Proceedings of "The 2nd International Conference on Applied Information and Communications Technology" - ICAICT 2014

ISBN: xxxxxxxxxxxx

ELSEVIER

A Patent Technique of Jaccard Discrete (J-DIS) Similarity Clustering Algorithm

Sharjeel Imtiaz, Prof Dr Saadiah Yahya

Phd Student at Asia e University,Malaysia, C70105130002@aeu.edu.my
Professor at UITM ,saadiah@tmsk.uitm.edu.my

Abstracts

Traditionally, the classification object yields homogeneous object to separate cluster. Few authors investigated clustering based on k-Means to distinguish intrusions based on the particular class. Mostly, k-Means algorithm finds out similarity between the object based on distance vector for smallest dataset. We proposed a new approach Jaccard Discrete (J-DIS) based approach which is combines with k-Means to find most similar measures over features attribute values in a larger dataset. Further, this paper is describing best suitable larger dataset taken from KDD CUP-99 dataset [1].Moreover, the J-DIS k-Means approach can be applied over clinical informatics and wireless clustering based routing protocols.

Keywords: Ecludian Distance, Jacord coefficient, Intrusion Detection, KDD CUP-99, k-Means;

1. Introduction

One of the pioneer algorithm k-Means was proposed by [5], where each cluster is represented by the center of the cluster. Further, some of the approaches have applied k-Means traditional approach for finding association rules among intrusion detection. But, the trend of discovering clusters from dataset is higher than association patterns from training data. The traditional intrusion detection approaches are taken from k-Means [4] [7] distance similarity vector. The traditional approach was based on square error [5] but there are few more approaches, which are based on different distance measures. The real problem of this algorithm is that it is sensitive to the selection of the initial partition and may conclude with a local minimum (not a global minimum) depending on the initial partition. To overcome the problem of local minimum the author studied four measure Ecludean, Cityblock, Cosine and correlation. The most popular metric for continuous features is the Euclidean distance which is a special case of the Minkowski metric (p =2). It works well when a data set has compact or isolated clusters. Similarly cityblock measures each partition based on median at centre [9][6].

Dataset chosen in this research is similar to clustering approach k-Means of BIRCH [8]. But, all approaches have mostly relied on feature condenses data rather data point means raw tcpdump data. But, the main criterion is used for clusters and their merging based on euclidean distance. To overcome the problem of local optima and cluster results produced depend on the initial values for the means, and it frequently happens that suboptimal partitions are found. The standard solution is to try a number of different starting points. A popular solution is to normalize each

variable by normalization attribute criteria. Though, this is desirable in current research. Our research is based on new J-DIS based similarity measure which is a variant of jaccard coefficient combine with k-Means to produce condense clusters. The aim of research is to overcome over fitting, the way of mean initialization, produce better similarity measurement criteria and to overcome the problem of defining k-Means always for small dataset rather for large dataset.

2. Problem Statement

- The way to initialize the means was not specified.
- k-means is suitable for small dataset which is not desirable for all datasets.
- The results depend on the metric used to measure $\| x - m_i \|$. A popular solution is to normalize each attribute between [0, 1] by its normalization criteria, though this is desirable.
- Over fitting is a phenomenon well-known in the area of statistical Machine Learning. It ensues when one trains a powerful learning model with many degrees of freedom on a relatively small data sample. In effect, the model utilizes the available parameters to "memorize" the available data sample, failing to derive proper generalizations from the training data. The observed effect of this is a poor performance on previously unseen data.

3. Proposed Solution

The metric is used to minimize and the choice of a distance measure will determine the shape of the optimum clusters. We proposed a J-DIS k-Means algorithm based on new approach Jaccard coefficient based similarity measure clustering technique [10]. The current equation d^{JAD} = J11/J10+J01+J11 is the binary based jaccard coefficient based on similarity approach. For two data records with *n* binary variables *y* the variable index *k* ranges from 0 to *n*-1. These combinations will replace current Jaccard coefficients J11, J10 and J01 by new J-DIS based coefficient formulation.

- Bj>Ai: Total number of variables Bj are having greater value than Ai
- Ai>Bj: Total numbers of variables Ai are having greater value than Bj.
- Ai ∩ Bj:The total number of variables Ai and Bj are having same values.
- c is the constant for boosting common measures to achieve smaller value of similarity

The similarity equation with new Jaccard Discrete (J-DIS) is as follows.

$$J - Dis = Ai \cap Bj \ / \ Ai > Bj \ + \ Bi > Ai \ + \ c^{*} Ai \cap Bj$$

The k-Means algorithm clustering similarity measure is based on euclidian distance. Based on euclidian distance the objective function of k-Means clustering is $J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i - cj||^2$. This research modifies cost or objective function with new proposed J-DIS based k-Means similarity equation. This objective function takes three equation forms. Equation 1 is $k * n$ over the same attribute A_i and B_j values. By combining (1), (2) and (3) eventually derived final form of equation (4), which is cost formulation of all three cases over the K_n .

1) This is the cost or objective function where Ai set values is equal to Bj.

$$J - DIS = \sum_{j=1}^{k} \sum_{i=1}^{n} c * Ai \cap B \quad (1)$$

2) This is the cost or objective function where Ai set values is greater than Bj.

$$J - DIS = \sum_{j=1}^{k} \sum_{i=1}^{n} Ai > Bj \quad (2)$$

3) This is the cost or objective function where Ai set values is less than Bj.

$$J - DIS = \sum_{j=1}^{k} \sum_{i=1}^{n} Bj > Ai \quad (3)$$

Combining (1), (2) and (3) Finally, we get J-DIS based k-Means objective function shown at (4).

$$J - DIS = \sum_{j=1}^{k} \sum_{i=1}^{n} = Ai \cap Bj / Bj > Ai + Ai > Bj + c * Ai \cap Bj \quad (4)$$

Finally, n points measures similarity based on coefficient based cluster closed points. The steps for K-mean clustering algorithm are as follows:

1- Input: A: between [0,1], K cluster, t=iteration, A adjacency matrix
2- Calculate initial mean by Jaccard discrete coefficient (J-DIS).
3- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
4- Assign each object to the group that has the closest centroid based on Jaccard discrete coefficient (J-DIS).
5- When all objects have been assigned, recalculate the positions of the K centroids.
6- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

However, it can be proved that the procedure will always terminate, the J-DIS based k-Means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-Means algorithm can be run multiple times to reduce this effect. Algorithm can be run multiple times to reduce this effect. This process iterates until the objective function converges. Where, objective function is the sum of square-error for all objects and every two subset A and B in the database. All subsets converge and produce compact clusters so that they can be well separated from others. The computation complexity of above algorithm is k<<n and t<<n the method mostly terminates at a local optimum. The method often terminates at O (n.k.t) computational complexity.

4. Suitable Data Set

Before, we are able to apply Jaccard Discrete J-DIS k-Means algorithm, we must decide which aspects of data we take into consideration, a process commonly referred to as "feature selection". Most of researchers choose common set of features which prove to have profound effect on the performance of the learning system. The system designer must take great care to pick features that carry a wealth of information. Conversely, considering too many features will invalidate the resulting classifier, as over fitting will result. A line of work about feature was adapted by W. Lee [3] and collaborators attempts to approach feature construction and data mining systematically, seen by.

Table 1: Typical features used for intrusion detection

Service	Service accessed (by port): http, ftp, telnet
Duration	Duration of connection
Src_ip	IP address of the initiator of connection
Dst_ip	IP address of the host
Src_bytes	number of bytes sent by initiator
dst_bytes	number of bytes sent by host
Protocol	TCP, UDP, ICMP, …
num_conn	number of open connections
tcp_flags	TCP Flags (SYN, ACK, RST, …)

In our case the dataset is about intrusion which is best suitable dataset for J-DIS base k-Means clustering algorithm due to intrinsic categorical attributes. Hence, we presented the sensitivity of k-Means to noise and outlier is for small set rather big set KDD Cup99 [1]. But, in this research we are proposing effective feature types for particular type of attack smurf. Therefore, KDD Cup 99 dataset is carrying over 26 features out of 41 features.

Evolution Methods

The intrusion detection model result mainly depends upon Detection Rate (DR) and False Positive Rate (FPR).
1) Detection Rate (DR): The rate at which classifier discriminate the intrusion data and normal data. *2) False Positive Rate (FPR):* It is defined as the rate at which normal data is identified as intrusion data.

Detection Rate: $DR = TP / TP + FP$

False Positive Rate: $FPR = FP / FP + TN$

Experiment Analysis

We have carried out 10% of KDD 99 which is correction version of original dataset. We have taken partial data out of 10% dataset consist of 1075 records only. The test was carried out for the validation of distance function and similarity method J-DIS. There are two main distance functions Euclidian and cosine suitable for KDD 99. Euclidian distance for n-dimensional space is carried out [11].

Distance formula: $\mathrm{d}(\vec{a}-\overrightarrow{b}) = \sqrt{\sum_{k=1}^{n} |a_k - b_k|^2}$

where d is the distance between vectors and is the sum of squares of difference between the coordinates of each vector in n-dimensional space. We are comparing J-DIS $J-Dis = Ai \cap Bj \ / \ Ai > Bj \ + \ Bi > Ai \ + \ c^{*}Ai \cap Bj$ while combining both approaches with k-Means for better classification of attacks and normal data.

Experimental Results and Discussion

KDD Cup 99 dataset is carrying over 26 features out of 41 features [1] and all continues features are carrying numeric values. These features types are having better information gain with better influence on intrusion prediction. We are considering DOS attach with particular type of attack SMURF. We are not considering U2R and U2L other attack types for our experiment because our experiment main objective is to validate J-DIS based k-Means in comparison to distance based approach. Moreover, dataset is partition equally in all initial clusters comprise of two categories; normal and attack as unseen.

Table 2: Feature set for Intrusion

Type	Features
Continues	duration, src_bytes, dst_bytes,hot, logged_in, root_shell, num_file_creations, num_access_files, num_outbound_cmds, count, srv_count, serror_rate, srv_serror_rate, rerror_rate, srv_rerror_rate, same_srv_rate, diff_srv_rate, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, st_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_ host_srv_rerror_rate

Those numeric values are containing values at high scale. There is requirement of normalization process, so we followed normalization scale into two steps.

Step 1: 1) Features with small scale were scaled linearly into range [0.0, 1.0] using the formula (13).

$$\overline{X} = \frac{X - MinX}{MaxX - MinX}$$

Where X is the numerical attribute value, MinX is the minimum value that the attribute X can get and MaxX is the maximum value that the attribute X can get.

Step 2: Features were normalized because of high scale values over fitting for normalization given at step1. Therefore, we applied log base 10 over src_bytes and dst_bytes to scale values between 0 and 1.0 [12].

The major objective is achieved by performing detection of intrusion named as DOS. The final clusters classify by detecting the numerical scaled training dataset consists of 1075 records which is divided into 808 (75%) normal, 264 (24%) DOS attack and 4 (0.003%) named attack. Each record has 26 attributes describing different features and a label assigned to each either as an 'attack' type or as a normal.

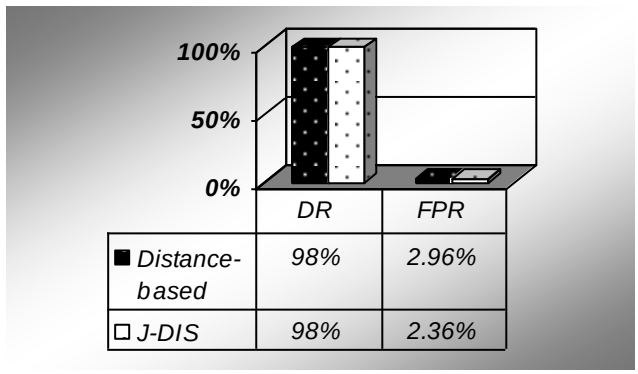

Fig. 1. Comparison of DR and FPR for Distance based k-Means and J-DIS based k-Means

It is clearly observed that theses two types of attacks measured over performance indicators named as Detection Rate (DR) and False Positive Rate (FPR). The distance based k-Means producing same detection rate over iteration k=2 as compare to J-DIS. But, FPR test shows better measures for J-DIS as compared to euclidian distance based function for n-dimension. Further, we have a plan to include various categories of U2R and R2L attacks with 26 features types to produce better percentage of attacks detection.

5. Conclusions

The current research proposed J-DIS based k-Means with an improvement in comparison to presented conventional k-Means Euclidian based distance. The J-DIS approach, thus achieved better performance by better False Positive Rate (FPR) at k=2 in order to determine the optimum number of attacks in all clusters. The results demand different similarity measure J-DIS rather distance based measure such as the Euclidean and Minkowski[10] distance. We like to explore current approaches with diverse and discriminated types of dataset belong to different categories such as clinical informatics and network intrusion detection [2]. Moreover, the J-DIS k-mean approach can be applied over sparse community in Ad Hoc network. Further, in future the approach can be tested to explore cluster based community detection approach where similarity among head nodes and neighbor nodes in clusters require better similarity measures.

References

[1] KDD Cup.1999 data. *http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html*.

[2] Y. Guan and A. A. Ghorbani. (2003). Y-means: *A clustering method for intrusion detection.* In Proceedings of Canadian Conference on Electrical and Computer Engineering, pages 1083{1086}.

[3] L. Portnoy, E. Eskin, and S. Stolfo (2001). Intrusion detection with unlabeled data using clustering. *In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001).*

[4] Xu, S. and L. Cai, 2008. Intrusion detection system based on data mining technology. *J. Elect. Des. Eng., 17(8): 3-5.*

[5] McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 281-297.*

[6] J. Mao, A.K.J. (1996). A self-organizing network for hyper ellipsoidal clustering (HEC). *IEEE Trans. Neural Netw. 7 (1996) 16-29.*

[7] Münz, Gerhard, Sa Li, and Georg Carle. (2007). "Traffic anomaly detection using k-means clustering." Correct Errors Monitor Changes by Gerhard Münz , Sa Li , Georg Carle , *In GI/ITG workshop MMBnet, Conference, Citsee.*

[8] Burbeck, Kalle, and Simin Nadjm-Tehrani. (2005)."Adwice–anomaly detection with real-time incremental clustering." *Information Security and Cryptology–ICISC 2004. Springer Berlin Heidelberg, 407-424.*

[9] Mao, J., Jones, A.k. (1996). A SELF-organizing network for hyper ellipsoidal cluster (hec), *IEEE trans. Neural Network. 7, 16-29.*

[10] A.k. Jain, M.N. Murty, and P.J. Flynn (1999). Data clustering: A survey. *ACM computer. Survey, 31:264-323.*

[11] T. Soni Madhulatha. (2012). An Overview on Clustering Methods, IOSR Journal of Engineering, vol. 2, no. 4, pp. 719-725.

[12] B. Mukherjee and T. Ramesh,(2011). Network Intrusion Detection System using Reduced Dimensionality, *Indian Journal of Computer Science and Engineering*, vol. 2 ,no. 1, pp. 61-6.

[13] H. Timm, C. Borgelt, and R. Kruse. (2004). An Extension of Possibilistic Fuzzy Cluster Analysis, *Fuzzy Sets and Systems*, vol. 147, no. 1, pp. 3-16.

[14] H. Kayacık, A. Zincir-Heywood, and I. M. Heywood. (2005). Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets," in *Proc. the 3rd Annu. Conf. onPrivacy Security and Trust,* ,pp.3-8.