Application research of improved K-means algorithm in intrusion detection

Liu Xiaoguo [1,a] , Tian Jing [2,b],

[1] Jilin Agricultural University Jilin, Changchun,130118 ,China

[2] Changchun Vocational Institute of Technology ,Jilin, Changchun,130033 ,China

[a]Liuxiaoguo@jlau.edu.cn, [b]35871773@qq.com

Keywords: Intrusion detection, Data mining, Clustering analysis, K-means clustering, Minimum spanning tree

Abstract. An improved K-means clustering algorithm is put forward on basis of the split-merge method for the purpose of remedying defects both in determination of value in K and in selection of initial cluster centre of traditional K-means clustering. At first , the concept of independence degree of date was incorporated into the experimental date subset construction theory , using independence degree to evaluate the importance of nature.Next ,the database is merged into several classes in respect of density of date points ,the combination of the minimum spanning tree algorithm and traditional K-means clustering algorithm is conducive to the achievement of splitting .Eventually ,the KDD Cup99 database is applied to conduct simulation experiment on the application of the improved algorithm in intrusion detection .The results indicate that the improved algorithm prevails over traditional K-means algorithm in detection rate and false alarm rate.

Introduction

With the popularity of computer application in the field of all over the world, the Internet is a unique way of changing people's study, work and daily life. However, with the improving of the network utilization, threats to network security is becoming more and more diversified, network security problem has become one of the important issues in the world today. At the same time, the static security technology can't meet the demand of modern network security problem day by day serious, therefore an act that can detect Intrusion and active security defense technology, Intrusion Detection System, IDS, arises at the historic moment. This study based on data mining technology applied in the intrusion detection system based on the related theory of, in the study of data mining algorithms-K-means algorithm of clustering algorithm, combined with the graph theory put forward by the Zahn clustering algorithm (called a minimum support tree clustering algorithm) to the traditional K-means algorithm to enhance and improve the algorithm.

Intrusion detection and clustering analysis

Intrusion detection, whether using analysis to identify outliers from the group, or by building a classifier classification of the intrusion events are trained to detect intrusion by unknown data, is a study of anomaly detection based on unsupervised clustering algorithm. Unsupervised clustering is a kind of learning method based on statistics theory, its biggest characteristic is based on V apn ik structural risk minimization principle, and try to improve generalization ability and learning algorithms, the unsupervised clustering algorithm applied to intrusion detection, can guarantee in the case of a priori knowledge is insufficient still has higher detection rate, so as to make the intrusion detection system has good detection performance. Clustering analysis and association rules and sequence pattern analysis combine to make the analysis of the data mining in three important ways. Clustering algorithm characteristics analysis was carried out on the training data set, the similar data into the same class. K-means cluster analysis as a classical algorithm of clustering algorithm in intrusion detection application in two stages: the first stage, the training module build classifiers; The second phase, detection engine based on discriminant classifier the behavior of each type.

© 2015. The authors - Published by Atlantis Press

Improved k-means algorithm

1、The traditional K-means algorithm

K-means clustering algorithm process: the first step is to randomly select K or initial clustering center; The second step, the rest of each point is assigned to the most similar, the closest cluster; The third step, the iteration step 2, the criterion function iteration convergence, output the clustering results. Commonly used square error criterion function formula is as follows:

$$J_C(m)= \sum_{j=1}^{k} \sum_{x_{i\in c_j}} |x_i - z_j|^2 \qquad (1)$$

All objects including JC (m) said data set the sum of squared error, JC (m) the smaller the instructions in the class the higher similarity, Xi said in multidimensional space point (of the given data object), Z_J shows the average value of the cluster $C_{j.}$

k-means inevitably have: 1) only when the cluster mean useful and pre-set number of clusters k value of the case in order to use K-means algorithm; 2) is not suitable for discovery of non-spherical clusters or clusters vary greatly in size; 3) for outliers and data points are sensitive to noise; 4) Regular termination shortcomings and deficiencies in the four major areas of local optimal solution.

2、Improved K-means clustering algorithm

Traditional K-means algorithm as the basis for improved algorithms, combined with the concept of minimum spanning tree, in the absence of pre-determined number of cluster centers and cluster K value, the training set for the merger, division and other operations to determine the final poly class of result sets, this paper will improve K-means algorithm named MSTK-means algorithm. MSTK-means algorithm is mainly related to: how to choose the initial cluster centers combined to produce three original cluster and divide and produce new and the like.

Set D $\{I_1, I_2, ..., I_n\}$ indicates the training data set, In the data set represents a data object. C $\{o_1, O_2, ..., O_m\}$ represents the cluster center set, O_m represents the center of a data object, namely the class all the data mean. T $\{E_1, E_2, ..., E_k\}$ represents the set of edges Minimum Spanning Tree, Ek = $<I_x, I_y>$, $I_x, I_y \in D$, wherein the length of the edge EK | EK | calculated by the formula 1 draw.

(1) The initial cluster centers Selection.

MSTK-means algorithm to obtain the initial cluster centers according to point density. Specific methods of operation are: given radius R and R ', traverse D, any object $I_x \in D$, $\{D- \{I_x\}\}$ Ix each point distance d is calculated by the formula distance formula, the statistics $d \leqslant R$ number of objects. All objects in descending order to obtain dot density data set D ', will join the collection point density maximum point C, D will not traverse any point C Ix radius R' point density within the range of the maximum point is added C, to obtain an initial cluster centers set C $\{O_1, O_2, ..., O_m\}$, C really included in D, and the number of elements is much smaller than D.

(2) Consolidated generates an initial class.

According center set C $\{O_1, O_2, ..., O_m\}$, as the center point O_x, the radius R 'all points within the range into a class, change the class mark these points, and calculate the mean O_x' as the center class values. Merge get the initial cluster centers set C '$\{O_1', O_2 ', \cdots O_x'\}$.

(3) dividing to produce new classes.

Traversing C 'get a subset of data contained in each class. D_x '$\{I_1 ', I_2 ', ..., I_k '\}$ expressed in O_x' as the center of all the data records to build point D_x 'minimum spanning tree Tx $\{E_1, E_2, ..., E_h\}$. Remove side grew to equal the average tree Lavg side edges to give forest Fx $\{T_1 ', T_2', ..., T_x '\}$, where jh, Computing Center O_y Meike subtree "= center (T_x') ($T_x \in F_x$), statistics and change the label for each data object class, get a new set of cluster center C "$\{O_1", O_2 ", O_3", ..., O_y "\}$.

$$Lavg = {}^{1}\!/_{h} \sum_{k=0}^{k=h} |E_K| \qquad （2）$$

Wherein, h represents the number of sides of the tree, | EK | EK represents the length of the edge. MSTK-means clustering algorithm called repeatedly thought these three aspects, until the criterion function (Equation 1) converges. MSTK-means algorithm is described as follows:

Step1 calculated according to the formula dot density data set D in the dot density of each data element, in descending order according to the dot density to obtain new data set D, point density is calculated as follows:

$$D(I_X)=countd(I_X,I_Y)<R),I_Y\in(D-\{I_X\}) \qquad (3)$$

Step2, starting from the first record, the center will not be set and not set in the center of each data element to a point within a given radius of the added initial cluster centers set O, the data set D of all elements in the continental divide from the shortest distance method to the corresponding cluster, in order to obtain original clusters. The shortest distance is to be divided into elements class center of Euclidean distance minimum, namely:

$$D(I_X)_{min}=\min\{d(I_x,O_y)|I_x\in d,O_y,\in C\} \qquad (4)$$

Wherein, Ix represents the data elements to be divided, O_y represents a collection of C in a cluster centers. Ok established with Continental shortest distance Ix, Ix will be merged into the Ok representative class and the mean recalculated Ok representatives of the class as the new value .

Step3 get each cluster set D_x (center O_x collection contains representatives of all the elements), separatist actions for each D_x MSTK-means algorithm according to the third aspect of thinking, calculating a new set of class C. Center Update Center Access to new processing data sets D (set of all clusters class centers that set C), empty set C.

Step4 criterion function (Equation 1) convergence, implementation of Step5; criterion function does not converge, perform Step1. Step5 output with the subject of the final data set of class D (last clades class center set after completion).

In MSTK-means algorithm performs Step1-Step3 merged with the splitting process will inevitably encounter can not be clearly classified into a cluster cluster boundary data points. These data points may have at least two clusters of cluster characteristics, and therefore can not be classified clearly. MSTK-means clustering algorithm based on the data points and the cluster class cosine similarity, mark such data points. Cosine similarity calculation formula as follows:

$$COS(X,Y)=\frac{\sum_{k=1}^{n}x_k y_k}{\sqrt{\sum_{k=1}^{n}x_k{}^2}\sqrt{\sum_{k=1}^{n}y_k{}^2}} \qquad (5)$$

Wherein, n represents the data dimension, X is the boundary data points, Y represents any one of a cluster of record. MSTK-means clustering algorithm boundary point X partition to an average cosine similarity largest cluster.

Experimental results and analysis

1、Experimental Data

Mining Intrusion Detection commonly used experimental data set -KDDCup99, comprising: 1) Denial of Service (DOS); 2) the distal unauthorized access (U2R); 3) local privileged user access (R2L); 4) detection or detection (PROBE); 5) normal (NORMAL) five types of data.

2、Experimental data preprocessing

Ⅰ) data dimensionality reduction. Since KDD Cup99 is a high-dimensional data sets, data must be processed before the data dimensionality reduction. Used in the experiment to eliminate redundancy from the relevant variables, weak property related or unrelated to retain the way to build a subset of the key attributes of high-dimensional data sets to reduce the dimension. According to the definition of formula Euclid D istance calculated for each class of property with respect to the properties of the degree of independence, the selected subset of key attributes construct experiments. Thus possible to reduce the amount of processing data in step Ⅱ.

Ⅱ) data preprocessing. To make the data more conducive for data mining process, modeled on Document .The data set is converted into a discrete attribute continuity property. Then follows are standardized and normalized.

Numerical standardization: AVG_j、S_j respectively attribute mean and mean absolute deviation value X_{ij} ,X_{ij} 'normalized , is calculated as follows:

$$AVG_j=\frac{1}{n}(x_1+x_2+x_3+\cdots+x_n) \qquad （6）$$

$$S_j=\frac{1}{n}(|x_j-AVG_J|+|x_j-AVG_J|+\cdots\cdots+|x_j-AVG_J|) \qquad （7）$$

$$X_{IJ}'=\frac{x_{ij}'-AVG_{min}'}{S_1} \qquad （8）$$

Value normalization: the value standard, then normalized to [0,1] value X_{IJ}'' normalized calculated as follows:

$$X_{IJ}'' = \frac{x_{ij}' - X_{min}'}{x_{max} - x_{min}} \qquad (9)$$

III) to select a subset of experiments. To be able to meet two preconditions cluster analysis assumptions built four experimental detection table than normal data for each table number and the number of pieces of data are invaded 2950: 50, the number of such intrusion data in the total data the proportion of the number is less than 2%. As shown in Table 1.

Testing set	Records	Normal Records	Invasion Records	Attack Types	Category Number attack
Test1	3000	2950	50	17	4
Test2	3000	2950	50	17	4
Test3	3000	2950	50	11	3
Test4	3000	2950	50	13	4

Table 1 test data sets of Hybrid Intrusion Detection

Experimental results and analysis

Criteria experimental use detection rate and false detection rate as the merits of the detection system, the formula is as follows:

Detection rate = detected intrusions data number / total number of data intrusion,

false detection rate = was detected for the invasion of normal data number / total number of normal data

First, according to cluster-standard training set, and then, using the clustering result set table in turn four detection testing, test results of traditional K-means algorithm and improved dot density clustering algorithm shown in Table 2. MSTK-means algorithm for many experimental results, when Radius = 0.01、 ROI = 2 clustering results were better, with a higher detection rate and low false alarm rate, the test results are shown in Table 3.

Checklist	Traditional K-means algorithm		Dot density based on improved algorithm	
	Detection rate	Error detection rate	Detection rate	Error detection rate
Test1	90.0%	35.02%	94.00%	11.15%
Test2	88.0%	35.83%	90.00%	11.25%
Test3	82.0%	35.63%	88.00%	11.15%
Test4	82.0%	37.62%	94.00%	12.00%

Table 2. compares the algorithm test results

As can be seen, the traditional K-means algorithm is very easy to have a "miscarriage of justice" phenomenon, the normal behavior is recognized as intrusion, a direct result of the false detection rate is higher, on average up to 36.03%, which is more difficult to accept. The dot density improved K-means algorithm based on the initial cluster centers still have a random selection, leading to instability in the test results.

Table Name	Detection rate (detection rate increase rate)	False detection rate (false alarm rate reduction rate)
Test1	98.00%（8.89%,4.26%）	0.78%（97.78%，93.00%）
Test2	90.00%（2.27%，0.00%）	0.54%（98.49%，95.20%）
Test3	94.00%（14.63%，6.81%）	0.75%（97.90%，93.27%）
Test4	94.00%（14.63%，0.00%）	0.71%（98.11%，94.08%）

Table 3 MSTK-means test results

From Table2,MSTK-means algorithm, compared with the traditional algorithm 8,2,12,12 percentage points increase in the detection rate. Reduce the false detection rate is more significant, with an average reduction rate of 97.07%. Meanwhile, based on the point density improvement compared to K-means algorithm, the effect can be seen in the false detection rate is the most obvious, the average reduction rate of 94.14%. Visible, according to a dot density as the initial cluster centers selected, both to ensure the selected point greater density, but also to avoid the phenomenon of the center of focus. So choose MSTK-means clustering algorithm, the initial success of the center to avoid the situation of K-means algorithm converges in a local minimum, reducing the dependence on initial clustering result, completely overcome the initial value of the random selection may lead to malpractice local optimal solution, but do not be given in advance the number of clusters k value, the detection algorithm is better than the first two.

Conclusion

This paper presents a traditional K-means algorithm to improve the algorithm MSTK-means algorithm. The algorithm is introduced the concept of dot density and the minimum spanning tree clustering algorithm, in the case of only the initial training data set, under the premise of not given the number of clusters to achieve clustering. Meanwhile, the cluster center selection certainty, completely avoid the volatility clustering phenomenon caused by random initial value. Experiments show that the algorithm is applied to the detection result of the data mining intrusion detection system module was more stable than the traditional algorithms, ensuring both a high detection rate have a low false alarm rate, detection effect is remarkable.

At the same time improved algorithm than the traditional method has a high time complexity, thus clustering operation consumes more time, work and study in the future, need further optimization to ensure that the effect of reducing the detection time complexity, and in the actual environment Using the algorithm, test its efficiency in a real environment.

Acknowledgement

This research was financially supported by Jilin Provincial, china Department of education "Twelve five" science and technology project the Foundation.

References

[1] KAMBER M. Data Mining : Concepts and Techniques [M]. Han jianwei,Translation Beijing:Machinery Industry Press,2012:338-465.

[2] xie zhuo. Study of network intusion detection based on clustering algorithm[J].Modern Electronics Technique,2012,29(35):91-93.

[3] DANG Xiao-chao,HAO Zhan-jun,WANG Xiao-juan.Intrusion detection based on cluster connectivity clustering algorithm.Computer Engineering and Applications,2010,46(21):82-85.

[4] Fu tao, Sun Yanmin Pos-based k-means Alogorithm and its App lication in Network Intrusion Detection System [J]. Computer Science, 2011,5(38):54-55.

[5] Zhou Haiyan, Bai Xiaolin.K-means Initial Clustering Center Optimal Algorithm Based on Graph Theory [J] Computer Measurement & Control,2010,18(9):2167-2169.

[6] PORTNOY L, ESKIN E, STOLTO S J. Intrusion detection with unlabeled data using clustering [C]//Proceedings of the ACM CSS Workshop on Data Mining Applied to Security.Philadelphia, PA, USA: ACM, 2001:56-60

[7] BRADLEY E, FAYYAD U. Refining initial points for K-means clustering[A]//Proceeding of fifteenth international conference on machine learning ICML98[C], San Francis-co:Morgan Kaufmann ,1988:91-99.

[8] BAO Ying. Research and application of cluster algorithm based on partitioning method [D].Da lian:Dalian University of Tech technology, 2008.

[9] Guo Ming,DingHuafu. Clustering method based on hybrid of SOM and K-means [J].Computer & Digital Engineering,2008,36(9):22-24.

[10] BRADLEY P S, FAYYAD U M. Refining initial points for K-means clustering [C]//Proceeding of the 15th International Conference on Machine Learning. San Francisco, CA , USA:Morgan Kaufmann,1998 :91-99.

[11] ZHOU Aiwu, CUI Dandan, PAN Yong. An optimization initial clustering center of K-means clustering algorithm [J]. Micro-computer & Its Applications ,2011, 30(13):1-3.