International Journal of Computer Networks and Communications Security

VOL. 3, NO. 5, MAY 2015, 225–228
Available online at: www.ijcncs.org
E-ISSN 2308-9830 (Online) / ISSN 2410-0595 (Print)

The Impact of Distance Metrics on K-means Clustering Algorithm Using in Network Intrusion Detection Data

HADI NASOOTI[1], MARZIEH AHMADZADEH[2], MANIJEH KESHTGARY[3] and S. VAHID FARRAHI[4]

[1, 2, 3, 4] Shiraz University of Technology, Department of Computer Engineering and IT, Shiraz, Iran

E-mail: [1]nasooti.ha@gmail.com, [2]ahmadzadeh@sutech.ac.ir

ABSTRACT

A Network Intrusion Detection System (NIDS) can detect suspicious activities that aimed to harm the network. Since, NIDS help us to keep the networks safer many researchers are motivated to propose more accurate NIDS. K-means clustering algorithm is a distance-based algorithm which widely used in IDS research area. This paper aimed to evaluate the impact of Euclidean and Manhattan distance metrics on K-means algorithm using for clustering KDD cup99 intrusion detection data. Experimental results indicate that Manhattan distance metric performs better in terms of performance evaluation metrics than Euclidean distance metric.

Keywords: *K-means Clustering, Network Intrusion Detection, Euclidean Distance, Manhattan Distance, Data Mining.*

1 INTRODUCTION

In computer security a threat is a possible danger that might use the system vulnerabilities in order to harm the system. An Intrusion Detection System (IDS) can detect suspicious activities that aimed to harm the network. So, an IDS can help network administrators and organizations to keep their networks safer. Up to now, the researchers are motivated in network and information security research area [1] because computer networks have many vulnerabilities and IDS can protect them from intruders.

IDS can be classified into two major categories [2, 3]: Misuse-based Intrusion Detection System (MIDS) and Anomaly-based Intrusion Detection System (AIDS). The major difference between these two categories is in the way of finding intrusive activities. A MIDS tries to identify intrusive behaviors based on the patterns that are available in a database. As a MIDS has well known patterns in a database, its accuracy is high. An AIDS tries to identify the intrusive behaviors that significantly different from normal behaviors. An AIDS is less accurate than a MIDS but can detect previously unseen attacks.

Network anomaly detection is the act of identifying intrusive behaviors in the computer networks. Data mining techniques can improve the accuracy and false alarm rate of Network-based IDS (NIDS). Clustering techniques are one of the most important data mining techniques that work with unlabeled data [4]. Clustering techniques are appropriate for Network anomaly detection since, they can handle unlabeled data.

Clustering algorithms group the data base on their similarities. In other words, the data which are grouped in the same cluster are similar to each other and dissimilar to the data which belong to other clusters. The ability of the clustering algorithms can help us in NIDS. Since, normal behaviors and anomalous behaviors are different from each other, clustering algorithms can group normal behaviors in the same clusters and assign anomalous behaviors to the other clusters [4].

In this paper, K-means clustering algorithm has been utilized to cluster network intrusion detection data. It is proved that K-means clustering algorithm has promising results in network intrusion detection [4]. K-means clustering algorithm needs a distance metric for calculation of the distances between data objects [5]. This paper evaluated the impact of two

different distance metrics on the clustering of network intrusion data.

The remainder of this paper organized as follows: Section 2 presents literature review .Section 3 explains K-means clustering algorithm and distance metrics .Section 4 presents the motivation of our paper. Section 5 presents simulation parameters and performance metrics. Finally, section 6 concludes the paper and explains the results.

2 RELATED WORKS

In [4] K-means clustering algorithm has been used for network intrusion detection. Simulation results show that K-means clustering algorithm is able to detect unknown intrusions in the network connections. Euclidean metric has been used as the distance metric.

In [6] the authors proposed an algorithm based on K-means clustering algorithm for network intrusion detection. The proposed algorithm can define the number of clusters automatically based on a predefined threshold.

Y-Means [7] is a clustering algorithm for network intrusion detection based on K-means clustering algorithm. Y-means is able to define the number of clusters automatically and overcome the shortcoming of producing empty clusters in K-means algorithm.

In [8] five different clustering algorithms including K-means clustering algorithm and four different classifiers have been evaluated for network intrusion detection. The results show that the distance-based clustering algorithms are more accurate in detecting unknown attacks in comparing to the classifiers.

All the previous works used Euclidean distance metric for calculation of the distances. None of the previous works had not employed another distance metric such as Manhattan distance in K-means algorithm. This paper is aimed to analysis the impact of different distance metrics on K-means clustering algorithm using in network intrusion detection. So, the impact of Manhattan and Euclidean distance metrics on K-means algorithm using in network intrusion detection has been evaluated in this paper.

3 K-MEANS CLUSTERING ALGORITHM

Clustering is one the most important data mining techniques that can handle unlabeled data. K-means is a distance-based clustering algorithm. K-means groups the data objects into *K* disjoint clusters. *K* is a user specified parameter. Since, K-means is a simple algorithm it has been used in a wide variety of applications [5] as well as network intrusion detection. It is one of the most important clustering algorithms in data mining area.

K-means algorithm is described as follows [5]:

1) Choose K data objects for the cluster centers, randomly.
2) Calculate the distances between each data object and all cluster centers base on a distance metric.
3) Assign data objects to the closet cluster center.
4) Update cluster centers.
5) Iterate steps (2)-(4) until there is no more updating.

A distance metric must be defined for distance calculation between data objects in K-means algorithm. The most common distance metric that uses in K-means clustering algorithm is Euclidean distance metric. Although, Euclidean distance metric is the most common metric, Manhattan distance metric can be used in K-means algorithm.

Consider X, Y are two data objects which described by n attributes. The Euclidean and Manhattan distances between data objects X, Y are denoted as Euclid(X,Y) and Manhat(X,Y), respectively. x_j and y_j are the j_{th} attribute values of data objects X and Y, respectively. Euclidean and Manhattan distance metrics are defined as follow:

$$Euclid(X,Y) = [\sum_{j=1}^{n} (x_j - y_j)^2]^{\frac{1}{2}} \qquad (1)$$

$$Manhat(X,Y) = \sum_{j=1}^{n} | x_j - y_j | \qquad (2)$$

4 MOTIVATION

Almost, heuristic methods are used for The distance metric is one of the key steps in the unsupervised learning process [9]. Many parameters effect the final clustering results of K-means algorithm such as, algorithm initialization and distance metric. As changing the distance metric might change the final clustering results so, it is important to evaluate the impact of other distance metrics on K-means algorithm in clustering network intrusion data. This paper is aimed to analysis the impact of Euclidean and Manhattan distance metrics on the application of network intrusion detection

5 SIMULATION DESIGN

In this paper, we used KDDcup 99 network intrusion data subset [10]. KDDcup intrusion detection data set consisted of 4 attack types plus normal type activities. The four attack types are Denial of Service (DoS), Remote to User (R2L), User to Root (U2R) and Probing. 10% KDDcup 99 has 494,021 data objects. We sampled 72,784 data objects for our experiments. The experiments have been done with five clusters (*K* is five in our experiments).

The data have been normalized based on Min-Max method [11]. Formula 3 shows Min-Max normalization method where MIN(i) and MAX(i) are the minimum and maximum value of i_{th} attribute, respectively. X corresponds to an unnormalized attribute value before transformation and X* corresponds to the normalized value of X. By using Formula 3 all data were scaled in the range [0-1].

$$X^* = \frac{X - MIN(i)}{MIN(i) - MAX(i)} \qquad (3)$$

5.1 *Performance Evaluation Metrics*

The impact of Manhattan and Euclidean distances metric in K-means algorithm have been evaluated in terms of accuracy, detection rate and false alarm rate. These performance metrics are as following:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (4)$$

$$\text{Detection Rate} = \frac{(TP)}{(TP + FP)} \qquad (5)$$

$$\text{False Alarm Rate} = \frac{(FP)}{(FP + TN)} \qquad (6)$$

- True positive (TP): number of attack data objects that are correctly classified as intrusion.
- True negative (TN): number of normal data objects that are correctly classified as normal.
- False positive (FP): number of normal data objects that are incorrectly classified as attacks.
- False negative (FN): number of data objects that are incorrectly classified as intrusion.

6 RESULTS

The performance of Euclidean and Manhattan using as distance metrics in K-means algorithm have been shown in Fig. 1. Manhattan distance metric performs better than Euclidean distance metric in K-mean clustering algorithm in terms of accuracy and detection rate.

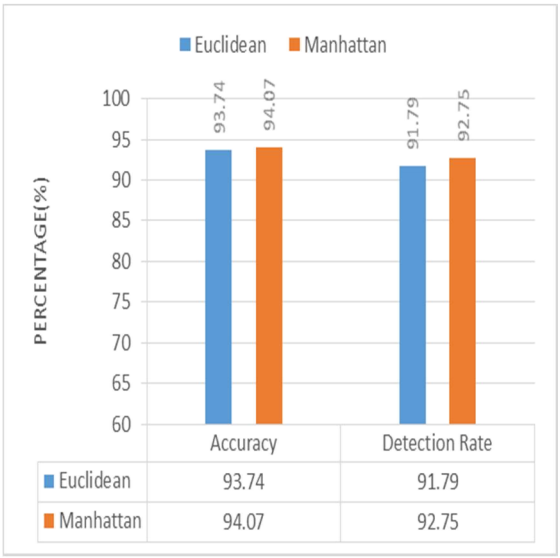

Fig. 1. Comparison of Detection Rate and Accuracy between Euclidean and Manhattan Metrics.

The false alarm rate of Euclidean distance metric is 5.23% and the false alarm rate of Manhattan distance metric is 4.72%. It means that Manhattan distance metric is superior to Euclidean distance metric in term of false alarm rate.

Generally, Manhattan distance metric has better performance than Manhattan distance metric using in K-means algorithm for clustering network intrusion detection data.

As a consequence, the impact of distance metrics should be evaluated for other distance based algorithm using in network intrusion detection area since, other distance metrics might have better results.

Many researchers such as [6, 7] proposed distance-based algorithms for network intrusion detection but the authors only used Euclidean distance metric for calculation of the distances. Other distance metrics such as Manhattan metric might have better results in the proposed algorithms.

7 REFERENCES

[1] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," Journal of Network and Computer Applications, vol. 36, pp. 16-24, 2013.

[2] P. Kabiri and A. A. Ghorbani, "Research on Intrusion Detection and Response: A Survey," IJ Network Security, vol. 1, pp. 84-102, 2005.

[3] S. Axelsson, "Intrusion detection systems: A survey and taxonomy," Technical report2000.

[4] M. Jianliang, S. Haikun, and B. Ling, "The application on intrusion detection based on k-means cluster algorithm," in Information Technology and Applications, 2009. IFITA'09. International Forum on, 2009, pp. 150-152.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM computing surveys (CSUR), vol. 31, pp. 264-323, 1999.

[6] L. Portnoy, "Intrusion detection with unlabeled data using clustering," 2000.

[7] Y. Guan, A.-A. Ghorbani, and N. Belacel, "Y-means: A clustering method for intrusion detection," 2003.

[8] I. Syarif, A. Prugel-Bennett, and G. Wills, "Unsupervised clustering approach for network anomaly detection," in Networked Digital Technologies, ed: Springer, 2012, pp. 135-145.

[9] K. Doherty, R. Adams, and N. Davey, "Unsupervised learning with normalised data and non-Euclidean norms," Applied Soft Computing, vol. 7, pp. 203-210, 2007.

[10] KDD Cup 1999 Data. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[11] G. W. Milligan and M. C. Cooper, "A study of standardization of variables in cluster analysis," Journal of classification, vol. 5, pp. 181-204, 1988.