the uncertainty. Inspired by the information entropy, in the study of computer network attack effect evaluation method can introduce the concept of [3], the corresponding network entropy to target performance variation of parameters before and after the network attack, the characterization of network performance index return one becomes Vi, and then to define its entropy

$$H = -\log_2 V_i \tag{2}$$

Therefore, the target network entropy changes before and after the attack can be used to calculate the entropy difference Δ H, with H-before and H-after respectively before and after the attack of network entropy, the entropy difference could be calculated through the type:

$$\Delta H = H_{after} - H_{before} = -\log_2 V_{after} - (-\log_2 V_{before}) = -\log_2(\frac{V_{after}}{V_{before}}) \tag{3}$$

Obviously, the target network after the attack, the greater the change of parameter values, calculated by the type of entropy difference is, the greater the attack, the more obvious effect. Therefore using the method of quantitative index of the entropy difference can provide normalized evaluation data and visually describe the change of network performance before and after the attack.

3 K-means clustering evaluation algorithm

3.1 The traditional K - means algorithm

Clustering analysis is an important research in the process of data mining and pattern recognition methods, in recent years, due to its in interdisciplinary aspects of strong adaptability makes clustering is widely used in machine learning and data analysis and statistical science and so on. Academics on the cluster does not have a clear definition, but clustering can be described as the test points in space according to its inherent characteristics are divided into different classes of cluster process, one kind of entities within the cluster is similar, not same cluster entity is not the same. Each class cluster mentioned here are the test points converge in space; Class is essentially a cluster contains relatively high density of the point set of multidimensional space connected area, they contain with relatively low density of point set separate area with the rest of the class cluster [5]. As in most clustering algorithms, K - means is a kind of typical partition clustering algorithm [6], this kind of clustering method is usually need to specify the number of clustering beforehand and cluster center, then through several iterations gradually reduce the error of the objective function value, the final clustering results are obtained when the objective function of convergence. The traditional K - means clustering process can be summarized as:

1) Select the clustering center;

2) Other points to the clustering center distance are calculated respectively, and will be divided into the clustering of the closest point;

3) According to the clustering results recalculation clustering centers, and according to the above steps to clustering;

Repeat step 1) to 3) until the results of convergence.

Though K - means has the advantages of fast convergence [7], but improper selection of the initial clustering center is easy to make the algorithm without the condition of the solution, in addition, according to the actual clustering objects correctly judging the validity of clustering is also influence the effect of K - means clustering is an important factor.

3.2 Improve the K means clustering evaluation algorithm

In the light of the problems mentioned above to the traditional K - means algorithm is improved, and the improved method is applied to the network attack effect evaluation, due to the attack data sets itself according to the target environment, the characteristics of different diversity and identity of analysis than attack data set and cluster test space point set can be found, in the phase of the opposite sex and identity between them exist in common: first, to the target environment or attack means change, under the condition of rather smaller attack data have similar properties, for example, in the continuous time same host of normal access to the Internet wiretapping, the data broadly similar, this is against the identity of the data; On the contrary, the target environment or attack means differences, will lead to attack the differences with the clustering analysis of data in the data between the opposite sex is strong, class conforms to the characteristics of the large data similarity, therefore, can make use of K - means the entropy difference algorithm, the attack on quantitative effect do the qualitative evaluation. Algorithm firstly the attack data preprocessing, calculate the specific performance index of the entropy difference before and after the attack and use the cosine theorem map it to the two dimensional vector space, unified scale while reflecting the characteristics of the data itself, and then after the preprocessing based on K - means algorithm of data clustering, classification characteristics of network attack effect as heuristic information to determine the clustering number, by calculating a two-dimensional vector space of each data point in Euclidean distance to determine the initial clustering center, after several rounds of iteration, to satisfy the minimum distance cost function clustering results in the evaluation of effect as the final output.

(a) Data preprocessing

Suppose you have a set of index data V = {V1, V2,... , vitamin k,... }, and attack of Vk index entropy respectively before and after the Hk and Hk 'to the index of different scale, although the size of the entropy difference can reflect the change of before and after the attack, but because of the different indicators measure, the value of the entropy difference itself the differences can be large, for example, for throughput and delay, the former numerical generally correlates to the above, while the latter is commonly a few milliseconds, only from the data on the size of the fail to reflect the effect of attack, so for each index in calculating the introduction of maximum entropy Hmax may attack effect to acquire normalization, according to the actual entropy difference and the ratio of ideal entropy difference to determine the degree of attack. Variable θk defined attack level, the attack data description for each group, which respectively before the attack, attack resistance entropy and entropy performance indicators ideal maximum entropy, with entropy difference formula, θk variables, for the attack rate and

$$\cos\theta_k = |\frac{H_k^{'} - H_k}{H_{\max} - H_k}|, \theta_k \in [0, \frac{\pi}{2}]$$.

In its domain is monotone decreasing function, the smaller the $\cos\theta_k$ said attack effect is more obvious, in the case of θk equal, actual amount attack effect is determined by the entropy difference numerical size. Through the above process introduced variable associated with entropy difference θk each attack data points can be mapped to a single data value of two dimensional vector space, on, with the size of the entropy difference said the length of the vector, and determine the direction of the vector x axis Angle θk.

(b) K average classification effect.

After data preprocessing to index, using the improved K - means algorithm clustering of data points in the evaluation of attack effect, specific steps are as follows:

Step 1: determine the clustering center. First of all, according to different evaluation object classification effect of prior knowledge to select clustering number, for example, the computer network attack effect evaluation, the final attack effect is commonly the Result = {good, better, general, poor}, according to this information to determine the clustering number of initial value k0 = 4.According to the data point k entropy difference and Angle (ΔHk,θk) calculating the coordinates of points in the Cartesian coordinate system (xk,yk)=(ΔHk*cosθk, ΔHk*sinθk), assuming that the initial clustering center in (xmin, ymin) ~ (xmax, ymax) equidistant distribution, to calculate the coordinates of, Can find out the coordinates for

$$C_i = (\frac{(2i-1)(x_{\max} - x_{\min})}{2k_0} + x_{\min}, \frac{(2i-1)(y_{\max} - y_{\min})}{2k_0} + y_{\min})$$ 。

Step 2: divide the clustering. Respectively calculated for each data point to all the Euclidean distance clustering center Ci, stored in the matrix D0, and divided the node to have minimum Euclidean distance clustering center belongs to. Calculate the clustering results of compact and separation effect of sexual function S (U, k) [8] as distance cost function,

$$S(U,k) = \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}^2 \| x_i - c_j \|^2}{\underset{i,j}{avg} \| c_i - c_j \|^2}$$

Where n is the total number of data points, K is the number of clusters, Xi denotes the ith data points, CJ represents the j cluster centers, data points Xi and cluster centers CJ 2-norm squared, in the two dimensional vector space said Xi and CJ distance vector, Rij is Xi of CJ's membership and

$$r_{ij} = \{ \begin{matrix} 1, x_i \in Clusterj \\ 0, else \end{matrix}$$

Can see, compact and separation effect function measure the ratio of the distance between the within class distance and class, intra class distance is small, inter class distance is large, effect function value smaller, indicating a clustering effect better, whereas clustering effect.

Step 3: recalculation clustering center. Because in the process of clustering need to ensure that the distance between the class as small as possible, assuming a new clustering center for μi, the data points mean square error of the dk to new clustering center, in the process of clustering need to find the optimal solution of thread μi to minimize SD.

The parameters of the rki obviously its division to the nearest cluster center clustering which can guarantee the minimal opportunity of SD, namely

$$r_{ij} = \{ \begin{matrix} 1, D_{ki} = \min D_k \\ 0, else \end{matrix}$$

For μi, μi derivation of the SD,

$$SD = \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ki} \| d_k - \mu_i \|^2 = \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ki} (d_k - \mu_i)(d_k - \mu_i)^T = \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ki} (d_k^T d_k - d_k^T \mu_i - \mu_i^T d_k - \mu_i \mu_i^T)$$

$$\frac{\partial SD}{\partial \mu_i} = 0 - \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ki} (\frac{\partial (d_k^T \mu_i)}{\partial \mu_i} + \frac{\partial (\mu_i^T d_k)}{\partial \mu_i} - \frac{\partial (\mu_i \mu_i^T)}{\partial \mu_i}) = \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ki} (\frac{\partial (\mu_i^T d_k)^T}{\partial \mu_i} + \frac{\partial (\mu_i^T d_k)^T}{\partial \mu_i} - \frac{\partial (\mu_i^T \mu_i)^T}{\partial \mu_i})$$

$$= -\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ki} (\frac{\partial (\mu_i^T d_k)}{\partial \mu_i^T} + \frac{\partial (\mu_i^T d_k)}{\partial \mu_i^T} - \frac{\partial (\mu_i^T \mu_i)}{\partial \mu_i^T} = \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ki} 2(d_k - \mu_i)$$

Order to $\frac{\partial J}{\partial \mu_i} = 0$ obtain $\mu_i = \frac{\sum_K r_{ki} d_k}{\sum_K r_{ki}}$ that μi cluster i should take the average coordinates of all the data points to calculate the after μi μi updating cluster centers Ci and repeat Step2 a new cluster center partition clustered to obtain new effects function S (U, k)', if S (U, k)'> S(U, k), repeat Step3, otherwise the algorithm stops and outputs the result of clustering Cluster0;

Step 4: After a complete process of clustering, clustering results obtained Cluster0, and then optimize the number of clusters k, since the number of clusters k determined empirically initialization, there may be slight deviations in the number of clusters to calculate the new number of clusters to new cluster number of repeating the process of clustering, clustering to get the final result set {cluster}, compare the clustering effect and has taken the smallest compact cluster results and effects of separation as a function of the final output.

4 Simulation results and analysis

Experiment using Matlab write data preprocessing and K-means algorithm and partial correlation effect grading simulation for the effect of consumption of resources to attack the case of the assessment, the data points in Figure 1 that meets the joint Gaussian distribution as the host after the attack CPU utilization,

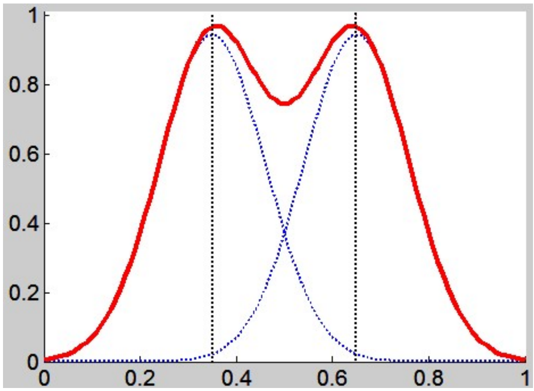

Figure 1. Simulation result.

Suppose the target host 20 per test launch 20 attacks, resulting in 400 data points (0,1) between the Gaussian distribution as shown in Figure 1, the input data points from the distribution function we can see that after the attack the host CPU utilization rate spread (0,1) between, and joint Gaussian distribution chart shows, most of the data is distributed (0, 0.4) between, so you can predict the final results of the assessment of the effects of the attack as "poor" or "fair" It will account for the majority. To test this prediction using the proposed method of data preprocessing and clustering, the initial number of clusters is set to 4, the process of distribution algorithm for the number of clusters k ~ {2,3,4,5,6} Iteration , obtained firmness and separation function value Sk = {33.9081,23.2042,10.9868,10.6196,16.4735} (k = 2, ..., 6), when k = 5 , Sk minimum, so take the number of clusters k = 5 is the optimal output result, as shown:

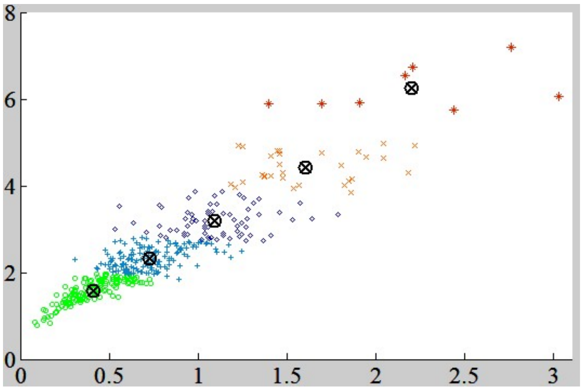

Figure 2. Optimal output result.

The results can be seen by the clustering algorithm input 400 data points according to how well the attack effect is divided into five categories, can be combined with the actual situation and the target network expert advice is given with reference to a qualitative description of these five attacks {excellent, good , fair, poor, had no effect}, where "excellent" effect produced by a handful of attacks, as shown in Figure 2 away from the coordinate origin "*" data points, the closer to the origin of coordinates of the data points the worse attack by final clustering result map can visually see the effect of the attack stepped classification.

To further verify the effectiveness of the attack algorithm for data processing using KDDCUP99 intrusion detection data set sampled data simulation. KDDCUP dataset contains five million records, and provides a subset of 10% of the training and testing subsets, where each data item includes 41 characterization, covering the basic features include a TCP connection to the network traffic statistics features, including related information, count feature of this experiment to extract KDDCUP99 dataset, count feature records the number of connections in the current connection with the detection time has the same target host, the range [0,511], the data centralized teardrop tab 979 data items sample extraction, to achieve the 600 attack data points clustering evaluation, simulation results as shown below:

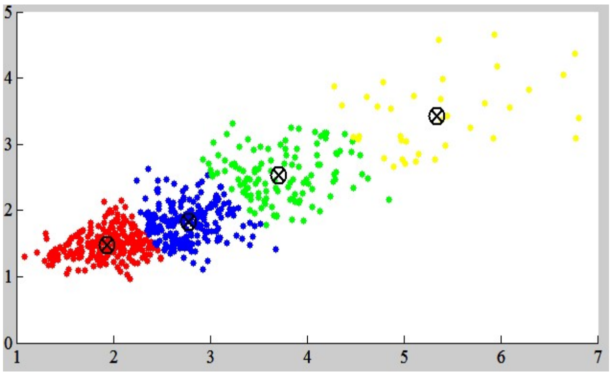

Figure 3. Simulation results of 600 attack data points clustering evaluation.

According to the clustering results, combined with KDD CUP dataset values, attack effect can be obtained as follows hierarchical description:

Table 1. KDD CUP attack dataset effect rated.

Count characteristic value range	Class in the data points	Effect rated
0~38	233	Attack poor, increase access number is not obvious, almost no effect on the target host
39~116	210	Attack effect, a smaller number of access increase, a slight impact on the target host normal communication
117~254	120	Attack better, a big increase access connections, significantly reducing communication performance target host normal
255~511	37	Attack effect, access to the rapid growth of the number of connections, the destination host cannot communicate properly

5 Conclusion

This paper draws on the idea of network entropy, maximum entropy is calculated by the ideal entropy difference associated with the introduction of the angle component, by the law of cosines data points are mapped to two-dimensional vector space, and then use the default Kmeans clustering algorithm cluster centers were attack effect grade calculation, simulation experiments show that the algorithm to maintain the traditional Kmeans algorithm for fast convergence characteristics, and by treating the cluster centers algorithm overcomes Kmeans easy to fall into local optimal solution, can effectively deal with attack data visualization and presentation the effect of the classification results. However, there are a number of algorithms to be the perfect place, for example, in the calculation of a variety of different attacks caused by the effect of the target network, in addition to the degree of attack, attack need to consider the dangers of empowerment to define a higher level of attacks on the network Effect property brought, in addition, quantification of the value of qualitative indicators, which are worthy of further study.

References

1. XiaoqianLiu, QianmuLi, TaoLi. Private classification with limited labeled data. Knowledge-Based Systems. 2017, 133: 197-207.
2. Zhang B, Li Q, Zhang Y, et al. The Proactive Defense of Energy Internet Terminals Edge-Access Using the Network Topology Autoassociation. IEEE Journal on Emerging & Selected Topics in Circuits & Systems, 2017, 99: 1-15.
3. Tang L, He SB, Li QM. Double-sided Bidding Mechanism for Resource Sharing in Mobile Cloud. IEEE Transactions on vehicular technology. 2017, 66 (2): 1798-1809.
4. Zhang B, Li Q, Ma Y. Research on dynamic heuristic scanning technique and the application of the malicious code detection model. Information Processing Letters, 2017, 117: 19-24.
5. Li, QM; Hou, J; Qi, Y. A classification matching and conflict resolution method on meteorological disaster monitoring information. DISASTER ADVANCES. 2013, 6 (2): 415-421.
6. Li QM. Multiple QoS Constraints Finding Paths Algorithm in TMN. INFORMATION. 2011, 14 (3): 731-737.
7. Li QM, Zhang H. Information Security Risk Assessment Technology of Cyberspace: a Review. INFORMATION. 2012, 15 (11): 677-683.
8. Li QM, Li J. Rough Outlier Detection Based Security Risk Analysis Methodology. CHINA COMMUNICATIONS. 2012, 9 (7): 14-21.