Processing of Information in Multidimensional Data Space by Projective ART Neural Network

Roman Krakovský*
Institute of Informatics
Slovak Academy of Science in Bratislava
Dúbravská cesta 9, 845 07 Bratislava, Slovakia
roman.krakovsky@stuba.sk

Abstract

The theme of paper is focused on the processing of information sources on the bases of data clustering using neural networks. In the recent years there were presented various research findings on processing of information sources based on the semantics, focusing on semantic Web and ontology, as well as models of neural networks based on the principle of adaptive resonance theory (ART). Published work in the area of classification and clustering of multidimensional data of various kinds confirmed the legitimacy of use and of further research potential of the ART neural networks models. Based on the review of current state of addressed problematics the submitted work is focused on analysis and proposal of possible improvements to the process of clustering and classification of multidimensional text and image data using the Projective ART neural network.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering; I.2.6 [**Learning**]: Connectionism and Neural Nets

Keywords

PART neural network, clustering, vigilance parameter, committed neuron, multidimensional data

1. Introduction

Due to the refinement of information retrieval based on the semantic research is in this area focused on the semantic web and ontology. The process of classification and clustering of information can be enriched in different ways, different types of neural networks are just one of them. In recent years, published models of neural networks based on the principle of projective adaptive resonance theory used for classification and clustering of multidimensional data confirmed the validity of different kinds of use, and further research. Therefore, this paper (in general) focuses on the analysis and proposes improvements to the process of clustering and classification of multidimensional data sets using PART neural network (hereinafter PART). We can conclude that for successful implementation of modified PART the idea is to support classification and clustering algorithms and their application in real data processing.

2. Overview of the Current Situation

2.1 Proccessing of Multidimensional Data Sources

A necessary condition for the proper processing of information is the appropriate representation of the data. Classification is the process in which recognized objects are incorporated into classes of the mutual affinity, the set of classes is given in advance. Clustering solves the problems of unsupervised learning, in which the objects are classified into groups, called clusters. The task of clustering is to create clusters, similar objects from the collection of inputs without prior knowledge about objects belonging to some classes [12] [13] [17] A particular problem is the clustering of multidimensional data, which tries to find clusters in different subspaces of the same dataset. Clustering of data is a process that classifies data into groups of similar objects. Object is conceived as a individual element of the dataset. Each cluster contains in its inside, objects that are very similar to each other, in contrast to other objects from other clusters, which are very different to each other. Degree of similarity of objects is based on a selected attribute values of individual objects. The aim of clustering is to obtain a possibility to assess individual objects as parts of one cluster and thus the opportunity to work with clusters instead of working with a number of objects. In the process of clustering we don't known in advance the number of generated objects clusters [1] [2] [13] [16]. Clustering of data in multidimensional space is an extension of the original clustering, which seeks to find clusters in different subspaces of the same dataset. Objects are mostly represented by vectors that characterize their basic attributes, or by the degree of similarity. The transformation of multidimensional data space into subspaces is based on the decomposition of multidimensional coordinate space to the coordinate system of planes of the respective subspaces.

*Recommended by thesis supervisor: Prof. Igor Mokriš Defended at Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava on August 24, 2015.

Krakovský, R. Processing of Information in Multidimensional Data Space by Projective ART Neural Network. Information Sciences and Technologies Bulletin of the ACM Slovakia, Vol. 7, No. 2 (2015) 1-9

2.2 Neural Networks in the Process of Classification and Clustering of Data

Neural networks belong to the universal means suitable for solving problems of classification and clustering of objects or processes prediction [5] [14]. They are used everywhere where inaccurate data is given (various forms of the word, occurence diverse keywords, changed shapes and sizes of images ...) [11] that make it difficult to search for information from objects and where it is needed to solve the problem of classification, search or prediction of processed, stored and retrieved data collections. Obviously, the structure of the classifier in large number of data is compiled and, from this is where other disadvantages come from, such as the large classification time difficulty, the capacity requirements of a classifier, etc. Therefore, the most frequently used is classification of objects by attributes.

In processing of information resources using neural networks research is mainly devoted to the application of feedforward neural networks [15]. For example, to solve the context in document processing, it is preferable to use predictive approaches that are accepted mainly in the structure of recurrent neural networks. The information retrieval models using neural networks were applied mainly in autoassociative Hopfield neural network [11], ART neural networks [9] hierarchical neural networks, as well as Self-organizing neural networks.

2.3 ART Neural Network

Adaptive Resonance Theory (ART) [8] [12] was used to analyze the problem of how the brain links can learn independently in real time in a changing world of rapid but stable manner. Key processes of ART networks are the selection and comparison. The selection process chooses the best category for input sample. If the template of selected category is sufficiently similar to the input sample to suit predefined parameters vigilance ϱ, then category resonates and learns. Weights of template are changed to match the new sample of data. Otherwise, the winning cluster is reset and the most likely category (cluster) is selected. If, after the transition of selected clusters does not match comparison criteria, a new cluster is created. For this reason, ART is able to incrementally create new clusters needed to represent all the clusters in the input samples.

Simplified structure of the ART network including input (processing or comparative) layer F_1 and output (recognition) layer F_2. Important parts of the network shall include vigilance test and reset element. In the network of this kind there are two groups of connections (each with its own weight) between each neuron of the input and output layer. The input layer is connected to the output by the means of bottom-up weights w_{ij}, and vice versa, F_2 layer is associated with F_1 through the reverse weights z_{ji}. Weights connection between the layers can be modified by two different learning rules. F_2 layer also called competition weight, which is subject to paradigm "winner takes all".

The main advantage of ART neural network is that it does not expect the number of clusters in advance and allows the user to manage the degree of similarity between samples located in the same cluster. A positive feature of the network is the low number of repetitions of the basic

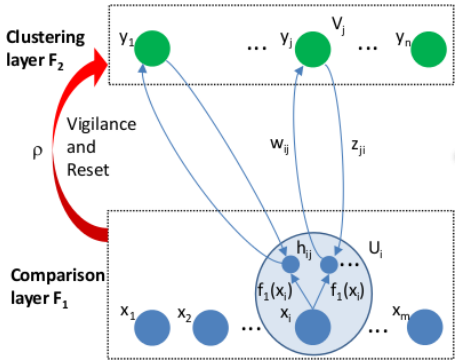

Figure 1: Model PART neural network.

network cycle compared to other recurrent neural networks. The disadvantage of ART neural networks is their low noise resistance and memory requirements for the implementation of the classification process while a strong noisy patterns can cause gradual resetting of bottom-up and top-down weights.

ART neural network deployment in practice to address the problem of aggregation and published experiments have shown that the architecture of the existing ART neural network must be modified in order to successfully address the role of subspace clustering in a multidimensional dataset.

2.4 Projective ART Neural Network

To solve the problem of searching data clusters in multidimensional space, a new data structure of recurent neural network without a teacher was established, which is based on the principle of adaptive resonance theory labeled PART (Adaptive Resonance Theory Projective Neural Network) [4] [3] [10] [5]. In principle, PART is based on a neural network system Proclus and ART neural network, which is very effective for aggregating data in a separate data space. The essential characteristic of PART network is the ability to switch between plastic and stable mode, without breach of learned information. Therefore the process of storing information in the PART network can be divided into two parts: the short term memory (STM Short Term Memory), which can be easily renewed, without leaving the stable state and long-term memory (LTM Long Time Memory), linked to the learning network. The second important feature of this network is projective clustering, whose main task is to find projective clusters, consisting of such a set of neurons which are sufficiently closely linked with the relevant subset of dimensions. The basic structure of the standard two-layer PART is shown in FIG. 1.

The neurons in the input F_1 (comparative) layer are labeled V_i, for $i = 1, ..., m$, the neurons of F_2 (aggregating) output layer V_j, neuron activation V_i from layer F_1 by x_i and activation layer V_j from layer F_2 by x_j. Neurons between individual layers of neurons V_i to neurons V_i are attached to each other through the weights z_{ij} and top-down weights z_{ji}.

Processing according to PART consists of initialization, comparation, recognition, search and adaptation phase in which the following formulas are applied:

$$T_j = \sum_{i \subset F_1} z_{ij} \cdot h_{ij} \tag{1}$$

$$h_{ij} = h(x_i, z_{ij}, z_{ji}) = h_\sigma(f(x_i), z_{ji}), l(z_{ij}) \tag{2}$$

$$h_s(f(x_i), z_{ji}) = \begin{cases} 1 & \text{if } d(f(x_i), z_{ji}) \leq \sigma \\ 0 & \text{else} \end{cases} \tag{3}$$

$$l(z_{ij}) = \begin{cases} 1 & \text{if } z_{ij} > \theta_c \\ 0 & \text{else} \end{cases} \tag{4}$$

$$r_j = \sum_i h_{ij} \tag{5}$$

$$z_{ij}^{new} = \begin{cases} L/(L-1+|X|) & \text{for } v_i \text{ commited} \\ 0 & \text{else} \end{cases} \tag{6}$$

$$z_{ji}^{new} = (1-\alpha) z_{ji} + \alpha I_i \tag{7}$$

$$z_{ij}^{new} = L/(L-1+m) \tag{8}$$

$$z_{ij}^{new} = I_i \tag{9}$$

Since its first publication, PART has undergone several changes, depending on the set of input data and applications in which it was used. Its modifications have found application in the area of processing of information sources on the Internet, text documents, data processing of real-time systems (PART with buffer), image data processing (PARTCAT) and also in the area of biomedicine (Bag-PART) [16].

3. The Thesis Objectives

The thesis objectives focuse on the analysis and proposal of possibilities for improving the process of clustering and classification of multidimensional data sets using unsupervised neural networks (without teacher). The motivation to write the work was the author's effort, based on the analysis of the original clustering by PART, to design own modification of PART. Modified PART would be designed to address the clustering of text documents. In addition to the implementation of the modified model design also algorithm for optimization of parameters PART, and not least to demonstrate the merits of using the proposed model PART also in processing of multidimensional non-text data. A prerequisite for the successful implementation of idea of modified PART is to support clustering algorithms and their applications in the processing of real text and image data.

The objectives of this article can be summarized into the following sections:

1. Analysis, design and implementation of the modified algorithm MA-PART for clustering of multidimensional text and image data.
2. Design and implementation of the algorithm to optimize the parameters modified PART.
3. Design and implementation of modified model PART clustering texts, including design and implementation of additional algorithms.
4. Design and implementation of hybrid neural network model for image classification by clustering, including the design and implementation of additional algorithms.

4. Modified Projective ART Neural Network on Clustering of Multidimensional Data Sets

The originating PART on the clustering of real data in the published applications have indicated that the plurality of the clustering needs to be supplemented by refinement algorithm to obtain a sufficiently accurate results. The idea of the proposed modified algorithm clustering MA-PART lies in the amendment of the original clustering PART algorithm. The structure of the neural network PART will remain the same.

The proposed modified PART by author does not change the original structure of PART. Change is in the adjustment of the calculation of selective output signal h_{ij} and adjustment of algorithm clustering.

Signaling function $f : R \to R$ must be non-decreasing function satisfying the Lipschitz condition:

$$|f(x) - f(y)| \leq M|x - y|;\ x, y \in R \tag{10}$$

Undertaken experiments with real collections of text documents confirmed that the calculation of the distance between the input vector and the value of output weights reached best results using the following equation:

$$d(x_i, z_{ji}) = \frac{|x_i - z_{ji}|}{1 + (z_{ji})} \tag{11}$$

In addition to altering the distance function are in a modified model PART also included other changes:

- In the case of aggregation of text documents can be the case that the two output neurons meet at the same time the condition for the winner (rear weights They have the highest value). In this case the algorithm applied to the a special section in which is gradually converted scalar product of vectors predicted winners for processing of vectors, the largest of the values determined by the winning neuron.
- In any projective cluster centroids are dynamically converted and maintained.

4.1 Optimization of Modified Projective ART Neural Network

Generating optimal parameters for aggregating multidimensional data by PART depends on the choice of used model PART but also on the nature of multidimensional data, which in this case is a set of training data. Applications and published clustering algorithms by PART [4] [5] did not contain a description of setting optimal parameters of network. Therefore, based on the author's own experience and realized experiments PART has created own algorithm for optimization of network parameters.

The idea of designing of optimization algorithm parameters PART essentially narrowed the search for optimal parameters ϱ and σ, because they have the greatest impact

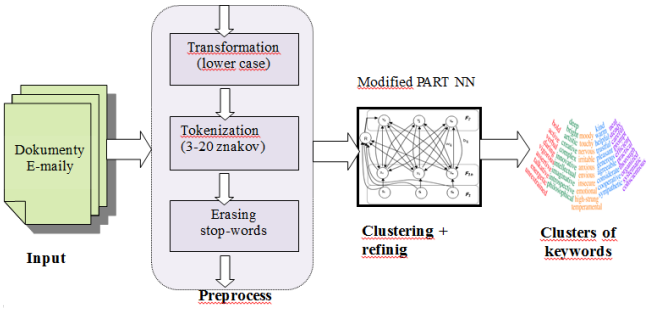

Figure 2: The process of generating glossary of keywords from text documents.

on clustering hierarchy of cluster generation and classification accuracy of various vectors into clusters. Optimization of these parameters is based on finding of such a range for each parameter, in which the desired number of clusters is formed and the accuracy of the classification of the individual clusters of samples is as high as possible.

The algorithm of search for optimal parameters PART has iterative in nature. In loops are gradually incremented parameters ϱ and σ (from *startsigma* after *stopsigma*) with iteration step *deltasigma*. Testing combinations of pairs of parameters ϱ and σ throughout the interval $\varrho = (1..m)$ and $\sigma = (0..1)$ is a computationally intensive task that can result in aggregation stored in matrix $Param$. Based on the evaluation of the data matrix param i.e. the maximum number of pure clusters (free from defective classified vectors), maximum number of lines in created clusters and the minimum number of vectors in outlier layer can be relatively accurately selected intervals of input parameters ϱ and σ in which PART clusters multidimensional data most precisely.

Testing combinations of pairs of parameters ϱ and σ throughout the interval $\varrho = (1..m)$ and $\sigma = (0..1)$ is a computationally intensive task that can result clustering stores in a matrix $Param$. Based on the evaluation of the data matrix $Param$ is maximum the number of pure clusters (error free classified vectors), the maximum number lines in the clusters and the minimum number of vectors in outlier cluster is can quite accurately algorithmically calculate intervals of input parameters ϱ and σ, in wich PART clusters multidimensional data accurately.

4.1.1 Modified Projective ART Neural Network for Generating Dictionary Keywords from Document

Suitability idea of modified PART in processing text documents was verified in an application for generating dictionary keywords, where except synthetic data were used shared and published text documents. Processed documents used for training and testing of the modified PART came from the repository 20 Newgroups. Processing of text documents in which for clustering was used MA-PART is shown in Fig. 2.

Clustering accuracy of words in example 4 documents in computer section by using MA-PART and PART is shown in Fig. 3.

Inaccuracies for generating resulting clusters in realized experiments lead the author to edit clustering algorithm

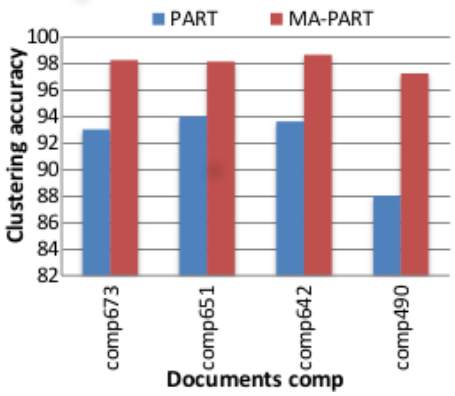

Figure 3: Clustering accuracy in example 4 documents by MA-PART and PART.

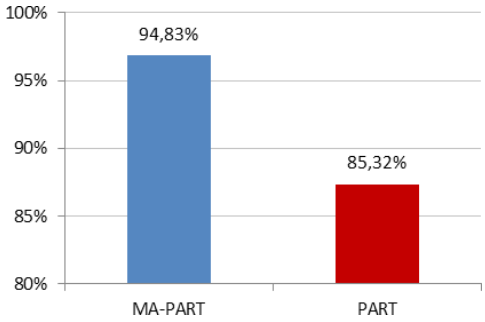

Figure 4: Clustering accuracy of words by MA-PART and PART.

to calculate the condition of selective output signal h_{ij}. It was supplemented by the condition of equality of the first three projective dimensions of input vector and one of the vectors being compared cluster:

$$h_{ij} = \begin{cases} 1 & \text{if} \quad (d(f(x_i), z_{ji}) \leq \sigma) \wedge (w_{ij} > \theta) \wedge \\ & \quad \wedge (x_1 = c_1) \wedge (x_2 = c_2) \wedge (x_3 = c_3) \\ 0 & \text{else} \end{cases} \tag{12}$$

Generated matrix of clusters of keywords achieved an average success rate of aggregation around 91%. To increase the accuracy of data clustering was created advanced algorithm based on the application of Porter's algorithm on the first keyword in the cluster and consistent comparison of remaining keywords with him.

In the experiments, was used in the training set of documents and the 30 test set of 340 documents. Choice of training and test sets was focused on more comprehensive, multi-kilobyte files free text, while the number of words The documents was in the range 250-4000.

The result of processing are generated clusters of keywords from which it is further possible to make the keyword dictionary. In the experiments, the whole testing set accuracy achieved clustering by MA-PART at 94.83%, an improvement compared to the result the original algorithm PART 9.51%. By applying the refinement algorithm the dictionary of keywords for individual clusters generated by MA-PART using the Porter algorithm was

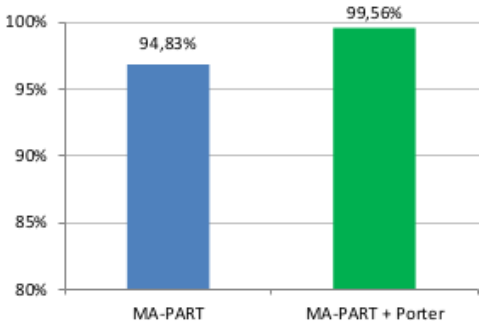

Figure 5: Clustering accuracy by MA-PART +Porter and MA-PART.

achieved recognition accuracy up to 99.56%. Success rate of classification of keywords into clusters is readable from graphs in the following figure.

4.1.2 Modified Projective ART Neural Network for Clustering Text Documents

Input of the whole system are text documents from various thematic areas obtained from the Internet. Processing results are clusters of documents that were created by the proposed model modified PART, aggregated to anticipated groups, i.e. thematic areas. For assigning of generated clusters to real thematic areas regrouping algorithm was used.

The processing of text documents is divided into separate parts:

- pre-processing of text documents using RapidMiner,
- clustering by MA-PART - creating of projective clusters with including relevant centroid for each created cluster,
- assigment of created clusters into real clusters using regrouping algorithm.

Based on the analysis results of processing with MA-PART, the set of documents with a considerably higher number of keywords failed to achieve success rate of classification into clusters above 90%. Modified PART model can not adequately respond to the increase in the number of dimensions of submitted vectors.

Pre-processing of text documents using RapidMiner consists of following parts:

- tokenize - divided contents of one document into sentences and sentences into tokens,
- transform case - document transform letters to lowercase,
- filter tokens - filtering of tokens, the length of which is within range minimum and maximum length,
- filter stop-words - delete the words belonging to the database of English stop-words,
- stem Porter - crop extension tokens using the Porter algorithm.

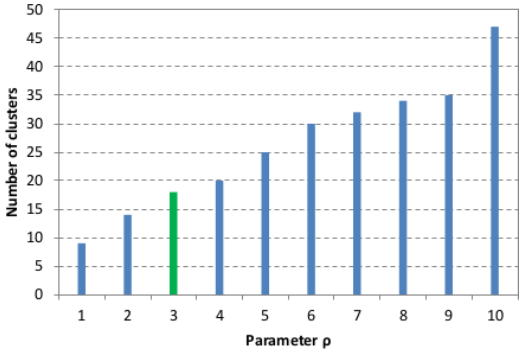

Figure 6: Impact parameter ϱ of the number of clusters created for $\sigma = 0.4$.

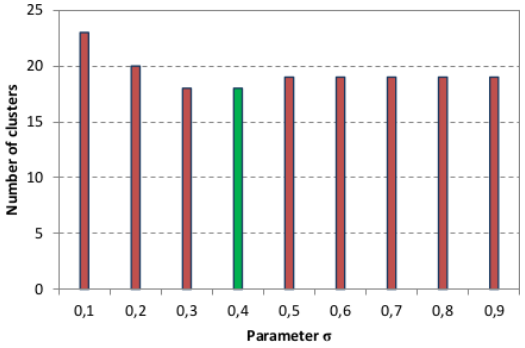

Figure 7: Impact parameter σ of the number of clusters created for ϱ=3.

The first step to improve the success rate of clustering was to increase stabilization time of network weights, by increasing the number of repetitions of *epoch*. Constant in the denominator function of distance was able to slightly increase the ratio of the distance attributes of the input vector x_i and top-down weights z_{ji} to attributes of top-down weights and thereby increase the overall accuracy of vector classification into existing clusters.

For the calculation of the centroid was used the least squares method:

$$dist(x_i, c_j) = \sum_{i=1} (x_{ik} - c_{jk})^2 \qquad (13)$$

On Fig. 6 and Fig. 7 is shown impact of parameter ϱ respectively σ for number of created clusters in pursuing maximum accuracy clustering.

The last phase in the processing of text documents is regrouping algorithm that maps created projective clusters in real clusters. For the purposes of clustering of text documents obtained from the web in regrouping algorithm were used and compared with each other Euclidean metric, cosine metric and Jaccard coefficient. Undertaken experiments have shown that the main cycle of clustering was necessary to repeat at least four times in order to achieve the most accurate results. The proposed experiments were carried out with a collection of text documents in English, which were obtained from the repository 20 Newsgroups. To train the network was used set of 30 documents from 4 areas. Test set with the number 96 text documents consisted of 4 selected thematic areas.

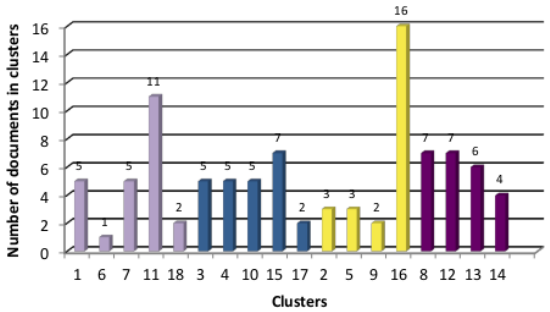

Figure 8: The results of applying the regrouping algorithm by using cosine metrics

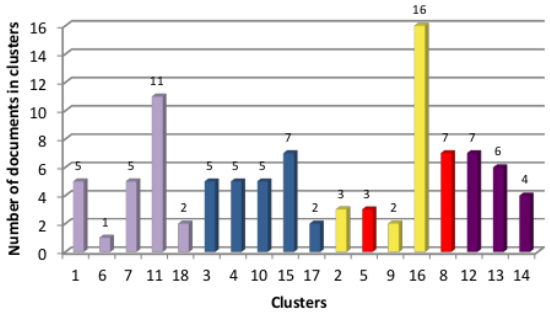

Figure 9: The results of applying the regrouping algorithm by using Jaccard metrics

The total number of keywords for the testing set was the 212.

In experiments there were compared also dependencies of incorrectly classified text documents by changing the distance parameter σ, with the correct setting of vigilance parameter $\varrho = 3$. Number of created clusters in changing parameter σ is around 18, but in fact the total number of incorrectly classified documents was for each of the values σ significantly different. In addition, was simultaneously monitored the impact of changing the number of repetitions clustering algorithm epoch on the total number of incorrectly classified documents. On the total number of 96 selected text documents from four thematic areas was for modified model PART reached clustering success rate 93.75%. The author proposed model used for clustering text documents in English combines the advantages of a modified PART and uses metrics to achieve the exact number of real clusters of documents. By editing function for calculation of distance of input vector and reverse weights of commited neurons was achieved improvement in success rate of clustering.

4.2 Modified Projective ART Neural Network for Clustering Image Data

Based on the analysis of previous modifications of PART is this part of the dissertation devoted to demonstration of PART usabalility in non-text data clustering of multidimensional data sets, namely multidimensional image data. For this purpose was by the author created and tested a hybrid model of PART in conjunction with optimized modified by model of coupled neural network (furher OM-PCNN).

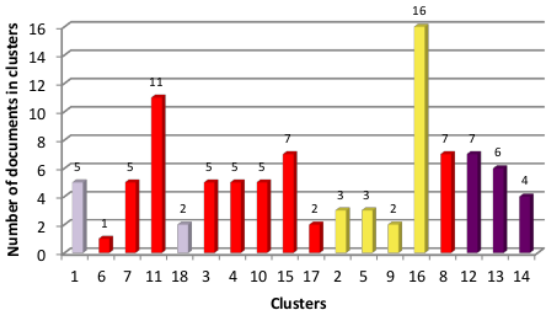

Figure 10: The results of applying the regrouping algorithm by using euclidean metrics

4.2.1 Hybrid Model of PART Neural Network for Clustering and Classification of Image Textures

The hybrid model has been designed to address the clustering of multidimensional image data using the PART together with decreasing dimension of the processed data by OM-PCNN. The main selection criterion used in neural networks for the purposes of research was the low number of network parameters and clearly defined structure of used neural networks. The proposed hybrid model was used in experiments for clustering images textures as well as for classification and filtering out of unnecessary data to outlier cluster.

PCNN advantages lie in its invariance to geometric transformations (translation, rotation and partly dilatation), the minimum set of used standards and firmly defined structure that maps the image matrix pattern recognition. PCNN does not requires a learning process, which is typical for standard neural network. Thanks to invariance of PCNN rigidly defined structure and uselessness of learning are eliminated problems of standard neural networks, which are most often solved experimentally. A precise mathematical model of OM-PCNN, algorithm for generating the symptoms can be found in [6].

The experiments with a modified PART were divided into two types of tasks:

- clustering the reduced matrix of image data into related clusters,
- filtering and classification of unnecessary data based on selected standards.

For clustering the reduced image data matrix was used a modified algorithm PART using a simplified function of distance:

$$|x_i - z_{ji}| < \sigma \tag{14}$$

The original database of sample images was processed to raster 450 x 450 pixels. The proposed hybrid model PART and PCNN also addresses the issue of classification of invariant image data [7]. Each class of original images was extended by a set of bracketed images, dilated images at interval of $\langle 30\%, 95\% \rangle$, rotated images at intervals of $\langle 15°, 345° \rangle$. For training and setting network parameters using the author proposed algorithm was used set of 260 images. The test database contained a total of 1612 images.

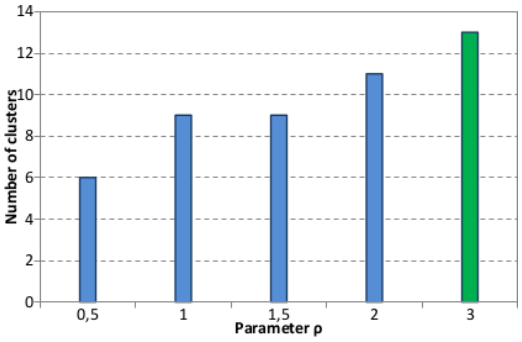

Figure 11: Impact parameter ϱ of the number of the clusters

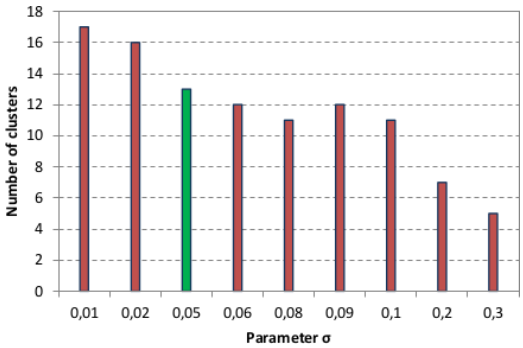

Figure 12: Impact parameter σ of the number of the clusters

The optimal setting of OM-PCNN network was for the input parameters α_0=0.6 and β_0=0.1. The process of clusetring and classification by PART was optimal by setting parameter ϱ=3 and σ=0.05. The clustering was assessed as successful if whole set of test images were classified into 13 groups, while each group having 124 texture image. The result of the clustering image textures with a hybrid model is summarized in Table 1.

The experiments were conducted repeatedly over various combinations of groups of image textures, but before every attempt was clearly intended that which groups of images were significant and which were insignificant. Experiments were conducted repeatedly in various combinations

Table 1: Precision of Classification by OM-PCNN+ED and OM-PCNN+PART

OM-PCNN+ED			OM-PCNN+PART			
d	2	3	4	2	3	4
Bark	0.927	0.976	1	1	1	1
Bricks	1	1	1	1	1	1
Bubbles	1	1	1	1	1	1
Grass	1	1	1	1	1	1
Leather	1	1	1	1	1	1
Pigskin	1	1	1	1	1	1
Rafia	1	1	1	1	1	1
Sand	1	1	1	1	1	1
Straw	1	1	1	1	1	1
Water	1	1	1	1	1	1
Wool	0.983	9.992	1	1	1	1
Wood	1	1	1	1	1	1
Wave	1	1	1	1	1	1

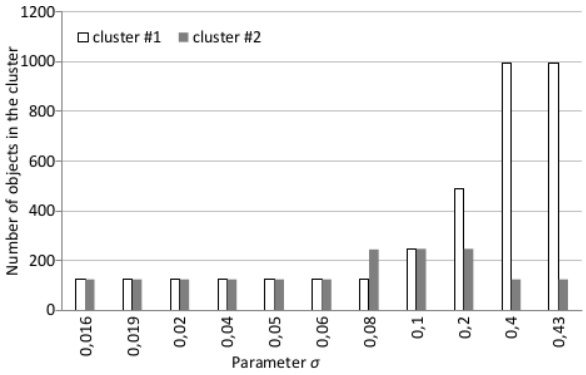

Figure 13: Impact parameter σ on the accuracy of image classification in clusters 1 and 2

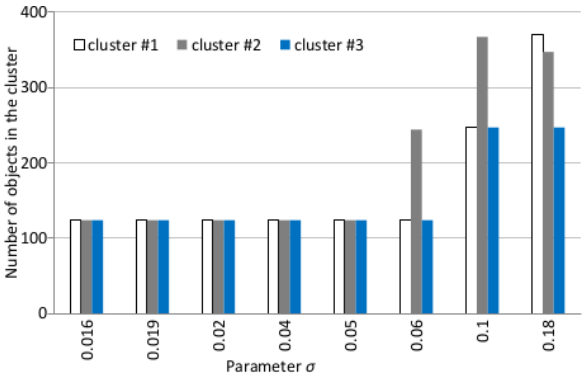

Figure 14: Impact parameter σ on the accuracy of image classification in clusters 1, 2 and 3

of groups of pictures of textures. Before each experiment, was preset desired number of clusters including the identification of a set of images pertaining to particular clusters. The number of iterations of the algorithm of aggregation MA-PART was set to *epoch* = 2. The following sections contain a more detailed analysis of the classification accuracy paintings textures in selected groups of clusters. The individual clusters are numbered sequentially from 1 to 13. For each group a significant and insignificant images was automatically constructed a graph of distance parameter σ on the number of generated clusters. Vigilance parameter in these cases was set to an optimal value of ϱ=3.

The Fig. 13 shows the detail of the classification of images of textures within the clusters 1 and 2. From the figure it follows that the classification takes place by means of incorrectly PART the distance parameter $\sigma >$ 0.06. Fig. 14 shows the classification result for the 3 clusters 1, 2 and 3. A plurality of 372 paintings textures were classified into the correct number of classes at the interval parameter σ between 0.016 after 0.18. Fig. 15 shows the progress of the texture image classification 5 clusters 1, 2, 3, 6 and 7 is an error in the classification in such case, the present value of the parameter $\sigma >$ 0,053. Last Fig. 16 includes 7 major classification for clusters 4, 5, 6, 7, 9, 10 and 12.

The classification was considered successful only when preselected classes of pictures fell all to belonging images textures.

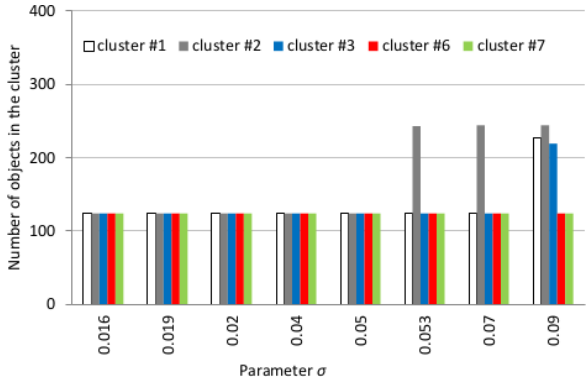

Figure 15: Impact parameter σ on the accuracy of image classification in clusters 1, 2, 3, 6 and 7

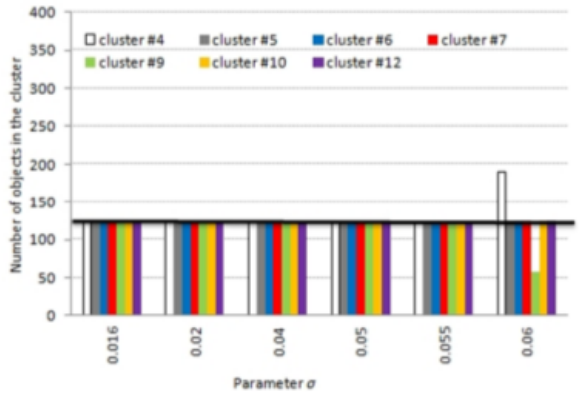

Figure 16: Impact parameter σ on the accuracy of image classification in clusters 4, 5, 6, 9, 10 and 12

5. Conclusion

Modified PART, based on the original PART, was proposed for clustering multidimensional text and image data, to achieve an adequate success rate of clustering and as low computational complexity of the model as possible. In addition, the aim of the author was to propose the algorithm design to find optimal network parameters. The proposed changes in the modified PART reached the author based on the analysis and change of the distance function $d(x_i, z_{ji})$, which is substantially applied in the process of winning neuron selection in clustering, and causes more accurate classification of input data into clusters. Another benefit was the introduction and maintenance of a matrix vectors of projective dimension of clusters, as well as establishing and maintaining centroids of the created clusters.

The improvement of the properties of the original clustering algorithm PART were verified on the system of generated dictionary of keywords from documents. Success rate vectors clustering of keywords into corresponding clusters reached by MA-PART 94,83% by original PART 85,32% . In the case, when MA-PART was added with refinement algorithm, success rate incerases to 99,56%.

Clustering of text documents using a model MA-PART were created clusters of documents assembled using regrouping algorithm based on the used similarity measure over a set of centroids of created clusters. In this case the accuracy achieved nearly 94%.

PART hybrid model in conjunction with OM-PCNN allows a significant reduction in dimensions of image space classification, while clustering by PART achieved 100% success rate aggregation of images with a known number of existing categories. The hybrid model using a modified clustering algorithm and using the PART outlier cluster to classify image data confirmed again the ability of PART to filter out unnecessary data when the number of created clusters and a set of etalons of significant classes is known in advance.

Future from the perspective of this dissertation opens up a number of ideas for a solution. The first of them could be testing of modified PART model on a large set of text documents using parallel processing. The second topic could be finding a suitable method for reducing the multidimensional space in tf-idf matrix of text documents and keywords and subsequent clusetring by PART. Another topic could be application of PART in the processing of audio data.

Acknowledgements. This work was partially supported by the Slovak Scientific Grant Agency (VEGA) in the research projects of No. 2/0211/09 - Service-oriented distributed computing and data management, No. 2/0184/10 - Intelligent methods for extensive information resources processing, and No. 2/0054/12 - Selected methods, approaches and tools for distributed computing.

References

[1] C. Aggarwal and C. Reddy. Data clustering algorithms and applicatons. 2014.

[2] R. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval. 1999.

[3] Y. Cao and J. Wu. Projective art for clustering data sets in high dimensional spaces. *Neural network*, Vol. 15(1):105–120, January 2002.

[4] Y. Cao and J. Wu. Dynamics of projective adaptive resonance theory model: the foundation of part algorithm. *Neural Networks*, Vol. 15,:245–260, March 2004.

[5] R. Chen and C. Chuang. Automating construction of a domain ontology using a projective adaptive resonance theory neural network and bayesian network. *Expert systems*, Vol. 25(4):414–430, 2008.

[6] R. Forgáč and I. Mokriš. Parameter influence of pulse coupled neural network for image recognition. *Journal of Applied Computer Science,*, Vol. 9(2):31–44, 2001.

[7] R. Forgáč and I. Mokriš. Feature generation improving by optimized pcnn. In A. Szakal, editor, *SAMI 2008 - 6 th International Symposium on Applied Machine Intelligence and Informatics,*, pages 203–207, January 2008.

[8] S. Grossberg and G. Carpenter. The art of adaptive pattern recognition by self-organizing neural network. *Computer*, Vol. 21(1):77–88, 1988.

[9] S. Grossberg and G. Carpenter. *Adaptive Resonance Theory. The Handbook of Brain Theory and Neural Networks.* MIT Press, April 2002.

[10] J. Hunter and J. Milton. Clustering neural spike trains with transient responses. In *Proceedings of IEEE Conference on Decision and Control*, pages 2000–2005, Cancun, Mexico, December 2008.

[11] V. Kvasnička and Ľ. Beňušková. *Úvod do teórie neurónových sietí.* Iris Bratislava, 1997.

[12] V. Mařík and J. Lažanský. *Umělá inteligence.* Academia Praha, 2003.

[13] E. Muller and T. Seidl. Evaluating clustering in subspace projections of high dimensional data. *Proceedings of the VLDB Endowment,*, Vol. 2(1):1270–1281, August 2009.

[14] L. Parson and H. Liu. Subspace clustering for high dimensional data. In *SIKDD Explorations 2004,*, pages 90–105, 2004.

[15] B. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996.

[16] H. Takahashi and H. Honda. Gene screening method for prognostic prediction using projective art model. *Genome Informatics*, 14:320–331, 2003.

[17] S. Zhong and J. Ghosh. A comparative study of generative models for document clustering. In *SIAM, International Conference Data Mining Workshop on Clustering High Dimensional Data and its Applications,*, pages 374–384, San Francisco, May 2005.

Selected Papers by the Author

R. Krakovský, R. Forgáč. Neural Network Approach to Multidimensional Data Classification via Clustering. In Aniko Szakal editor, *Proceedings of SISY 2011 – 9 th IEEE International Symposium on Intelligent Systems and Informatics*, pages 169–174, Subotica, Serbia, 2011.

R. Krakovský, R. Forgáč. Neural network model for multidimensional data classification via clustering with data filtering support. In Aniko Szakal editor, *Proceedings of SISY 2012 – 10 th IEEE International Symposium on Intelligent Systems and Informatics*, pages 169–174, Subotica, Serbia, 2012.

R. Krakovský, R. Forgáč, I. Mokriš. Influence of Cluster Center Selection on Clustering by Hybrid Neural Network Model. In Aniko Szakal editor, *Proceedings of LINDI 2012 - 4 th IEEE International Symposium on Logistics and Industrial Informatics*, pp. 5, Smolenice, Slovakia, 2012.

R. Krakovský, I. Mokriš. Clustering of Text Documents by Projective Dimension of Subsapces using PART Neural Network. In Aniko Szakal editor, *Proceedings of SACI 2012 - 7 th IEEE International Symposium on Applied Computational Intelligence and Informatics*, pp. 203-208, Timisoara, Romania, 2012,

R. Krakovský, I. Mokriš. Clustering of Text Collections Based on PART Neural Network and Similarity Measure. In Aniko Szakal editor, *Proceedings of ICSSE 2013 - IEEE International Conference on System Science and Engineering*, pp. 253-257, Budapest, Hungary, 2013.

R. Krakovský, R. Forgáč, I. Mokriš. Modified Clustering Algorithm for Projective ART Neural Network. In Aniko Szakal editor, *Proceedings of INES 2014 - 18 th IEEE International Conference on Intelligent Engineering Systems*, pp. 245-250, Tihany, Hungary, 2014.