http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (2): 375-379, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.375.379

Enhanced Dynamic Self-organizing Maps For Data Cluster

Li Feng and Li-Quan Sun
School of Computer Applied Techniques,
Harbin University of Science and Technology, Harbin, China

Abstract: In the algorithm of Kohonen's Self-Organizing Maps (SOM) at the beginning of cluster, the number of input vectors in training set has to be settled down, which leads to the bad flexibility and is against the unsupervised principle. Also the fixed output network structure will lead to over-use or lack-of-use to the neuron node. To improve the exist defect of SOM and at the same time keep its advantages, an enhanced dynamic self organizing maps algorithm is proposed. This new method based on the idea of classical Growing Hierarchical Self-organizing Map (GHSOM), take advantage of GHSOM's feature of self determine the structure reflects the variability of data. By put forward a new cycle network structure EDSOM overcome the limitation of neutron under-utilize and over-utilize caused by the boundary effect. The experiment of intrusion detection proved the efficiency of the algorithm.

Key words: Growing hierarchical self-organizing map, enhanced dynamic self organizing maps under-utilize, neutron

INTRODUCTION

As an unsupervised machine learning method, clustering has become an important means for organization, summarization and navigation of text information. The advantage of text clustering is that, after coordinating document collection automatically, hidden structural information can be mined out and shown to people, thus text clustering technology makes it easier for users to browse many documents with less time. In addition, text clustering technology has been used in many fields such as multi-document summarization, post-processing of the returned results of search engine, information filtering and the information recommendation services, Digital library services and so on.

Among many text clustering methods, Self-Organizing Map (SOM) constitutes one of the most popular methods for unsupervised processing of high-dimensional and complex data based on principles of prototype-based vector quantization. Cluster with SOM has become an important means for organization, summarization and navigation of text information. Under certain conditions the learning scheme generates a model which allows a maybe nonlinear mapping of the given data set onto a low-dimensional regular lattice in a topology-preserving fashion (Kohonen, 1995), which it allows an easy analysis and interpretation of the data (Vesanto, 1999).

The power of the SOM stems from its robustness and the visualization possibilities for the results, both profiting from the key feature-the neighborhood cooperativeness incorporated in the adaptation scheme. During the last years, several extensions of the basic SOM have been established to make the scheme more flexible or to assess the quality of the generated model. These extensions are related to flexible grid structures, the processing of structured data and handling of labeled data by supervised learning. Thereby, the handling of appropriate, problem dependent data metrics becomes more and more important (Depren *et al*., 2005).

It is expected to generate the right cluster result automatically by using SOM model. Whereas the node number of Kohonen's SOM must be determined at the very beginning, so it has less flexibility and the clustering quality will often be adversely affected. When network structure is fixed beforehand, the nodes in output layer are prone to be overused or underutilized. People always expect to find the topic distribution of document collection automatically by using clustering model. Whereas the node number of Kohonen SOM must be determined before clustering, so it has less flexibility and the clustering quality will often be adversely affected. When network structure is fixed beforehand, the nodes in output layer are prone to be overused or underutilized.

This article presented an enhanced dynamic SOM clustering algorithm of incremental gradient descent for clustering large-scale data (EDSOM). The method is based on one of the classic dynamic self organizing maps method, GHSOM. The number of inserted neurons shows decreasing trend until only a neuron needs to be inserted

Corresponding Author: Li Feng, School of Computer Applied Techniques, Harbin University of Science and Technology, Harbin, China

at last. The incremental gradient decent is defined as the size of network in the algorithm can be gradually reduced and dynamically by inserting suitable number of neurons. To avoid neutron over-utilize and under-utilize, a closed circle structure is applied to replace the rectangle grid structure for GHSOM. The proposed method has successfully implemented the proposed algorithm and the experimental results are also presented to measure the efficiency.

SOM AND RELATED WORKS

The SOM was first proposed by Kohonen (1995) as a biologically inspired method to generate useful representations of data objects, now it is one of the most popular neural networks used in unsupervised learning. An SOM network consists of neurons organized in a lattice. The neurons are connected to adjacent neurons by a neighborhood relation, which dictates the topology of the map (Ouadfel and Batouche, 2007). The network implements a nonlinear projection from the high-dimensional input space to the low-dimensional lattice of neurons. SOM can serve as a clustering tool for high-dimensional data, which constructs a topology that the high-dimensional space is mapped onto the lattice of neurons in such a way that relative topology distances between input vectors are preserved (Poonguzhali and Ravindran, 2008). In an SOM, the neurons are arranged into the nodes of a lattice that is shown in the Fig. 1.

The lattice is usually one or two-dimensional. Higher dimensional maps are possible but not as common because it is limited by the visualization ability and also the applications of this kind of network.

As is known, it is before training that the number of the neurons should be identified, that is because Kohonen SOM generally uses a fixed output layer structure. It, however, destroys the unsupervised principles and self-adaptive of clustering. In addition, the problem of neurons underutilized or overused may be brought easily by the fixed structure. Thus the practical application of the algorithm has been greatly affected.

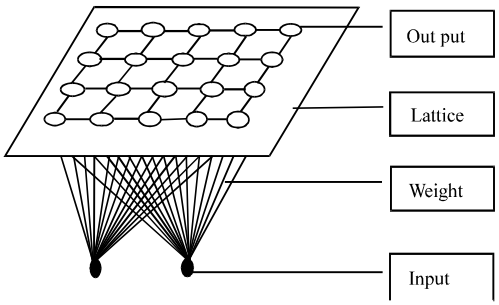

Fig. 1: Structure of Self organizing maps

Several dynamic SOM algorithms are proposed, Dittenbach (2000)'s Growing Hierarchical Self-organizing Map (GHSOM), extend the network by insert new row or column in output layer, using rectangular network structure, adaptively perform the dataset of input data. Because of the fast incensement of the network scale, the neurons in the module may be lack-of used. Growing Self-organizing Map (GSOM) applied a method to control the expansion of network scale, in which accumulative error of square structure is calculated (Alahakoon *et al*., 2000), so the expanding of network scale is get slower but by using the square structure, lack-of used neurons can not be totally avoid. Tree-structured Growing Self-Organizing Maps (TGSOM) applied more flexible tree-structure, which can generate new node based on the request efficiently (Li and Zhengou, 2003), the disadvantage is the network growing speed is controlled by experience parameter. To summarize, currently the clustering researches are limited to a small number of corpuses and a minority of the systems is fulfilled. The reasons include the larger memory, the greater dimension of vector texts (generally larger than ten thousand-dimension), the relatively slow processing speed, the relatively poor clustering results and so on (Ouadfel and Batouche, 2007).

ENHANCED DYNAMIC SOM

The process of insertion of units is shown in Fig. 2. The left side of the figure presents the structure of output layer before the insertion with unit e being the error unit and unit d being it's most dissimilar neighboring unit. The output nodes {N1, N2 …N6} are currently exist neutron elements. In this situation, a new column of units is inserted. The map after the insertion is shown on the right side of Fig. 2. Note that the new inserted nodes {N12, N11, N10} may not be fully used but anytime we need to extend the structure, we have to add more nodes to keep the structure of rectangle. With more operations of insert, the under-utilized getting bigger and bigger.

Process of EDSOM for units' insertion is shown in Fig. 3 In this contribution GHSOM is been extended,instead of using rectangle grid, a closed circle structure is applied for output nodes structure ({N1, N2 …N9}), every sector present one neuron. Here,

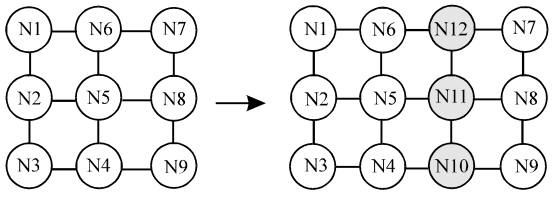

Fig. 2: Dynamic growth for new node added

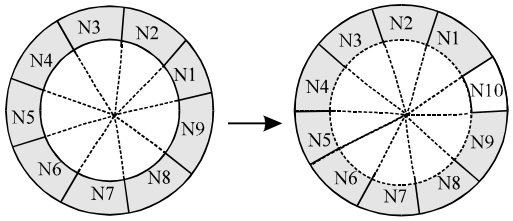

Fig. 3: Structure of enhanced dynamic SOM

we take advantage of number of sector can be easily customized, so that the information of category distribution for input dataset can be clearly reflected. By this way, when the output layer have to be dynamic expand, only one node N10 will be added based on the request, which can avoid the under-utilized.

Furthermore, each sector in the module has same number of neuron been adjacent to, so this is a symmetrical structure, which avoid edge effect problem. The underutilized neural problem can also be avoided by the flexibility of insert any number of neural node when the output layer is going to expand.

For the algorithm, first we initialized a relatively small scale network, then combine the decomposition and expand strategy together in the guidance of cluster criterion, to dynamic adjust the network structure. Decomposition strategy cites the idea of hierarchy clustering method, at the very beginning, set the output layer only contain 2 neutrons, at this time, one neutron may map multiple input node, which is a kind of over utilized. To balance the structure, new neutrons will be generated near the over utilized node, so the presentation of input layer is refined.

Apply R^2 cluster criterion coefficient as the basis of judgment, such criterion can efficiently control the network scale by evaluate the relationship between input and output dataset, find a proper balance among over-utilized and under-utilized and avoid infinite growth.

Let |Ci (t)| present the number of input nodes related with neutron C_i , m_i is the vector of C_i. The dispersion square sum inside the class formula can be like:

$$S = \sum_{d_j \to C_i} (d_j - m_i)^t (d_j - m_i) \quad (1)$$

Obviously, the smaller of S_i, the input data have bigger possibility to be same class. At time t, assume there are n neutrons, define, $\sum_{K=1}^{c} S_k$ we get:

$$T = \sum_{i=1}^{|D|} (d_i - x)^t ((d_i - x)) \quad (2)$$

T is the total dispersion square sum of input sample:

$$x = \frac{1}{|D|} \sum_{i=1}^{|D|} d_i$$

represents the mean vector of all the training samples, |D| is the number of input sample, so:

$$R^2 = 1 - \frac{P_c}{T} \quad (3)$$

The range for cluster criterion coefficient R^2 is [0, 1], the value is monotone increasing with the expanding of network scale. We set a threshold μ in order to terminate such expand, so that neutron under utilize can be avoid. If R^2 is less then threshold μ, then new neutron will be inserted near the biggest DSS neutron C_{max}, so that the representation of input data can be refined.

Let's generally describe the basic SOM algorithm:

1: Initialize a cycle structure output layer
2: Give an input set:

$$X(t) = \{X_1 (t), X_2 (t,...,X_N (t)\}$$

3: Calculate the cluster criterion coefficient R^2 of current output layer, if $R_2 > \mu$, then stop the training and quit, else perform the following steps:
4: Find the position of neutron with biggest square sum value inside the class, insert new neutron around it and then initialize the vector weight
5: Randomly pick the input sample and start training for all of the neutrons in output layer

RESULTS

The datasets used for the training and evaluating of EDSOM neural network were the KDD Cup 99 datasets (Http://kdd.ics.uci.edu/databases/kddcup99),which provide designers of intrusion detection systems with a benchmark to evaluate different methodologies. The datasets contain 24 types of training attacks, with an additional 14 types in the test data. All the attacks fall into four main categories:

- **Denial of service (DOS):** Attacker tries to prevent legitimate users from using a service. e.g., SYN-Flood attack
- **Remote to local (R2L):** Unauthorized access from a remote machine. e.g., guessing password
- **User to root (U2R):** Unauthorized access to local root privileges. e.g., buffer overflow attacks;
- **Probe:** Attacker tries to gain information about the target host. e.g., port scanning

Each record in the KDD Cup 99 datasets includes 41 features, some of which may be redundant or contribute

little to the detection process since the information they attach is contained in other features. In this study, 21 features are selected. The set of selected features contains numerical and symbolic features (per protocol type). However, a new metric including both numerical and symbolic data is introduced into the improved GHSOM algorithm and it is not necessary to convert the values of symbolic features to numerical values.

The only thing to do is to normalize the values of numerical features in the whole training and to test datasets into the interval (0, 1). The preparation of training and testing datasets is shown in Table 1.

To examine the performance of the improved GHSOM, This study use it to classify the normal and abnormal data and obtain detection rates on the testing dataset which is subset of Corrected KDD dataset. The parameter of the learning process of the improved GHSOM was: $\tau 1 = 0.0035$. The learning rate was defined as $\eta(t) = 0.1/(1+0.001t)$. All programs were coded in MATLAB 7.0 and run in a personal computer with Intel CPU 1.86(2) GHz, 2GB memory and Windows XP.

To compare the performance of EDSOM with GHSOM, this article implemented EDSOM and also applied the algorithm of GHSOM and compares the best result of the EDSOM with that of GHSOM for detection rates. Comparison results of attack specific detection as cluster key are shown in Table 2.

From Table 2, it is concluded that the improved GHSOM, the original SOM and GHSOM can detect attack types unavailable in the training dataset. For example, the°mail bomb" attack type is not available in the training dataset and the three algorithms classify it as a type of DOS attacks. Meanwhile, the number of training data will influence the detection rates. For instance, the number of training data for "buffer overflow" is too small to form a certain cluster, which results in a bad detection rate.

To make a longitudinal comparison, use category as cluster key to measure GHSOM and EDSOM, the result is shown in Table 3.

Table 1: Training and testing datasets

Category	Attack name	Training dataset	Testing dataset
Normal	Normal	16000	8000
DOS	Back	2142	845
	Neptune	7453	27
	Smurf	4559	2247
	Mailbomb	0	831
R2L	Guess_passwd	53	85
	Sendmail	0	17
U2R	Buffer overflow	30	22
	Xterm	0	13
Probe	Ipsweep	753	306
	Nmap	159	84
	Satan	851	704
	Mscan	0	143

Table 2: Comparison results of attack-specific detection rates

		Detection rate (%)	
Category	Attack name	GHSOM	EDSOM
Normal	Nrmal	94.3	95.2
DOS	Bck	93.5	94.5
	Neptune	96.3	95.9
	Smurf	95.7	96.3
	mailbomb	92.3	94.1
R2L	Guess_passwd	18.0	46.0
	Sendmail	26.4	37.0
U2R	Buffer overflow	46.5	40.9
	xterm	36.4	30.8
Probe	Lpsweep	92.1	94.0
	Nmap	93.6	94.8
	Satan	92.7	94.6
	Mscan	90.0	91.9

Table 3: Comparison results of category-specific detection rates

	Detection rate (%)	
Category	GHSOM	EDSOM
Normal	94.3	95.2
DOS	95.7	96.2
R2L	20.1	43.6
U2R	42.9	38.4
Probe	90.3	92.3

From Table 3, it can be seen that the EDSOM has shown better results for DOS, R2L, U2R, Probe attacks and normal compared to the original GHSOM. However, it remains difficult to identify R2L and U2R attacks.The number of training data will influence the detection rates. If the number of training data for specific attack is too small to form a certain cluster, it results in a bad detection rate. This is the most serious drawback for all the neural networks.

CONCLUSION

In this study, an enhanced dynamic SOM algorithm is presented, implemented and evaluated. The validity and feasibility of the enhanced dynamic SOM are confirmed through experiments on KDD Cup 99 datasets. The experimental result shows that the detection rate has been increased by employing the EDSOM compared to the GHSOM.

A new dynamic node growth structure has been introduced into the EDSOM after a investigation on the advantage and disadvantage of GHSOM algorithm. The comparison results between the EDSOM and GHSOM show that the EDSOM has much higher detection rate. Both improved algorithm have much better performance compared to the classical GHSOM respectively. Part of our future research will focus on the selection of features.

ACKNOWLEDGMENT

This study is supported by National Nature Science Foundation of China under Grant No. 60173055.

REFERENCES

Alahakoon, D., S.K. Halgamuge and B. Srinivasan, 2000. Dynamic self-organizing maps with controlled growth for knowledge discovery. IEEE Trans. Neural Networks, 11: 601-614.

Depren, O., M. Topallar, E. Anarim and M.K. Ciliz, 2005. An intelligent Intrusion Detection System (IDS) for anomaly and misuse detection in computer networks. Expert Syst. Appl., 29: 713-722.

Dittenbach, M., D. Merkl and A. Rauber, 2000. The growing hierarchical self-organizing map. Proceedings of the International Joint Conference on Neural Networks 2000, July 24-27, 2000, Como, Italy, pp: 15-19.

Kohonen, T., 1995. Self-Organizing Maps. 1st Edn., Springer, New York, USA.

Li, W. and W. Zhengou, 2003. TGSOM: A new dynamic self-organizing maps for data clustering. Electron. Inform. Technol., 25: 313-319.

Ouadfel, S. and M. Batouche, 2007. AntClust: An ant algorithm for swarm-based image clustering. Inform. Technol. J., 6: 196-201.

Poonguzhali, S. and G. Ravindran, 2008. Automatic classification of focal lesions in ultrasound liver images using combined texture features. Inform. Technol. J., 7: 205-209.

Vesanto, J., 1999. SOM-based data visualization methods. Intell. Data Anal., 3: 111-126.