NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

An Unsupervised Clustering Algorithm for Intrusion Detection

Guan, Y.; Ghorbani, Ali-Akbar; Belacel, Nabil

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

NRC Publications Record / Notice d'Archives des publications de CNRC:
https://nrc-publications.canada.ca/eng/view/object/?id=2820b823-8731-4927-a235-4050e28fe6bf
https://publications-cnrc.canada.ca/fra/voir/objet/?id=2820b823-8731-4927-a235-4050e28fe6bf

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada | Conseil national de recherches Canada

Canada

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*An Unsupervised Clustering Algorithm for Intrusion Detection**

Guan, Y., Ghorbani, A., and Belacel, N.
June 2003

* published in Advances in Artificial Intelligence, The 16th Conference of the Canadian Society for Computational Studies of Intelligence (AI 2003). Halifax, Nova Scotia, Canada. June 11-13, 2003. pp. 616-117. NRC 45843.

Copyright 2003 by
National Research Council of Canada

Permission is granted to quote short excerpts and to reproduce figures and tables from this report, provided that the source of such material is fully acknowledged.

Canada

AN UNSUPERVISED CLUSTERING ALGORITHM FOR INTRUSION DETECTION

Yu Guan[1], Ali A. Ghorbani[1], and Nabil Belacel[2]

[1] Faculty of Computer Science
University of New Brunswick
Fredericton, NB, E3B 5A3
{guan.yu, ghorbani}@unb.ca
[2] E-health, Institute for Information Technology
National Research Council
Saint John, NB, E2L 2Z6
{nabil.belacel}@nrc.ca

1 Introduction

As the Internet spreads to each corner of the world, computers are exposed to miscellaneous intrusions from the World Wide Web. Thus, we need effective intrusion detection systems to protect our computers from the intrusions. Traditional instance-based learning methods can only be used to detect known intrusions since these methods classify instances based on what they have learned. They rarely detect new intrusions since these intrusion classes has not been learned before. We expect an unsupervised algorithm to be able to detect new intrusions as well as known intrusions.

In this paper, we propose a clustering algorithm for intrusion detection, called Y-means. This algorithm is developed based on the H-means+ algorithm [2] (an improved version of the K-means algorithm [1]) and other related clustering algorithms of K-means. Y-means is able to automatically partition a data set into a reasonable number of clusters so as to classify the instances into 'normal' clusters and 'abnormal' clusters. It overcomes two shortcomings of K-means: *degeneracy* and *dependency on the number of clusters* .

The results of simulations that run on KDD-99 data set [3] show that Y-means is an effective method for partitioning large data set. An 89.89% detection rate and a 1.00% false alarm rate were achieved with the Y-means algorithm.

2 Y-means Algorithm

The amount of normal log data is usually overwhelmingly larger than that of intrusion data. Normal data are usually distinguished from the intrusions based on the Euclidean distance. Therefore, the normal instances form clusters with large populations, while the intrusion instances form remote clusters with a relatively small populations. Therefore, we can label these clusters as normal or intrusive according to their populations.

Y-means is our proposed clustering algorithm for intrusion detection. By splitting clusters and merging overlapped clusters, it is possible to automatically

partition a data set into a reasonable number of clusters so as to classify the instances into 'normal' clusters and 'abnormal' clusters. It also overcomes the shortcomings of the K-means algorithm.

We partitioned 2,456 instances of KDD-99 data using the H-means+ algorithm with different initial values of k. The decline of SSE is fast when the value of k is very small. When the value of k reaches 20, the decline of Sum of Square Error (SSE) becomes slow. In this experiment, the optimal value for k is found to be around 20. At this point, we obtained a 78.72% detection rate and a 1.11% false alarm rate. This result is probably the best that we can get with H-means+.

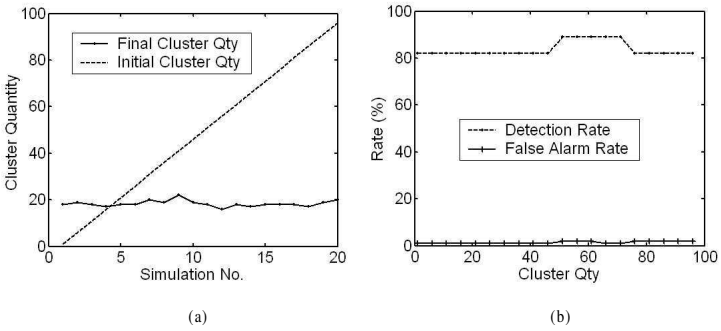

Fig. 1. a. Initial number vs. final number of clusters; b. Y-means with different initial number of clusters.

Y-means partitioned the same data set into 16 to 22 clusters as shown by the approximately horizontal line in Figure 1 (a), when the initial number of clusters varied from 1 to 96. On average, the final number of clusters is about 20. This is exactly the value of the 'optimal' k in H-means+. On average, the Y-means algorithm detected 86.63% of intrusions with a 1.53% false alarm rate as shown in Figure 1 (b). The best performance was obtained when detection rate is 89.89% and false alarm rate is 1.00%.

In conclusion, the Y-means is a promising algorithm for intrusion detection, since it can automatically partition an arbitrarily sized set of arbitrarily distributed data into an appropriate number of clusters without supervision.

References

1. MacQueen, J.B. "Some methods for classification and analysis of multivariate observations." Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol.2, pp.28-297, 1967.
2. Hansen, P. and N. Mladenovic "J-Means: a new local search heuristic for minimum sum-of-squares clustering." Pattern Recognition 34 pp.405-413, 2002
3. KDD Cup 1999 Data. University of California, Irvine. http://kdd.ics.uci.edu /databases/kddcup99/kddcup99.html, 1999.