Analysis and Application of Data Mining Based on Clustering Algorithm

LAI Honghui[1, a], LAI Xiao tao[1]

[1]Faculty of Information Engineering, Gannan Medical University, Jiangxi GanZhou，341000, China

[a]laihonghui1215@163.com

Keywords: Soccer Robot; Mechanical Analysis; Optimal Design

Abstract. As the ART2 neural network clustering occurs normalization in the data inputting mode by vector and nonlinear transformation pretreatment process is easy to be filtered as a substrate for an important, but a minor component of the noise, while there are still phenomenon of the drift mode in the learning process due to the correction of the value of weight, this paper proposes an improved method of ART2 neural network. On the one hand, as the inputting mode enters into the network to start learning, in the learning process of the input mode to enter the network, it saves their amplitude information, relaxes negative real nonlinear conversion, and considers the shortest distance of the input pattern to each cluster center.; on the other hand, there is also a need to make the corresponding treatment on the non-linear transfer function, so that it can properly handle the input of negative and retains its negative form after the stable F1 layer, not causing loss of information in the inputting mode; in another aspect, in order to eliminate outliers' influence on clustering results, this paper also carried out on the input mode to determine outliers.

Introduction

In order to carry out the clustering process to the text data, people have used a number of effective clustering methods, such as the classic k-means clustering algorithm, which wad based on text clustering algorithm of SOM neural network. However, these methods often require a lot of previous knowledge to determine the number of clusters; it is not a way of dynamic learning, and learning the new vector will affect the learnt vectors and other issues. According to the advantages of ART2 neural network, it can efficiently go on the dynamic learning, and it not only realize the balance of memory and learning, but also determine the number of clusters adaptively. But ART2 network remains worthy improvements, such as the entering sensitivity to data will greatly affect the clustering results of ART2 network.

Improved Art2 Neural Network Method

The traditional ART2 neural network's clustering is based on the phase information, irrelevant to amplitude information. The effect of the traditional ART2 neural network is far from ideal when dealing with the same phase information and amplitude information of two different clusters. Some also are mentioned in the article by comparing the weights of the model and the inputting samples to recover the amplitude information, but this does not reflect the weight of the prototype model's amplitude information, so it is still unable to use the amplitude information.

For the data samples that the original data are both positive and negative, due to the limitations of inputting fields of the traditional ART2 network. In the F1-layer of the traditional ART2 network, non-positive real number of sample data is suppressed to 0, so the traditional ART2 network can not effectively classify data samples locating two, three, four quadrants.

Meanwhile, the traditional ART2 neural network is not sensitive to the presence of outliers. In order to try to eliminate outliers' influence on clustering results, this improved algorithm, by using the outlier as an additional class, reduce the effects of outlier on the clustering results.

As for such existing shortcomings in the traditional ART2 neural network, in the inputting activation process, the study calculates the shortest distance to each cluster from the center, considering its amplitude information. Only when both the values of phase and amplitude exceed

the corresponding threshold limit alerting value, the resonance occurs and the adjustment of the weights begins. So does non-linear transfer function. So much so that it can properly handle negatives' input and retain its negative form after the stability of F1-layer, avoiding loss of information in the inputting mode; in order to eliminate the impact of outliers on clustering results, the paper also carried out on the sure of outliers in inputting mode and increased a threshold limit value $R-dis$ to detect outliers. Improved ART2 neural network matches the phase and amplitude as shown in Figure 1.

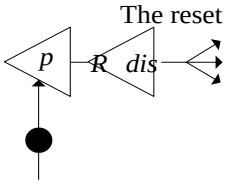

Figure1. The matching of phase and amplitude of improved ART2 neural network

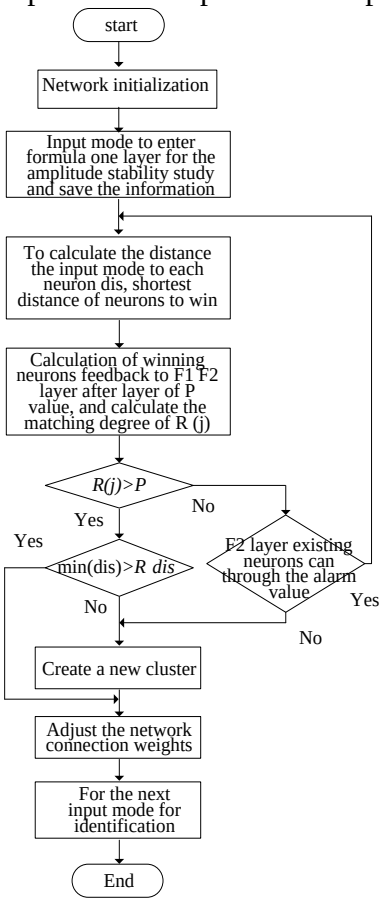

Figure2. Flowchart of Improved ART2 algorithm

Improved ART2 network also includes attention subsystem and orientation subsystem. Attention subsystem includes STM, two short storage units of F1 and F2, and long storage unit LTM connecting F1 and F2 layers, that is, connected weight vectors $w_n \times m$ and $T_n \times m$. In this time, the connection weight vector w_{ij} from the bottom to up is a record of the amplitude information of the cluster's center; the j denotes $j's$ center point. The role of the orientation subsystem is to calculate the phase degree of matching between the inputting mode and the memory mode, that is, the matching extent between intermediate mode U that the F1-layer is stable and the winning neuron's feedback mode P from the top to down is compared, as well as detecting inputting mode's outliers. It is for determining the next action of network: resonate or reset.

First, initialize the network's settings. In the improved ART2 network, the initialization of F1-layer from the top to down and the initialization of weight vector $T_n \times m$ is the same with the traditional ART2 network. The number of clusters m is set to 1, the connection weights vector

$w_n \times m$ from the bottom to up initializes the first inputting pattern as the first cluster center, i,e .

$$w_n \times m = \left(t_1^1, t_2^1, ..., t_n^1\right)$$

It also need set two threshold limit values ρ and R_ dis. ρ is the alerting value of the phase matching, $R-dis$ as outliers' alerting values for determination.

When the n-dimensional inputting mode t= (t_1, t_2 ,..., t_n) entering into F1-layer, the formula (1) to (3) F1 layer can calculate the steady states of F1-layer. As non-positive real number in the traditional network ART2 is unified t_y processed as 0, nonlinear transferring function takes it as a noise to deal with, making the network lost that part of the information and affecting the whole results of the clustering. Therefore, there is a need to adjust non-linear processing function to correctly handle non-positive real numbers, to prevent the misuse of the useful information. The non-linear processing function is adjusted to

$$u(x) = \begin{cases} \dfrac{2\vartheta t^2}{t^2 + \vartheta^2}, 0 \le t \le \vartheta \\ t, |t| \ge \vartheta \\ -\dfrac{2\vartheta t^2}{t^2 + \vartheta^2}, -\vartheta \le t \le 0 \end{cases} \tag{1}$$

Or

$$u(x) = \begin{cases} 0, 0 \le |t| \le \vartheta \\ t, |t| \ge \vartheta \end{cases} \tag{2}$$

When F1-layer reaches a steady state, the inputting mode I connects with F2-layer through a bottom-up weight vector $w_n \times m$ and conducts competitive learning for finding neurons with the shortest distance as the winning neuron, i,e .

$$w_j = \min\{w_i\}, i = 1, 2, ..., n, w_j = \sum_{i=1}^{m} \left\| i_1 - w_{ij} \right\| \tag{3}$$

Winning neuron is activated, while the other neurons are in the state of inhibition. F2-layer selects the winner neuron j and returns a feedback signal, calculating the degree of phase matching ‖R‖ between the processed STM signal U of F1-layer and the feedback value LTM P of active neuron signal. Since ‖R‖ reflects the overall matching degree between U and P, regardless of the difference between P and U, which are the various components. Equation (1) is used for phase-matching calculation in the paper. If ‖R‖ is greater than the setting threshold limit value ρ, then the inputting mode is determined by the outliers; if k_j is larger than a presetting threshold limit value $R-dis$, then the inputting mode is processed as the outlier point, and the inputting mode is considered as a separated class; the inputting mode is classified into the class j. When the network enters into the learning phase, bottom-up weight vector w_{ij} is updated to a new class $j's$ center point. The effect is to be that it is the average value of data samples' weights, weight vector t_{ij} from the top to down is to be updated according to formula (3).

Experiments

In this paper, the above algorithm for the horizontal and vertical coordinates in [0, 1] is generated randomly within five characteristics distinct groups, each containing 30 data samples cluster, using traditional ART2 and improved ART2 network respectively to cluster the data sample. The clustering results are shown in Figures 3 and 4, wherein each parameter is set as shown in Table 1, where the parameters a, b, c, d, e values of both parameters, these parameters may also be determined by the experience of experiments.

Table1. Setting table of network parameters

	a	b	c	d	e	t	$R-dis$
Traditional ART2	10	10	0.12	0.93	0	0.98	×
Improved ART2	10	10	0.12	0.93	0	0.94	0.21

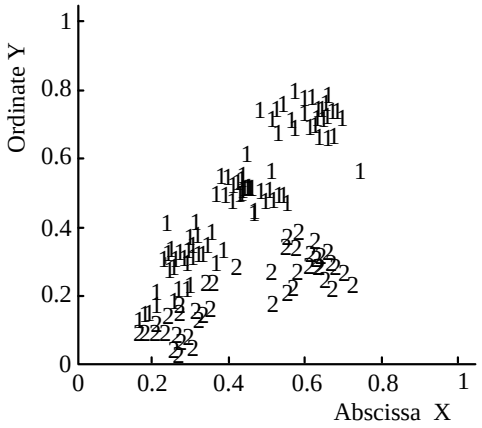

Figure3. Clustering results of traditional ART2 network

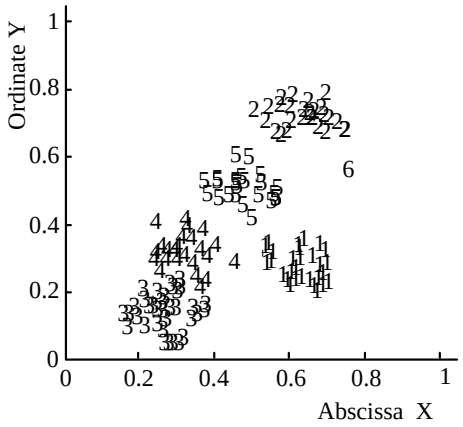

Figure4. Clustering results of improved ART2 network

Conclusion

Through the theoretical and experimental results, it shows that when the improved ART2 neural network is in the same phase of the two clusters, the performance is better than the traditional ART2. Meanwhile, the network takes the phase information of data and amplitude information of a prototype data into account and eliminates outliers of the clustering results. By changing the nonlinear transforming function, an improved ART2 network can handle negative data, and the four quadrants of the data can be efficiently clustered. The experiments show that the improved ART2 network is to be significantly better than the traditional ART2 network in dealing with outliers amplitude information and data samples' performance.

References

[1] Xin Huang, Xiao Ma, Bangdao Chen, Andrew Markham, Qinghua Wang, Andrew William Roscoe.Human Interactive Secure ID Management in Body Sensor Networks. Journal of Networks, Vol 7, No 9 (2012), 1400-1406

[2] Jian Wu, Jie Xia, Jian-ming Chen, Zhi-ming Cui, "Moving Object Classification Method Based on SOM and K-means. Journal of Computers", vol.6, no.8 , pp.1654-1661, 2011.

[3] J. He, Y. Geng and K. Pahlavan, Modeling Indoor TOA Ranging Error for Body Mounted Sensors, 2012 IEEE 23nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Sydney, Australia Sep. 2012 (page 682-686)

[4] Y. Geng, J. Chen, K. Pahlavan, Motion detection using RF signals for the first responder in emergency operations: A PHASER project[C], 2013 IEEE 24nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), London,Britain Sep. 2013

[5] S. Li, Y. Geng, J. He, K. Pahlavan,Analysis of Three-dimensional Maximum Likelihood Algorithm for Capsule Endoscopy Localization, 2012 5th International Conference on Biomedical Engineering and Informatics (BMEI), Chongqing, China Oct. 2012 (page 721-725)