A Low Complexity Intrusion Detection Algorithm

Lin Yao[1] and Kai Yao[2]

[1] Software College of Dalian University of Technology, Dalian 116023, China
lin_yao@eyou.com
[2] Library Information Center of Shenyang University of Technology, Shenyang 150001, China
frantodd2002@yahoo.com.cn

Abstract. A low complexity clustering algorithm for intrusion detection based on wavecluster is presented. Using the multiresolution property of wavelet transforms, we can effectively identify arbitrarily shaped clusters at different scales and degrees of detail, moreover, applying wavelet transform removes the noise from the original feature space and make more accurate cluster found. Experimental results on KDD-99 intrusion detection dataset show the efficiency and accuracy of this algorithm. A detection rate above 98% and a false alarm rate below 3% are achieved. The time complexity of the wavecluster algorithm is $O(N)$, which is comparatively low than other algorithms.

1 Introduction

Network Intrusion detection is the process of monitoring the events occurring in a computing system or network and analyzing them for signs of intrusions, defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanism of a computer or network. Recently, many researchers turned into data mining techniques to attack the problem. Data mining can improve variants detection rate, control false alarm rate, and reduce false dismissals. A wide variety of data mining techniques has been applied to intrusion detections. In data mining, clustering is the most important unsupervised learning process used to find the structures or patterns in a collection of unlabeled data. Until now, the clustering algorithms can be categorized into four main groups: partitioning algorithm, hierarchical algorithm, density-based algorithm and grid-based algorithm[1],[2],[3]. Gholamhosein proposed wavecluster approach, which is a grid-based approach, and successfully used in image processing[4].We extend wavecluster to use in intrusion detection, but traffic in a network is never stagnant, for example, new services can be activated, work patterns can change as projects start or finish, and so on. Consequently, an intrusion detection system needs to be able to adapt to these changing traffic patterns while still maintaining a high level of detection accuracy. To deal with these issues, we modify wavecluster and develop a low complexity wavecluster algorithm with little prior knowledge, the proposed method allows the discovery of clusters of any shape and can detect timely not only the known intrusion types, but also their variants.

Y. Shi et al. (Eds.): ICCS 2007, Part III, LNCS 4489, pp. 461–464, 2007.

The rest of the paper is organized as follows. Section 2 discusses the low complexity wavecluster method, in section 3, we present the experimental evaluation of the wavecluster method using KDD-99 intrusion detection dataset. Finally, we present concluding remark and suggestions for future study.

2 Low Complexity Wavecluster Algorithm

Given a set of spatial objects, the goal of the adaptive algorithm is to detect cluster and assign labels to the objects based on the cluster that they belong to, thus it can detect whether there is intrusion exists. Applying wavelet transform to transform the original feature space and then find the dense regions in the new space. It yields sets of clusters at different resolutions and scales, which can be chosen based on user needs. The main step of adaptive wavecluster algorithm are shown in Algorithm 1.

Algorithm 1.
Input: Multidimendional data objects feature vectors
Output: clustered objects
1. Quantize feature space, then assign objects to the cells.
2. Apply wavelet transform on the quantized feature space.
3. Find the connected components in the subbands of transformed feature space.
4. Assign labels to the cells
5. Map the objects to the clusters

The first step of the adaptive wavecluster algorithm is to quantize the feature space, where each dimension A_i in the d-dimensional feature space will be divided into m_i intervals. Then, corresponding cell for the objects will be determined based on their feature values. A cell $C_i=[c_{i1},c_{i2},...c_{id}]$contains an object $O_i=[o_{k1},o_{k2},...o_{kd}]$.

We recall that c_{ij} is the right open interval in the partitioning of A_j. For each cell, we count the number of objects contained in it to aggregation of the objects. The quantization m_i effect the performance of this algorithm, we set the correct value of m_i for intrusion detection dataset by simulation. In the second step, discrete wavelet transform will be applied on the quantized feature space. Applying wavelet transform on the cells in $\{C_j:1<j<h\}$ results in a new feature space, that is new cells. Given the set of new cells, wavecluster detects the connected components in the transformed feature space. Each connected component is a set of new cells and is considered as a cluster. For finding the connected components, we define that a new neighborhood for intrusion detection application, that is ,the neighborhood is defined in Euclidean space, a significant cell a in the transformed feature space is neighbor of another cell b if a lies within one of the four grid cells surrounding cell b,with the shortest Euclidean distances is defined to be its nearest neighbors. With an indexing scheme such as an R-tree, it is very easy to find the neighborhoods in short time.

Each cluster w will have a cluster number w_n, the adaptive wavecluster algorithm labels the cells in each cluster in the transformed feature with its cluster number. Calculate the mean of every cluster, remark as x_n. The clusters that are found are in the transformed feature space and are based on wavelet coefficients. Thus, they cannot be directly used to define the clusters in the original feature space. We make a lookup table to map the cells in the original feature space. Each entry in the table specifies the relationship between one cell in the transformed feature space and the corresponding cell in the original feature space. Finally, the algorithm assigns the label of each cell in the feature space to all the objects whose feature vector is in that cell, and thus the cluster as determined.

When the objects are assigned to the cells of the quantized feature space at step 1 of the algorithm, the final content of the cells is independent of the order in which the objects are presented performed on these cells. Hence, the algorithm will have the same results for any different order of input data, so it is order insensitive with respect to input objects.

3 Experimental Results

In order to evaluate the low complexity wavecluster algorithm, we test the algorithm on a benchmark dataset, the network traffic data from the KDD Cup 1999 dataset. In the experiments, the values of each features are normalized with the minimum and maximum values of that features so that they fall in the range of [0,1]. In our experiment, beside the normal instances, instances of three popular attacks are involved: ipsweep, smurf, neptune. For each type, only one seed point is labeled at the beginning, and four nearest neighbors are defined as the nearest neighborhood of each instance. As shown in Table I, most attacks can be distinguished from the normal activities and the detection rate is as high as 98.3%. At the same time, the false alarm rate is approximately 1.8%. Assuming m cells in each dimension of feature space, there would be $K=m^d$ cells. Complexity of applying wavelet transform on the quantized feature space will be $O(dK)$.Since we assume that the value of d is low, we can consider it as a constant, thus $O(dK)=O(K)$.Making the lookup table requires $O(K)$ time. The time complexity of last step of our low complexity wavecluster algorithm is $O(N)$. Since this algorithm is applied on very large databases with a low number of dimensions, we can assume that $N>K$.Thus, based on this assumption, the overall time complexity of the algorithm will be $O(N)$.

Table 1. Experiment Result

Experiment	Detection Rate	False Alarm Rate
First	98.3%	1.8%
Second	98.7%	1.7%

4 Conclusion

In this paper, we present a low complexity wavecluster algorithm for intrusion detection, which can adapt to changes in normal traffic. Experimental results on a subset of KDD-99 dataset showed the stability of efficiency and accuracy of the adaptive wavecluster algorithm. With different setting, the detection rate stayed always above 98% while the false alarm rate was below 3%. The time complexity of adaptive wavecluster is low, which is $O(N)$,N is the number objects in the database.

Acknowledgment

The authors are gratefully acknowledge support from NSF grants 60673046,and thank Professor Mingchu.Li for providing helpful comments and information.

References

1. Pauwels EJ, Fiddelaers P,Van Gool L. DOG-based unsupervised clustering for CBIR. In Proceedings of the 2nd International Conference on Visual Information Systems, Vol.2, San Diego,Calif (2003) 137-145.
2. Wang W, Yang J,Muntz R.STING: A Statistical Information Grid Approach to Spatial Data Mining. In Proceedings of the 23rd VLDB Conference, Athens, Vol.30, Greece(2000) 186-195.
3. Xu X,Ester M,Kriegel H. A distribution-based clustering algorithm for mining in large spatial database. In Proceedings of the 14th International Conference on Data Engineering, Vol.41,Orlanda,Fla.(2002)324-331.
4. Gholamhosein S. et al, WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. Journal of VLDB(2000) Vol.33, 289-294.