Data Clustering Using ART-like Neural Networks

Agáta Bodnárová

Department of Information Technologies
Faculty of Informatics and Management
University of Hradec Králové
Rokitanského 62, 500 03 Hradec Králové, Czech Republic
agata.bodnarova@gmail.com

Tyler Frank

Faculty of Economics
Technical University of Košice
Letná 9, 040 01 Košice, Slovak Republic
katkafrank@yahoo.com

Abstract: This paper is focused on the data clustering using the ART-like neural network. Firstly is concerned with the problems from the theoretical point of view by explaining what the clustering is, giving an idea of accumulation methods and mentioning the place of the neural networks in the data clustering. From among the neural network it is focused on the MF ARTMAP. Firstly it explains the principle of the ART neural networks namely unSupervised ART, Supervised ARTMAP which is created by uniting two ART neural networks using MAPFIELD together with giving the basis of the fuzzy set of the ARTMAP. The aim of the thesis was not only to explain the theory of the MF ARTMAP networks but also to implement them and suggest their improvement. The experiments were evaluated on the data sets circle in square, spiral and economical data. Improvements are related to data which are categorized into two classes. Improvements roots is addition of third class, which contains classified contradictory. Examples, and their subsequently separation into two classes which are concretized whether there are data which belong to the both classes or are contradictory. In case of the multiclasses classification, the functionality of the MF ARTMAP on the economical data is verified by neural network. The results are processed, visualized as well as explained in the individual sections.

Keywords: ART, ARTMAP, data clustering, neural networks

1 Project Definition and Task Determination

This paper analyzes data clustering using ART-like neural networks.

Main tasks are the following:

- Present overview to data clustering methods
- Theoretically describe ART-like neural networks
- Theoretically analyze and describe MF ARTMAP
- Implement a MF ARTMAP neural network
- Realize experiments on data sets circle in square and spiral, visualize it and propose enhancement of these experiments
- Realize experiments on economical data sets
- Make conclusions

2 The State of the Art in the Domain

ART (Adaptive Resonance Theory) neural networks for fast, stable learning and prediction have been applied in a variety of areas. Areas of their technological application includes industrial design and manufacturing, control of mobile robots, face recognition, remote sensing, land cover classification, target recognition, medical diagnosis, electrocardiogram analysis, signature verification, tool failure monitoring, chemical analysis, circuit design, protein/DNA analysis, 3-D visual object recognition, musical analysis, as well as seismic, sonar and radar recognition. ART systems are used in VLSI microchips.

Supervised ART architectures, which are called ARTMAP systems, feature internal control mechanisms that create stable recognition categories of optimal size by maximizing code compression while minimizing predictive error in an on-line setting. Special-purpose requirements of various application domains have led to a number of ARTMAP variants, including fuzzy ARTMAP, ART-EMAP, Gaussian ARTMAP, and distributed ARTMAP. ARTMAP has been used for a variety of applications, including computer-assisted medical diagnosis. Medical databases presents many of the challenges found in general information management settings where speed, efficiency, ease of use, and accuracy are at a premium.

3 Selected Methods and Approaches

3.1 Data Clustering

Clustering is the process of unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult problem combinatorially, and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur.

Clustering can be defined as the process of separating a set of objects into several subsets on the basis of their similarity. The aim is generally to define clusters that minimize intracluster variability while maximizing intercluster distances, i.e. finding clusters, which members are similar to each other, but distant to members of other clusters. Two clustering strategies are possible: hierarchical or non-hierarchical. In this master thesis we talk about non-hierarchical clustering.

3.2 Computational Intelligence in Data Clustering

Computational Intelligence represents a part of Artificial Intelligence and mainly integrates three different technologies concerning artificial neural networks, fuzzy systems and evolutionary systems. Integration of these systems results in so called hybrid intelligent systems. The most known systems of computational intelligence are neural networks.

Neural Network (NN) is an information processing paradigm that is inspired by biological nervous systems, such as the brain. The key element of this paradigm is the novel structure of the information processing system. It is composed from a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. Neural Network analogous to people, learns by example. NN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. One of the most important feature of neural network is universal approximation of function. Neural network is also used in classification into classes, classification of various situations, solving predict and process control problems, signal transformation, association problems and simulation of memory.

3.3 ART-like Neural Networks

A central feature of all ART systems is a pattern matching process that compares an external input with the internal memory of an active code. ART matching leads either to a *resonant* state, which persists long enough to permit learning, or to a parallel memory search. If the search ends at an established code, the memory representation may either remain the same or incorporate new information from matched portions of the current input. In case of search ends at a new code, the memory representation learns the current input. This match-based learning process is the foundation of ART code stability. Match-based learning allows memories to change only when input from the external world is close enough to internal expectations, or when something completely new occurs. This feature makes ART systems well suited to problems that require online learning of large and evolving databases.

Supervised ART architectures, called ARTMAP systems, self-organize arbitrary mappings from input vectors, representing features such as spectral values and terrain variables, to output vectors, representing predictions such as vegetation classes in a remote sensing application. Internal ARTMAP control mechanisms create stable recognition categories of optimal size by maximizing code compression while minimizing predictive error in an on-line setting.

3.4 MF ARTMAP

MF-ARTMAP calculates the membership function of the point from the feature space to the 'fuzzy class'. Representation of knowledge appears from presumption when data at input space are organized at fuzzy clusters. It is possible to bind random dot x from input space value of membership function to fuzzy cluster μA(x). Given value of membership function from dot x to fuzzy set, which fuzzy cluster is representing. Because each cluster introduces relation between inputs and is defined by fuzzy set, it is possible to describe it through the use of fuzzy relation, where exact parameters of fuzzy relation will be adapted in learning process, so they are carriers knowledge. Some requirements are desired on selection of parametric function that will represent fuzzy relation.

In regard to these criterions fuzzy relation is defined as:

$$A(\bar{x}) = \int_Y \frac{1}{1+\left|\left(\frac{\left|\bar{x}_S-\bar{x}\right|}{\bar{E}}\right)^{\bar{F}}\right|} / \bar{x} \tag{1}$$

where Y is input space, vector X present random dot from input space and vectors Xs, E and F are parameters of fuzzy relation. Example of this function for 2-D input space is displayed on the picture. Vector Xs represents the centre of fuzzy relation for each dimension of space.

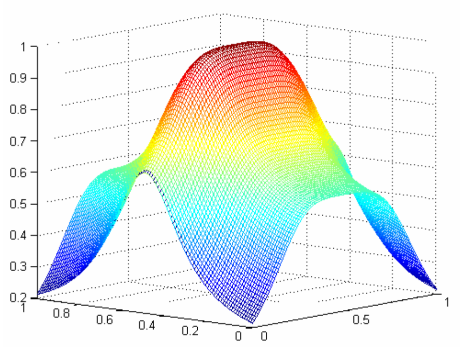

Figure 1
Fuzzy membership function

4 Design and Implementation

Supervised learning based on MF ARTMAP neural network was realized on two data files where each of them included its own train and test set with different number of samples.

During learning process data from train set was used as input and by testing resultant efficiency of classification was checked on test set. Improvements are related to data summarized into two classes. They are based on addition of the third class, where the third one contains classified contradictory examples, and its subsequent separation into two classes which are concretized whether there are data which belong to the both classes or are contradictory.

Functionality of the MF ARTMAP neural network in case of multiclasses classification is realized on economical data. Results indicates the usability of MF ARTMAP as gradually learning system.

Contingent table presents detailed analysis of classification process and indicates both the number of correctly and incorrectly classified vectors for each class.

5 Experiments

5.1 Circle in the Square

Benchmark data contains train set of 1000 elements and test set with 10000 elements. It is dichotomic classification of two-dimensional real input vector. Input vectors represents coordinates of circle in the square.

The best results of experiments are:

Table 1
Parameters of NN MF ARTMAP

Parameter	VALUE	ParameteR	VALUE
E	0,05	Number of classes	217
F	1	Good results	9747
Threshold	0,5	Incorrectly classified objects	253

Table 2
Contingent table

Real class/ Classified class	0	1
0	4855	111
1	142	4892

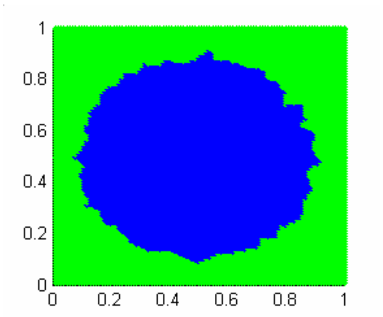

Figure 2
Circle in the square trained in MF ARTMAP

After the addition of the third class, in which are classified contradictory examples, and its subsequently separation into two classes which are concretized whether there are data which belong to the both classes or are contradictory are results the next:

Table 3
Parameters of NN MF ARTMAP after addition of third and fourth class

PARAMETER	VALUE
Number of clusters in first two classes	217
Number of clusters in the third class	71
Number of objects in the third class	72
Number of clusters in the fourth class	57

Number of objects in the fourth class	61
Good results	9648
Incorrectly classified objects	219

Graphical result is represented on the following figure. Blue color means circle, green means square, white color means the third class, red the fourth class and yellow means incorrectly classified objects.

Table 4
Contingent table after the addition of the third and the fourth class

Real class/ ClassifYed class	0	1
0	4795	95
1	124	4853
2	31	41
3	47	14

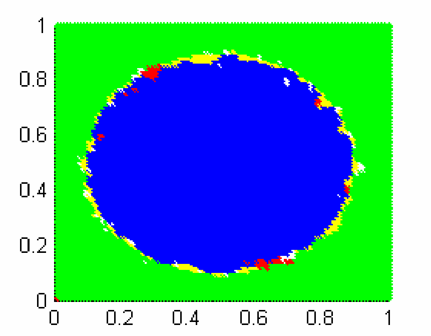

Figure 3
Circle in the square trained in MF ARTMAP after the addition of the third and the fourth class

5.2 Economical Data

Economical data contains answer sheets from 109 individual company. These companies the are divided into 6 classes scaled from 0 to 5. Zero rating indicates a company is in Bankruptcy, severe financial difficulty or has gone out of business. A rating of 1 indicates that the company and the industry are below average in performance. A rating of 2 indicates that a company is of average performance but in an industry that is declining in performance. A rating of 3 indicates that a company in below average in an industry that is performing well. A rating of 4 indicates average performance in a good performing industry. 5 indicates superior performance in a growing industry. These 109 companies represents the train set. The test set was generated from MF ARTMAP after training and presenting centre of each generated fuzzy set. 90% of objects was successfully classified.

6 Contribution to the Research Domain

There are two contributions to the research domain:

1 Improve NN MF ARTMAP with adding of the third class, in which are classified contradictory examples, and its subsequently separation into two clasees which are concretized whether there are data which belong to the both classes or are contradictory. Importance of this improvement is in using class 2 approach (means either 1 OR 2) beside class 1 and class 2. We do expect in this case better results. This idea is drawn on the figure below.

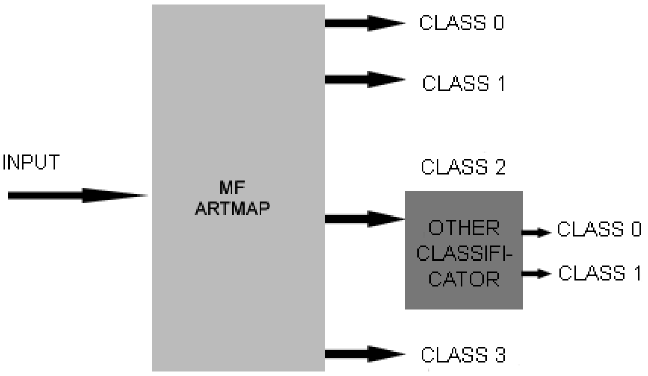

Figure 4
Two classificators

2 **Representation of NN MF ARTMAP as gradually learning system. In case of economical data it means that with raising number of answer sheets this network branch out and gives better results of classification**. During the following few years, the evolution of each company will be predicable based on these answer sheets. Also, conditions for being prosperous will be revolved for these companies.

Table 5
Results of experiments in economical data

PARAMETER	1	2	3	4	5
E	2,3	3	2	3,3	3,6
F	2	2	3	2	2
Threshold	0,8	0,8	0,8	0,8	0,8
Number of clusters	103	102	97	90	60
Good results in train data	105/109 96,3%	104/109 95,5%	103/109 95%	100/109 92%	78/109 72%
Good results in generated test.data	94/103 91%	94/102 92%	90/97 93%	82/90 91%	54/60 90%

Conclusions

The main object of this thesis was to check the possibilities of NN MF ARTMAP in two classes classification and to design its upgrading. Additional object was to introduce NN MF ARTMAP as gradually learning system. Gradually learning systems are perspective way of neural network development in the future.

References

[1] BODNÁROVÁ, A.: Metódy zhlukovania dát s využitím neurónových sietí typu ART, Diplomová práca, TUKE, FEI, KKUI, 2006

[2] CARPENTER, G. A., MILENOVA, B. L., NOESKE, B. W.: Distributed ARTMAP: a Neural Network for Fast Distributed Supervised Learning. Neural Networks, Vol. 11, No. 5, Jul. 1998, pp. 793-813

[3] CARPENTER, G. A., MILENOVA, B. L., NOESKE, B. W. (1998). Distributed ARTMAP: a Neural Network for Fast Distributed Supervised Learning. Neural Networks, 11, 1998. 793-813

[4] FRANK, T.: Infrastructures for Sharing Information Between Firms, Dissertation work, May 2004

[5] HAKL, F., HOLEŇA, M: Úvod do teorie neuronových sítí. Ediční středisko ČVUT Praha, 1997

[6] HRIC, M.: Diplomová práca – Integrácia neurónových sietí typu ARTMAP s prvkami fuzzy systémov pre klasifikačné úlohy. TU FEI KKUI, 2000

[7] HRIC, M.: Minimová práca. TU FEI KKUI

[8] HRISTEV R. M.: The ANN Book. GNU Public Licence, 1998

[9] KVASNIČKA, V.. et al. Uvod do teorie neuronovych sieti. Iris: Bratislava, 1997

[10] OCELÍKOVÁ, E.: Multikriteriálne rozhodovanie, ELFA, Košice, 2002

[11] SINČÁK, P., ANDREJKOVÁ G.: Neurónové siete, Inžiniersky prístup (1. diel). Elfa: Kosice, 1996

[12] SINČÁK, P., ANDREJKOVÁ G.: Neurónové siete, Inžiniersky prístup (2. diel). Elfa: Kosice, 1996

[13] SINČÁK, P., HRIC, M.: Neural Networks for Fuzzy Clustering, IJCNN 2001, International Joint Conference on Neural Networks, Washington, July 2001, pp. 291-29