Chaos, Solitons and Fractals 110 (2018) 33–40

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Nearest neighbors based density peaks approach to intrusion detection

Lixiang Li[a,b,c,*], Hao Zhang[b,c], Haipeng Peng[b,c], Yixian Yang[b,c,d]

[a] School of Computer Science and Technology, Henan Polytechnic University, 2001 Century Avenue, Jiaozuo, Henan, 454003, China
[b] Information Security Center, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[c] National Engineering Laboratory for Disaster Backup and Recovery, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[d] State key Laboratory of Public Big Data, Guizhou, 550025, China

ARTICLE INFO

Article history:
Received 28 November 2017
Accepted 5 March 2018

Keywords:
Intrusion detection
KDD-CUP 99
k-nearest neighbors (kNN)
Density peaks clustering (DPC)
Density peaks nearest neighbor (DPNN)

ABSTRACT

Intrusion detection systems are very important for network security. However, traditional intrusion detection systems can not identify new type of network intrusion for example zero-day attack. Many machine learning techniques were used in intrusion detection system and they showed better detection performance than other methods. A novel clustering algorithm called Density peaks clustering (DPC) which does not need many parameters and its iterative process is based on density. Because of its simple steps and parameters, it may have many application fields. So we are going to use it in intrusion detection to find a more accurate and efficient classifier. On the basis of some good ideas of DPC, this paper proposes a hybrid learning model based on *k*-nearest neighbors (kNN) in order to detect attacks more effectively and introduce the density in kNN. In density peaks nearest neighbors (DPNN), KDD-CUP 99 which is the standard dataset in intrusion detection is used to the experiment. Then, we use the dataset to train and calculate some parameters which are used in this algorithm. Finally, the DPNN classifier is used to classify attacks. Experiment results suggest that the DPNN performs better than support vector machine (SVM), *k*-nearest neighbors (kNN) and many other machine learning methods, and it can effectively detect intrusion attacks and has a good performance in accuracy.

1. Introduction

With the rapid development of computer network technology and the dramatic increment of computer users, the information security becomes more important on computer network. In order to protect the security of computer and Internet, traditional security defense measures, such as identity authentication and access control, have exposed many defects or vulnerabilities [1]. Intrusion detection which is emerged on the information and network security becomes one of the key technologies. It is not affected by time and space limitations, the characteristics of attack strike means hidden and more intricate, internal commit crime continually. Intrusion detection becomes a new solution for network security strategy because of its new character [2].

James Aderso firstly introduced intrusion detection in 1980, which was the beginning of the intrusion detection research. Since then, many methods were used in intrusion detection, such as intelligent algorithm [3–5] and clustering [6–8]. In the past few years, many machine learning methods were used in intrusion detection [9–11]. Chih-Fong Tsai came up with the triangle area based nearest neighbors approach (TANN) [12]. In TANN, the *k*-means clustering was firstly used to obtain cluster centers corresponding to the attack classes respectively. It can detect attacks more effectively. Wei-Chao Lin introduced an intrusion detection system based on combining cluster centers and nearest neighbors[13]. CANN classifier was not only similar to but also performed better than kNN and supported vector machines which were trained and tested by the original feature representation in terms of classification accuracy. Abdulla Amin Aburomman gave a new idea that used PSO to generate weights to create ensemble of classifiers [14]. There are many algorithms based on SVM or deep learning which are applied to intrusion detection [15–18]. These methods demonstrate the advantages of machine learning in intrusion detection.

Although intrusion detection is constantly developing, it still exists many problems and challenges [19]. Firstly, the KDD-CUP 99 is the standard data set in the field of intrusion detection, it has a total of 41-dimensional feature [20,21]. Such a high dimension has an enormous influence on efficiency, so the dimension must

[*] Corresponding author at: Information Security Center, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

E-mail address: li_lixiang2006@163.com (L. Li).

https://doi.org/10.1016/j.chaos.2018.03.010
0960-0779/© 2018 Elsevier Ltd. All rights reserved.

be reduced [22–24]. Secondly, many methods have been proposed to improve the accuracy of Dos that is one of the most common attacks in intrusion detection. Although Probe and R2L often appear in the attack, there is few work considering them. To solve these two problems, many hybrid methods appeared [25].

kNN is one of the most basic methods of machine learning which is based on distance. It has many advantages, such as ease of implementation, suitable for multi-class classification and less estimate parameters [26]. But kNN has a bad performance when the data is unbalanced. The peak density clustering is an algorithm which can efficiently solve the problem of many kinds of clustering [27]. Similar to *k*-means clustering methods [28], the algorithm is only based on the distance between the two classes. Meanwhile the same as DBSCAN [29] and mean shift clustering algorithm, the peak density clustering can not only classify the non-spherical accurate data sets into different clusters, but also can automatically find the number of clusters [30]. But it is different from mean shift algorithm, the peak density algorithm does not need to embed data into vector space, or increase the density of each class [31]. As far as we know, these two methods are effective. But there is few attention on mixing DPC and kNN together.

In this paper, we propose a novel method DPNN using the idea of density [27,32] and the process of the k-nearest neighbors to detect attacks [33]. First of all, the data is preprocessed and the features are selected by SVM. Then the density is introduced into kNN as a standard. And it is used to filter the data with distance. Due to the density, kNN has stronger robustness to some specific data. And it can solve the problem which is hard to choose some parameters of DPC. The experiment results prove that this method can greatly improve the accuracy for intrusion detection. And DPNN can reduce testing time by screening some insignificant data. The time of algorithm mainly comes from the parameters calculation in training set, practical application when we have already trained the training set. DPNN spends little time on classifying the abnormal data. So DPNN has a very strong practicability.

The rest of this paper is organized as follows: Section 2 briefly describes two classical methods used in this paper. Section 3 introduces the feature selection and DPNN. Section 4 gives many groups of comparing experiment results. Conclusion and future work are provided in Section 5.

2. Basic method

Although data mining has developed for many years, it still has many areas that can be consummated [30,34]. We proposed a method DPNN combining the core ideas of kNN and DPC [27]. This section provides brief reviews of kNN and DPC.

2.1. kNN classifier

The k-nearest neighbor is a kind of typical nonparametric algorithm [35]. The judgement standard of kNN is the distance [36]. Compared with other classification algorithms, it is simpler and more efficient [37]. For a given point, kNN records the neighbors of a testing sequence among the training data, and uses the labels of the nearest neighbors to assign the testing sample. The parameter *k* of kNN is the core of the method. So the variation of *k* will influence the accuracy which is the standard evaluation criteria in machine learning. In the contrast experiments, to ensure the diversity of methods and to maximally explore the potential of kNN, we will choose different values of parameter *k*.

2.2. Density peak clustering

Density peak clustering is a kind of clustering algorithm based on density [38,39]. One of the main assumption is that each cluster center is surrounded by a lower local density of the adjacent points, and the adjacent points distance density is higher than the other distances which are relatively far from the center of the clusters [40]. Two quantities need to be calculated for each object in density peak clustering: the local density of object is higher than the density of points. The size of the amount of ρ_i and $\varphi(x)$ are only related to the distance between two objects. We defined the local density of the object as formula:

Table 1
Dataset size of KDD-CUP 99.

	Normal	Probe	Dos	U2R	R2L
Testing data	60,593	4166	231,455	88	14,727
Training data	97,277	4107	391,458	52	1126

$$\rho_i = \sum_j \varphi(d_{ij} - d_c), \tag{1}$$

$$\varphi(x) = \begin{cases} 1 & x < 0, \\ 0 & x > 0. \end{cases} \tag{2}$$

d_c is to specify a threshold, $\varphi(x)$ is defined as a formula. Generally speaking, ρ_i is equal to the distance from the object i which is smaller than d_c by the number of all the classes. Through the analysis of formula, the peak density algorithm in large data sets is only sensitive to d_c. Then it calculates the shortest distance and the high density of the class.

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i}(d_{ij}) & i \text{ is not density maximum points}, \\ \max(d_{ij}) & i \text{ is density maximum points}. \end{cases} \tag{3}$$

For the highest density of points, we usually use $\delta_i = \max(d_{ij})$ to represent the density of δ_i which is greater than the adjacent local or global density of maximum density. So the point of δ_i is regarded as a cluster center.

3. Preparatory work and rule evaluation

3.1. Dataset description

In this paper, we used kDD-Cup 99 data set which was proposed in 1998 by MIT Lincoln Laboratory. Since it was put forward, it has been regarded as the standard data set for intrusion detection up to now. KDD-CUP 99 data set contains 494,020 samples. Each sample which represents a network connection is composed by 41-dimensional feature vector.

Each sample of kDD-Cup 99 data set is labeled as normal or abnormal. Abnormal is subdivided into four major categories, a total of 39 types of attacks:

- Probe: A connection attempts to search potential dangerous on target machine.
- Denial of Service(Dos): A connection attempts to cause the disruption of service or let the target machine crashed.
- User to Root(U2R): An attacker who has already gained access wants to gain the privileges of super user.
- Remote to Local(R2L): An attacker attempts to gain illegal access to a remote computer.

We choose five data sets which are taken from testing and training kDD-CUP 99 randomly. All data sets detailed written in Table 1 are the same size. And the original data sets of kDD-CUP 99 are presented in Table 2.