CLUSTERING ALGORITHM IN MULTIDIMENSIONAL DATA SETS USING PART NEURAL NETWORK

Roman Krakovský [1] and Igor Mokriš [2]

[1] Catholic University in Ružomberok, Faculty of Pedagogy
Námestie A..Hlinku 56/1, 034 01 Ružomberok, Slovak Republic
e-mail: roman.krakovsky@ku.sk
2) Institute of Informatics, Slovak Academy of Sciences
Dúbravská cesta 9, 845 07 Bratislava, Slovak Republic
e-mail: igor.mokris@savba.sk

Abstract: The article is concerned with the problematic of Projective ART neural network (PART NN) and their use in the area of non-controlled learning for creation of cluster. The article states description of neural network PART, principle of Projective Adaptive Resonance Theory in the process of learning of neural network and describes its individual phases. In the next part the article focuses on use of PART neural network for processing of multidimensional data stored in text documents, system real-time databases and biomedicine.

Keywords: PART neural network, clustering algorithm, multidimensional data space.

1 INTRODUCTION

For the purpose to increase the accuracy information retrieval the research of the semantic web has been developed. Keywords are usual means to formalization of documents for information retrieval on Internet. Nowadays the research of information retrieval utilizes neural networks where it is possible to generate clusters from input samples. Traditional clustering methods don't work efficiently for data sets because of the inherent sparsity of data. This motivated the concept of subspaces clustering whose advantage is to find cluster formed in subspaces of the original high dimensional space.

Clique was first known algorithm proposed for automatically discovering clusters in isolated subspaces of a multidimensional space. In 1999 was designed faster algorithm for clustering in different subspaces PROCLUS (PROjective CLUStering) (Aggarval,1999). PROCLUS solves the problem finding projective cluster. Each of which consists of a subset C of data points together with subset of dimension D, such the points of C are closely correlated in subspace of dimensions D. This algorithm exploits medoids-based optimization approach and combines the greedy search algorithm with locality analysis technique to find set of dimensions associated with each medoid. Medoid substitutes object in cluster to serve as a surrogate center in cluster. The clustering quality is evaluated as a middle-value of Manhattan segment distance from each points to the center of corresponding cluster. This distance is defined relative to the dimensions, where cluster is generated. PROCLUS works as follows:

- initialization phase – small medoid set M is generated using greedy serach algorithm, such that points inside this subset are sufficiently separated from each other,
- iterative phase – set of medoids is arranged by hill climbing algorithm, so the bad medoids are replaced another medoid from the set M. Bad medoids have shorter distance of other medoids than predefined threshold,

- final phase – pick up clusters with the smallest average distance along dimension and set D_i asociated with medoid m_i. Next the clusters are formed by grouping every data poinsts to its closest medoid according Manhattan distance relative to set of dimensions associated with medoid.

Experiments with syntetic data sets showed that PROCLUS is sensitive to the choice of input parameters but PROCLUS performed better than CLIQUE in term of quality and running time.

2 PART NN DESRIPTION AND PART CLUSTERING ALGORITHM

In order to deal with the feasibility-reliability dilemma in clustering data sets of high dimension, Cao and Wu presented an approach based on new neural network architecture – PART (Projective Adaptive Resonance Theory). The basic architecture PART NN is similar to the ART neural network, it proves very effective in a self-organizing clustering in full dimensional spaces (Mařík, 2003). PART NN tackles this problem to selective output signalling mechanism to increase the accuracy of clustering in multidimensional spaces (Cao, 2004). Fig. 1 illustrates the basic PART NN architecture while the PART algorithm is described below.

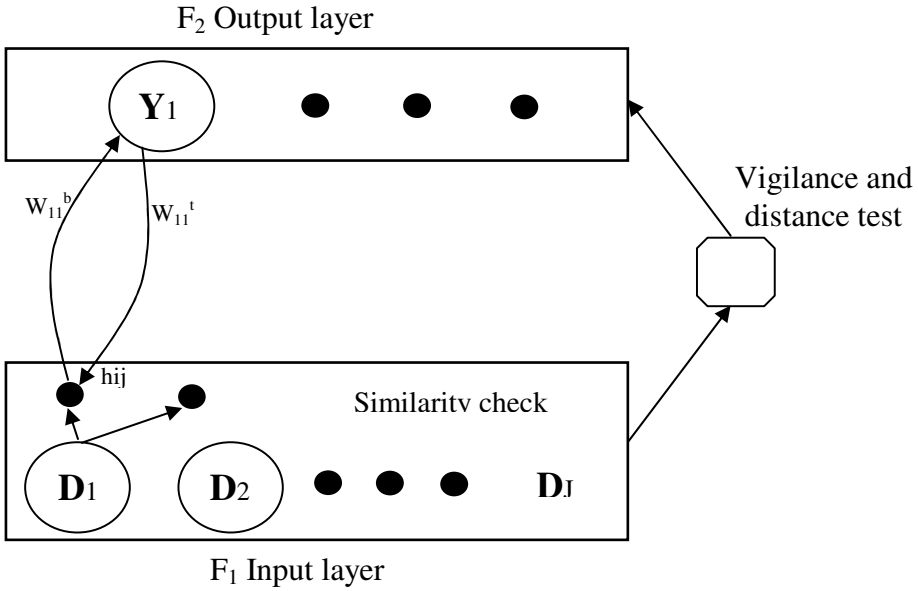

Figure 1. PART NN architecture

Definition parameters appearing in the PART algorithm:

m	pattern amount	
n	term amount	
W_{jk}	bottom-up weight	
W_{kj}	top-down weight	
σ	distance parameter	
L	constant parameter	$L>=1$
θw	threshold of weight	$0<\theta w<=L/(L-1+n)$
θc	threshold of cluster	$0<\theta c<=m$
ρ	vigilance parameter	$0<=\rho<=n$
α	learning rate	$0<=\alpha<=1$

PART algorithm to cluster form is presented in details as follows:

- Initialization – set the internal parameters *L, ρ, α ,δ, θw, θc* and on input layer are introduced samples *D1,D2,...Dj,...Dm.* Output notes are Y_k, for *k=1,2,...m.*
- For each data point in input data sets is executed similarity check.

$$h_{\mathrm{jk}} = h(D_j, W_{jk}, W_{kj}) = h_\sigma(D_j, W_{kj}) l(W_{jk}) \quad (2.1)$$

$$h_\sigma(a,b) = 1 \quad \text{if} \quad d(a,b) \le \sigma \quad (2.2)$$

$$h_\sigma(a,b) = 0 \quad \text{if} \quad d(a,b) > \sigma \quad (2.3)$$

$$l(W_{jk}) = 1 \qquad \text{if} \quad W_{jk} > \theta_w \quad (2.4)$$

$$l(W_{jk}) = 0 \qquad \text{if} \quad W_{jk} \le \theta_w \quad (2.5)$$

  In case that $h_{jk}=1$, document D_j is similar to output node Y_k. Otherwise, *Dj* is not sufficiently similar to Y_k.

- Is select the winning node to satisfy next equation

$$Tk = \Sigma WjkHjk = \Sigma WjkH(Dj, Wjk, Wkj) \quad (2.6)$$

  maximum $\{ T_k \}$ is winner node.

- In follow step is evaluated vigilance test and reset mechanism for winner node.

$$R_{\mathrm{k}} = \sum h_{jk} < \rho \quad (2.7)$$

  The input pattern will be c;ustered into the winner node, if winner node passes the viligance test. Otherwise is used reset mechanism and input pattern will be clustered into a new node.

- After finishing previous tests and checking started learning process in a net. If Y_k has not learned a pattern before, are updated the bottom-up and top-down weights for winner node as following:

$$W_{jk}^{new} = L/(L - 1 + n) \quad (2.8)$$

$$W_{kj}^{new} = D_j \quad (2.9)$$

  Otherwise :

$$W_{jk}^{new} = L/(L - 1 + n) \quad \text{if} \quad h_{jk}=1 \quad (2.10)$$

$$W_{jk}^{new} = 0 \qquad \text{if} \quad h_{jk}=0 \quad (2.11)$$

$$W_{kj}^{new} = (1 - \alpha) W_{kj}^{old} + \alpha D_j \quad (2.12)$$

- Are repeated steps before, until amount of nodes in each cluster fall under threshold θc.
- Finally are returned the clusters.

The degree of similarity of patterns is controlled by both vigilance parameter and distance parameter which control the size of dimensions of the projected subspaces and the degree of similarity in a specific dimension involved. In difference of PROCLUS these viligance and distant parameters are the only required input parameters for PART algorithm. The PART algorithm with a wide range of input parameters enables to find the correct centers of clusters, the correct number of cluster and sufficiently large subset of dimmensions where clusters are formed. So that is possible to fully reproduce the original input clusters, after reassigns procedure, which bind set every data point to its closest cluster center according to the distance on the subspace of the found dimensions.

PART algorithm is based on the assumptions that the model equations of PART (a large scale and singularly perturbed system differencial equations coupled with reset mechanism) have quite regular computational performance described by the dynamical behaviors during each learning trial. In the experiments with 1523 web pages consist of 900 pages from Google and 623 pages from ESPN as domain data was shown, that PART NN gets better result than ART NN when the quentity of data is large and PART NN is better in web page clustering.

3 PART NN IN APPLICATIONS

Was proposed a principle based on advantages of processing the huge text documents placed in multidimensional space through neural networks. The base is built on Projective Adaptive Resonance Theory published on Chen and Chuang (Chen, 2008). In this study was proposed and testified through experiments automatic generated domain ontology based. PART NN clustered the collected web pages and then found representative keywords for each cluster of the web pages using entropy value. The system next used a Bayesian network to insert the terms and complete the hierarchy of the ontology. Finally the system used a resource description framework to store and express the ontology result.

The major contribution Projective ART with buffers (Liu, 2009) is introducing a buffer management mechanism that allows data sets not to be immediately clustered into one cluster and partly achieve an independent purpose order without very correct parameters. The purpose of the average similar degree is to successfully works with high similar noise data sets and partly achieve an order independent objective without correct parameters. In the experiments was putted some noise data sets between the input files. Those noise data sets have a characteristic that they share partial dimensions of one cluster, but they are completely different at the other dimensions of the cluster. Was designed the experiment file with 10000 data sets and 5 clusters in 20-dimensional space and all 5 clusters have the 7 projected dimensions in different subspaces. Next was reported result of both PART and Buffered PART (BPART) in each file. So the clustering result doesn't depend on the precise choise of input parameters. In the study was experimented BPART in the database of the Hang Seng Composite Index Series (HSGI) from 3 October 2001. There are 481 transacion days and each transaction day contains all 100 constituent stocks.

Neural networks enforce in various domains of medicine and in gene analysis also. However in most gene expression datasets, the number of samples is less than that of gene. At University in Nagoya was developed and put into use new filtering method BagPART, based on Projective adaptive resonance theory (Kawamura, 2008). In addition it was found showed that BagPART is more effective than traditional filtering methods when sample size is small. In order to correctly select genes, BagPART method applies an idea modifying PART NN introducing idea of Bagging. In addition it was found showed that BagPART is more effective when sample size is small. Last advances in DNA microarray technology have made measures that express thousands of levels genes together. Artificial neural networks and Fuzzy ART neural networks

combined with SWEEP operator (FNN-SWEEP method) are helpful for building cancer class prediction model with high accuracy (Takahashi, 2005). However, in gene expression data are easy to include experimental error. Therefore, it's necessary to find significant genes and eliminate non-significant genes to prevent the model from over fitting for learning data before modelling. It was applied PART NN for eliminating nonspecific genes, furthermore was built model for cancer class prediction.

4 CONCLUSION

The article is devoted to clustering problem in multidimensional data sets through Projective ART neural network, based on S. Grossberg published theory of ART NN (Grossberg 2002) and PROCLUS clustering algorithm (Aggarwal, 1999). Examples from various areas suggest that with daily increase of information, as well as kinds of methods of storing and data processing with huge data sets on multidimensional spaces, PART NN and modifications of PART NN find increasing application.

REFERENCES

AGGARWAL C.C., PROCOIUS, C., WOLF, J.L., YU, P.S., PARK J.S. (1999): Fast Algorithm for Projected Clustering, *SIGMOD '99,* pp.61-72.

CAO, Y., WU, J. (2004): Dynamics of Projective Adaptive Resonance Theory Model: the Foundation of PART Algorithm, *Neural networks, Volume 15, March 2004*, pp.245-260.

GROSSBERG, S., CARPENTER, G.A. (2002): Adaptive Resonance Theory. *The Handbook of Brain Theory and Neural Networks, MIT Press.*

KAWAMURA, T., TAKAHASHI, H., HONDA, H. (2008) : Proposal of New Gene Filtering Method, BagPART for Gene Expression Analysis with Small Sample. *Journal of bioscience and bioengineering, Vol.105, No.1,* pp. 81-84.

LIU, L., HUANG, L.(2009): Projective ART with Buffers for the High Dimensional Space Clustering and an Application to Discover Stock Associations. *Neurocomputing , Elsevier*, pp.1283-1295.

CHEN, R.CH., CH.H. CHUANG, CH. H. (2008): Automating Construction of a Domain Ontology Using a Projective Adaptive Resonance Theory Neural Network and Bayesian Network, *Expert systems, Vol. 25, No. 4,* pp. 414-430.

MAŘÍK, V., ŠTEPÁNKOVÁ, O., LAŽANSKÝ, J. (2003): Umělá inteligence (4), *Academia Praha, ISBN 80-200-1044-1.*

TAKAHASHI, H., HONDA, H. (2005): A New Reliable Cancer Diagnosis Method Using Boosted Fuzzy Classifier with SWEEP Operator, *Journal Chemical Engineering, Japan, No. 38*, pp.763-777.

Acknowledgement

This work was supported by Slovak Science Agency VEGA No. 1/0692/08, VEGA No. 2/0211/09 and VEGA No. 2/0184/10.