When a family of 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. This article is not a report of success but rather a challenge to those who claim to have designed a network intrusion detection system successfully by means of a machine learning technique using artificial dataset to train and to test the system.

1 Introduction

Those highly qualified hackers who provide security services to companies during the daytime and then go home at night to conduct totally illegal hacking are the ones who are the most dangerous. – by Enis Senerdem from Turkish Daily News on 29 March 2006.

Designing a network intrusion detection system is one of the hottest issues in our computer network society. Designing such a system with so-called a soft-computing now seems to became a sort of fashion these days. To try one of these approaches, we usually need a dataset to train the system and to test the system afterwards. Sometimes an artificial dataset is employed for the purpose. In fact we could find a lot of such databases in public domain. Spearman's iris database is one of those examples. We doubt, however, these dataset cannot necessarily reflect transactions in a real world. We usually don't know what does a coming intrusion looks like until it has completed the illegal intrusion when it is too late. In this paper, we give it a consideration to this issue.

2 Iris Database

Iris database[1] is made up of 150 samples consists of three species of iris flower, that is, *setosa, versicolor* and *virginica*. Each of these three families includes 50 samples. Each sample is a four-dimensional vector representing four attributes of the iris flower, that is, *sepal length, sepal width, petal length*, and *petal width*.

[1] University of California Urvine Machine Learning Repository. ics.uci.edu: pub/machine-learning-databases.

As is often mentioned, this iris flower database is perhaps one of the most often used datasets in pattern recognition/classification, machine learning, data mining, etc. As a matter of fact, there have been fair amount of studies in which this iris flower database is employed as a dataset to train and to test the system. Quite naturally, all of these papers report their success in designing a system.

Let us take an example. Castellano et al. [1] assumed one family of this iris flower to be normal whilst the other two to be abnormal. The whole dataset was divided into 10 parts each of which has 15 samples and are uniformly drawn from the three classes. The system is trained by the remaining 135 samples. The originally picked up 15 samples are used to test the results. After this 10-fold cross validation, the authors concluded that the system shows the invasion detection rate is 96% while the false alarm rate is 0.6%.

In reality, however, it is not so simple. Just imagine that a hacker always tries to explore an unlearned region to invade the network.

3 Challenges

First of all, let us formalized the standard way of using iris database for designing a network intrusion detection.

Problem 1 (One is normal while others abnormal) *Assuming one out of three families of iris flower to represent illegal transactions while the remaining two families represent legal ones. Is it possible then to design a system for network intrusion detection by using part of the dataset to train and remaining data to test the system?*

Let us now take a look at how those iris data are distributed in the whole search space. We tried a Sammon mapping to see those data in 2-dimensional space.

Sammon Mapping is a mapping a set of points a in high-dimensional space to the 2-dimensional space with the distance relation being preserved as much as possible, or equivalently, the distances in the n-dimensional space are approximated by distances in the 2-dimensional distance with a minimal error.

One of the results is shown in Figure 1. Just a brief look at the figure reveals us that there remains an enormously big region of unlearned.

Mutation

We now introduce a mutation which is applied to the data. in order to avoid for the data to be deterministic. Real data are never expected to be deterministic. Each of the iris database is given as the form of

$$(x_1, x_2, x_3, x_4)$$

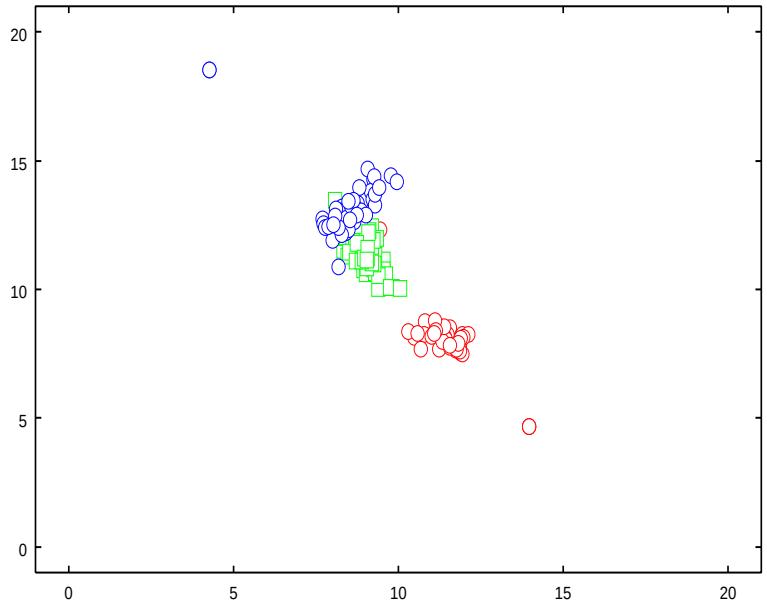

Fig. 1. A 2-D visualization of iris flower data by Sammon Mapping

With a small probability called mutation-rate we modify record by

$$x_i^{\text{new}} = x_i^{\text{old}} + r\sigma_i$$

where σ_i is standard deviation of 150 values of x_i, and r is a small random number anewed each time when mutation occurs. Then our problem turns into as the following.

Problem 2 (One normal while others abnormal but both mutated) *All points of one family of iris flower are given mutation and assumed to be normal while the remaining two families are also mutated and are assumed to be abnormal. Then is it possible to use these sample to train and to test the system?*

Still it would be not so difficult. Furthermore even if for a large mutation rate and/or large range of value r we still have large region of unleared samples. The more realistic and challenging one would be

Problem 3 (Is the system immune to mutants of previous invasion?) *A set of normal and abnormal samples are specified at random one by one. Then can the system trained by those samples recognize a mutant of abnormal sample?*

Let us now be more specific. We assume the whole universe is the 4-dimensional Euclidean space all of whose coordinate lie between 0 and 1. Then we create 50

normal samples and 100 abnormal samples all at random. We train the system using those 150 samples. Then the question is, "A test with 100 samples, 50 of which are mutants from normal samples and yet another set of 50 mutants from abnormal samples, will satisfy us?"

This is reminiscent of the experiment by Ayara et al. [2] who created 8-bit binary random strings, as a set of training samples of normal patterns assuming other comparable amount of 8-bit binary patterns as abnormal with their pattern-recognition results being very successful. Still this is not difficult, however, as the data used were fixed ones. How about, then, our case of mutants of abnormal sample? Does the system which is trained by an random abnormal sample still immune to those mutants? It would be really hard to believe.

4 Experiment

Each of the above 3 different test-sets is tried by C4.5 here. However, our goal will not be to show the success but be opposit. We will try to specifies the samples so that successful detection rate become as low as possible, and false alarm rate becomes as high as possible. (Results are not shown in this submission paper but sure to be appeared in the final version if this submission is accepted.) Not performed yet but we also plan to create abnormal samples by using a co-evolution of pledetor-and-prey type. (Not shown here either)

Also by using this iris flower database, we will explore an issue of learning where we doubt that training using previous abnormal data is really practical. The question is like, "If a wine taster who trained only with real champagne can recognize bootleg or other sparkling wine, or not."

5 Concluding Remarks

We feel sorry that we have described the above not so optimistically. However, we should be careful to draw a conclusion from our experiment or simulation. We sometimes tend to overestimate our results so that we like it. This article is not to deny the possibility but challenge for a new innovative approach to be emerged.

References

1. 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.
2. 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.

This article was processed using the LaTeX macro package with LLNCS style