When an iris flower is normal then are others abnormal?

Akira Imada

Brest State Technical University
Moskowskaja 267, 224017 Brest, Republic of Belarus
akira@bsty.by

Abstract: *When we design a system for anomaly detection, we sometimes have to train and then test the system by using an artificial data samples. We could find a lot of such databases like iris database in public domain. And we have so many such reports. However, we doubt, more or less, these data cannot be necessarily representative of normal and abnormal data of real transactions. For example, we usually don't know what does a coming anomaly pattern look like until it has completed the illegal intrusion. In this paper, we give it a little consideration to this problem using some lately proposed artificial immune based technique of network intrusion detection by exploring the iris database.*

Keywords: *artificial immune system, negative selection, network intrusion detection, detection of self and non-self cells, iris database*

1 Introduction

Network Intrusion detection is one of the hottest issues in our computer society nowadays. Some of the approaches such as Hong et al. (2004) [1] used a real data, some others such as Gomez et al. (2004) [2] used the 1998 DARPA intrusion detection evaluation data-set prepared by MIT, also known as KDD cup 99 data-set.[1] We have also a newer such public domain data-set called KDD cup 2003 which was used in the data mining competition held in conjunction with the Ninth Annual ACM SIGKDD Conference.[2] On the other hand, an artificial data-set to simulate normal and anomaly patterns is often employed. Gomez et al. [2], for example, used the Mackey-Glass equation to generate time series data as artificial transactions.

Since we usually don't know what does a coming anomaly pattern look like until it has completed the illegal intrusion, what we think of the key issue is if these anomaly detection systems can eventually evolve to detect an unknown anomaly pattern only by training using a known set of normal patterns. As such training samples Ayara et al. [3] used randomly generated 8-bit binary patterns assuming they are normal patterns. We picked up anormal patterns from what we call *a-tiny-island-in-a-huge-lake* [4] or as its extreme case we assumed the needle of so-called *a-needle-in-a-haystack* as anomaly (See the other of our paper in this volume) [5].

In this paper, we give it a consideration to the idea of using iris database[3] among others as an example of normal samples as well as abnormal samples. Castellano et al. (2000) [6] evaluated their network intrusion system using this iris database. They assume the data of one iris family as normal while others as abnormal, and trained the system using both classes, which is what we doubted, more or less. One reason why we thought in that way is if usually unknown illegal patterns can be represented by certain specific pattern or not, as already mentioned. The other reason is the sparseness of this data-set. What if the system came across a pattern which does not belong to either of the two classes

[1] http://kdd.ics.uci.edu/databases/kddcup99
[2] http://www.cs.cornell.edu/projects/kddcup

[3] ics.uci.edu: pub/machine-learning-databases

2 Iris Database

Castellano et al. [6] clearly described their data-set as:

> *The validity of our approach to fuzzy inference and rule extraction has been tested on the well-known benchmark Iris data problem. The classification problem of the Iris data consists of classifying three species of iris flowers (setosa, versicolor and virginica). There are 150 samples for this problem, 50 of each class. A sample is a four-dimensional pattern vector representing four attributes of the iris flower (sepal length, sepal width, petal length, and petal width).*

Then our concerns, as already mentioned in the previous section, are two fold. One is "Can usually unknown illegal patterns be represented by certain specific pattern?" The other is "What if the system comes across a pattern which does not belong to either of the two classes?"

3 Algorithm

We employ, among others, the following two algorithms proposed by Zhou Ji and Dasgupata (2004) [7]. Principally the method is made up of (1) generating candidate detectors randomly; (2) checking them one by one if it matches self patterns; (3) eliminating if it matches self, otherwise put it in the repertoire. This constructs detectors set which detects non-self. The goal of these algorithms is to create detectors which cover non-self space as much as possible. One is called *"Augmented Negative Selection Algorithm with Variable-Coverage Detectors"*, and the other is its simpler version called *"Constant-Coverage Detectors"* in which detector size is constant instead of variable. The followings are these two algorithms that we paraphrased the original ones with the semantics being intact. First, the simpler version is:

Algorithm 1 (Constant-sized Detectors) *After setting (i) N_t, the number of training samples; (ii) r_d, the radius of detector; and (iii) N_d, the total number of detectors:*

1. *Create N_s samples of self cells at random.*
2. *Create a hyper-sphere which has the radius r_d and whose center locates at random in $[-1, 1]$. This is a candidate detector to detect non-self cells.*
3. *If this-hyper sphere does not contain any sample self cells, then put it as a detector in D, the detector's repertoire. Otherwise delete the hyper-sphere.*
4. *Repeat 2-3 until we find N_d detectors.*

Then second,

Algorithm 2 (Variable-sized Detectors) *After setting (i) N_t, the number of training samples; (ii) r_s, the radius of self cells; (iii) c_0, expected coverage, i.e., the degree to how much those created detectors cover non-self cells; (iv) $c_{\max}$, the upper bound of self coverage; and (v) N_d, the maximum number of detectors:*

1. *Empty D, the detector's repertoire.*
2. *Try to find a point $\mathbf{x} = (x_1, \cdots, x_n) \in [-1, 1]^n$ which is not contained by any of the valid detectors so far created, unless the number of those trials exceeds $1/(1 - c_0)$. If no such $\mathbf{x}$ is found, then terminate the run.* [4]
3. *If r, the distance between $\mathbf{x}$ and its closest self sell in the training sample, is larger than the radius r_s, i.e., if the candidate doesn't include any of the sample self cells, then add the sphere whose center is $\mathbf{x}$ and radius is r to D as a new valid detector.*
4. *If such $\mathbf{x}$ cannot be found within the consecutive trials of $1/(1 - c_{\max})$ time, then terminate the run.* [5] *Otherwise repeat 2 and 3, until we find a total of N_d detectors.*

We do not think these two algorithms strongly reflect a concept of immune system, despite the title of the original paper indicates it. However at least the title holds true in the sense that detectors are chosen by trying to match them to the self strings and if a detector matches then it is discarded, otherwise it is kept. This is, above all, what we call a natural selection algorithm.

4 Evaluation of How it Works

We use a measure originally proposed by Lopes et al. [8] in which four quantities, i.e., (i) true-positive, (ii) true-negative, (iii) false positive, and (iv) false negative are used. Here we assume positive sample is non-self and negative sample is self, since detectors is designed to detect non-self cells. Hence, these

[4] This is because when we have sampled m points and only one point was not covered, the expected coverage is $1 - 1/m$. Hence the necessary number of tries to ensure expected coverage c_0 is $m = 1/(1 - c_0)$.

[5] See also the footnote above replacing c_0 with $c_{\max}$.

four terms are defined in a sence that (i) t_p (true positive) — true declaration of positive sample, i.e., non-self declared as non-self (ii) f_p (false positive) — false declaration of positive sample, i.e., self declared as non-self (iii) t_n (true negative) — true declaration of negative sample, i.e., self declared as self (iv) f_n (false negative) — false declaration of negative sample, i.e., non-self declared as self. Under these definitions $d_r = t_p/(t_p + f_n)$ implies detection rate, and $f_a = f_p/(t_n + f_p)$ implies false alarm rate. Then we plot d_r versus f_a, and the resultant graph is called *Receiver Operating Characteristics (OCR)* [9] which reflects a tradeoff between false alarm rate and detection rate.

5 Summary

We have described our doubt about an usage of an artificial data-set. We think this issue remains un open and important in considering how we design a network intrusion system. Although we have not started this work, it almost ready to start. We hope a lot of experiments which might result in positive and negative observation await our exploration.

References

[1] Xuan Dau Hoang, and Jiankun Hu (2004) "An Efficient Hidden Markov Model Training Scheme for Anomaly Intrusion Detection of Server Applications Based on System Calls." Proceedings of Internation Conference on Networks, Vol. 2, pp. 470-474.

[2] J. Gomez, F. Gonzalez, and D. Dasgupta (2003) "An Immuno-Fuzzy Approach to Anomaly Detection" proceedings of the 12th IEEE International Conference on Fuzzy Systems, Vol. 2, pp. 1219-1224.

[3] M. Ayara, J. Timmis, R. D. Lemos, L. N. D. Castro, and R. Duncan (2002) "Negative Selection: How to Generate Detectors." Proceedings of 1st International Conference on Artificial Immune Systems pp. 89–98.

[4] A. Imada (2005) "Can a Negative Selection Detect an Extremely few Non-self among Enormous Amount of Self Cells?" Proceedings of International Conference on Advanced Computer Systems (ACS) and Computer Information Systems and Industrial Management Applications (CISIM), to (appear).

[5] A. Imada (2005) "Can a negative selection detect unique non-self cell in an infinitely large number of self cells?" Proceedings of International Conference on Pattern Recognition and Information Processing, (in this volume).

[6] G. Castellano and A. M. Fanelli(2000) "Fuzzy Inference and Rule Extraction using a Neural Network." Neural Network World Journal Vol. 3, pp. 361–371.

[7] Z. Ji and D. Dasgupata (2004) "Augmented Negative Selection Algorithm with Variable-Coverage Detectors." Proceedings of the Congress on Evolutionary Computation Technical Report 94-07, University of New Mexico, Albuquerque, NM.

[8] H. S. Lopes, M. S. Coutinho, W. C. Lima (1997) "An Evolutionary Approach to Simulate Cognitive Feedback Learning in Medical Domain." Genetic Algorithms and Fuzzy Logic Systems. World Scientific, pp. 193–207.

[9] F. Provost, T. Fawcett, and R. Kohavi (1989) "The case against accuracy estimation for comparing induction algorithms." Proceedings of international conference on machine learning, pp. 445–453.