Unsupervised Anomaly Detection for KDD-Cup99 Data
Using Self-Organizing Map and Convolutional Encoder
Hiroshi Dozono, Koki Yoshioka, Saga University JAPAN
Contact Email: hiro@dna.ec.saga-u.ac.jp

1. Introduction
- Many tailor made deep learning systems or libraries are applicable openly.
- Many of those are designed for 2 dimensional image classification which employs deep Convolutional Neural Networks (CNN).
- Some researchers uses raw data of their field of study with arranging them in 2 dimensional array regardless of the structure of data which are 1 dimensional, more than 3 dimensional or unstructured multi-modal data.
- However, CNN can extract the features of local regions, thus performance may be degrade if the data is not arranged properly in 2 dimensional array.
- If data is arranged properly in 2 dimensional array, performance may be improved.
- The experiments of unsupervised anomaly detection using convolutional encoder for KDD-CUP 99 data are conducted

2. Proposition
- We propose a method for arranging arbitrary data in 2 dimensional array using Self Organizing Map (SOM).
- Usually, SOM is applied to the unsupervised clustering and visualization of the relationship of input vectors.
- In this method, SOM is used to convert the unstructured data to 2 dimensional array.
- With using smaller map size compared with the original data, the input data can be compressed, and the time for computation is expected to be shortened.

3. Method
- As for conventional SOM, the input vectors are given for each data in row data shown as (a) in Fig.1, and each data is mapped according to the similarities among them.
- In this method, for converting the input data to 2 dimensional array, the input data are given to SOM for each attribute in column as shown as (b) in Fig.1.
- After learning SOM, the map represents the relationship among the attributes according to the set of input data.
- The layer composed of the element of n-th position in reference vectors represents the features of n-th input data arranged on 2 dimensional array of map size as shown in Fig. 2.

4. Experimental Result
- The experiments are conducted using KDD cup 99 data which is used as standard benchmark data for detection of malicious IP packets.
- Each data is composed of 42 attributes including the label, and 38 numerical attributes are used after normalizing to the range 0 to 1.
- The experiments are conducted with changing the size of maps.
- Preprocessed data in 4x4 images are shown Fig.3.
- The converted data of normal packets are given to convolutional encoder which is composed of 2 convolutional layers and 4 dense layers as training data.
- The converted data of normal packets and malicious packets which are different from training data are given to learned convolutional encoder, and the predicted data whose loss exceeds the threshold are determined as malicious packets.

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FN</th>
<th>FP</th>
<th>TN</th>
<th>ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0.999121</td>
<td>0.000879</td>
<td>0.116557</td>
<td>0.883443</td>
<td>0.905405</td>
</tr>
<tr>
<td>4x4</td>
<td>1</td>
<td>0</td>
<td>0.058836</td>
<td>0.941164</td>
<td>0.952619</td>
</tr>
<tr>
<td>8x5</td>
<td>1</td>
<td>0</td>
<td>0.016619</td>
<td>0.983381</td>
<td>0.986486</td>
</tr>
<tr>
<td>16x16</td>
<td>0.991957</td>
<td>0.008043</td>
<td>0.008215</td>
<td>0.991785</td>
<td>0.991817</td>
</tr>
<tr>
<td>32x32</td>
<td>0.998326</td>
<td>0.001674</td>
<td>0.01542</td>
<td>0.98458</td>
<td>0.987321</td>
</tr>
</tbody>
</table>

5. Conclusion
- SOM can convert arbitrary data to 2-dimensional data which can be applied to Deep neural network.
- The accuracy of detection becomes superior than those of original data with preprocessing using SOM.

Reference
Hiroshi Dozono, Application of Self Organizing Map to Preprocessing Input Vectors for Convolutional Neural Network; ICANN2019: Deep Learning, Lecture Notes in Computer Science, vol 11728, p.96-100