24

Dynamic Data Mining: Synergy of Bio-Inspired Clustering Methods

Elena N. Benderskaya[1] and Sofya V. Zhukova[2]
[1]*St.Petersburg State Politechnical University,*
[2]*Graduate School of Management,*
Russia

1. Introduction

Dynamic data mining (DDM) comprises advantages of static methods used to reveal implicit structure of classes and at the same time benefits from high quality results obtained in the field of time series analysis. Clustering problem is recognized to be the most crucial in almost any knowledge domain: telecommunications and networking, nanotechnology, physics, chemistry, biology, health care, sociology, economics, etc (Aliev & et. al., 2008; Ceylan & et. al., 2009; Chee & Schatz, 2007; Ghosh & et. al.; 2008; Pedrycz & Weber, 2008; Xu & et. al., 2010). Scientists are in chase of new materials and new decision making techniques that manage data, information, devices or people on the fly. Centralized management techniques are mainly ineffective if we need to operate heaps of redundant information under uncertainty in real time. Decentralized control of interconnected elements, called networks, collectives, colonies, ensembles, maps is based on self-organization and bio-inspired principals that underlie amazing effects applied in highly interdisciplinary environment.

In the paper we extend the thorough comparative analysis of bio-inspired methods provided in resent research (Blum & Merkle, 2009; Budyan & et. al., 2009; Dressler & Akan, 2010) for benefit of clustering problem. Under consideration are the following bio-inspired approaches, used to reveal implicit data structures: small-world networks, ant-based networks, fuzzy logic, neural networks, chaotic map lattices, classical data mining, self-organizing maps. The general view on advantages and delimitations on various bio-inspired methods combinations is proposed in the form of a decision tree. There are so many combinations of bio-inspired methods (Crespo & Weber, 2005; Georgieva & Klawonn, 2008; Jaimes & Torra, 2010; Kaiser & et. al., 2003, 2007; Li & Shen, 2010, Sussillo & Abbott, 2009) with various extent of effectiveness that we tried to propose a systematic approach to reveal best practices. We state that significant advantages in terms of high quality clustering results can be obtained when complexity of both structure and dynamics is commensurable with complexity of problem. This is possible when we tune the harmony of more than two or three techniques. Distributed manner of decision-making processes in nature dictate multiform compensation of possible ineffective functioning of separate system's element by collective dynamics of all other elements.

Detailed analysis of simultaneous clustering techniques application within one method is given on the example of chaotic neural network (Benderskaya & Zhukova 2008, 2009). The

synergy of bio-inspired methods combination makes possible the solution of general clustering problem (no a prior information about topology and number of clusters is available). For the first time we managed to demonstrate the flexibility of the developed dynamic data mining technique as it happens to solve not only clustering problem, but classification problem as well. Fragmentary synchronization of thousands nonlinear elements stands to be stable when new players appear in the collective. We found out that chaotic neural network can classify simultaneously not one but many more objects. To demonstrate wide set of CNN applications we introduce the results on texts categorization that aims to improve the quality of search engines in the Internet.

2. Bio-inspired clustering methods

For centuries humans admire animate nature and accessories applied by life creatures to fulfil various functions. At first it was just formal resemblance and mechanistic imitation, then along with sciences maturity the focus shifted on inner construction of living systems.

However due to the complexity of a living system it is reproduced partly. Separate subsystems embody limited set of functions and principals. Just independently showed up artificial neural networks (attempts to mimic neural system), genetic algorithms (data transfer by means of inheritance), artificial immune systems (partial reproduction of immune system), evolutionary modeling (imitation of evolution development principals). The idea of natural self-organization within individuals is the basis for swarm and ant colony technologies (Handl & Meyer, 2007; Blum & Merkle, 2009). It is important to note that nearly all mentioned technologies deal with distributed parallel data processing thanks to numerous simple processing units comprised into self-organized networks that adapt to ever-changing environment (input information).

Of course there exit substantial peculiarities in the types of local cooperation and global behavior mechanisms predetermined by system's goal (as it is well-known systems demonstrate not only interconnectivity of elements but their ability to serve one purpose).

Evolution of society, new computer technologies have in common the idea of small worlds modelling. Communities of various natures (interests clubs, computer clusters, marketing networks, etc.) speak up for strong local linkage of units and weak connectivity outward nearest neighbors (nodes of the net).

Recent research on brain activities gives evidence for its cluster organization (Kaiser, 2007). So we can generalize that small-world models reflect both animate nature and abiocoen. Originally the notion *bio-inspired* comprised problem solving approaches borrowed from living systems but nowadays it is understood more widely. Results in physics in the field of chaos theory and nonlinear dynamics contribute greatly to bio-inspired methodology as soon as nonlinear chaotic models find their application in data mining – first and foremost bio-inspired scientific area. We propose to classify bio-inspired methods on different issues:

a. *structure and connection:* neural networks (macro level) and artificial immune systems (micro level);
b. *collective behaviour:* ant-based networks, swarm methods, multi agent systems, small-world networks;
c. *evolution and selection*: genetic algorithm, evolutionary programming and evolutionary modelling, evolutionary computations;
d. *linguistics*: fuzzy logic.

To step forward with generalization one can note that nearly all mentioned methods realize collective data processing through adaptation to external environment. Exception is fuzzy logic more relative to classical mathematics (interval logic reflects the diversity of natural language descriptions) (Choi & Chung-Hoon Rhee, 2009; Mendel, 2009).

Though bio-inspired methods are applied to solve a wide set of problems we focus on clustering problem as the most complex and resource consuming. The division of input set of objects into subsets (mainly non-overlapping) in most cases is interpreted as optimization task with goal function determined by inter and inner cluster distances. This approach obliges the user to give the a priori information about priorities: what is of most importance - compactness of clusters and their diversity in feature space or inner cluster density and small number of clusters. The formalization process of clustering problems in terms of optimization procedures is one of the edge one in data mining (Handl & Meyer, 2007; Herrmann & Ultsch, 2008, 2009).

Recent modifications of bio-inspired methods are developed as heuristics. The desire to enlarge the abilities of intellectual systems a separate knowledge domain within artificial intelligence field revealed (Lin & Lee, 1998; Georgieva & Klawonn, 2008; Pedrycz & Weber, 2008; Boryczka, 2009). Soft computing (SC) considers various combinations of bio-inspired methods. As a result there appeared such hybrid methods like: neural-fuzzy methods, genetic algorithms with elements of fuzzy logic (FL), hybrid comprised by genetic algorithms (GA) and neural networks (NN); fuzzy logic with genetic algorithm constituent, fuzzy systems with neural network constituent, etc. One of the main ideas of such combinations is to obtain flexible tool that allow to solve complex problems and to compensate drawbacks of one approach by means of cooperation with another.

For example, FL and NN combination provides learning abilities and at the same time formalized knowledge can be represented due to fuzzy logic element (Lin & Lee, 1998). Fuzzy logic is applied as soon as we want to add some flexibility to a data mining technique. One of the main drawbacks of all fuzzy systems are absence of learning capabilities, absence of parallel distributing processing and what is more critical the rely on expert's opinions when membership functions are tuned. In advance to input parameters sensitivity almost all methods suffer from dimension curse and remain to be resource consuming. The efficiency of these methods depends greatly on the parallel processing hardware that simulate processing units: neurons of neural networks, lymphocyte in artificial immune systems, ants and swarms, agents in multi-agent systems, nodes in small-world networks, chromosomes in genetic algorithms, genetic programming, genetic modeling.

We can benefit from synergetic effects if consider not only collective dynamics but also physical and chemical nature of construction elements – nonlinear oscillators with chaotic dynamics. As it is shown in numerous investigations on nonlinear dynamics: the more is the problem complexity the more complex should be the system dynamics. All over the world investigations on molecular level take place to get new materials, to find new medicine, to solve pattern recognition problem, etc. Most of them consume knowledge from adjacent disciplines: biology, chemistry, math, informatics, nonlinear dynamics, and synergetics.

3. Clustering challenge

During the last decade three curses formed an alliance: great volume of information, its increasing variety and velocity of data processing. These 3Vs predetermine strict quality requirements to data mining systems. The costs of wrong decisions increase exponentially as

the environment changes rapidly. Under this condition the development of automatic clustering systems seems to be one of the most pressing problems. At the moment the greater part of existing clustering systems are semiautomatic. And the key reason for this is the multiformity of datasets that cannot be formalized in one unified way.

Clustering problem is the most complex problem among those defined in Data Mining. The set of elements division into non-overlapping groups (clusters) is provided via criterion of similarity that predetermines the result. In terms of neural networks it is solved by means of unsupervised learning or learning without a teacher (Dimitriadou & et. al., 2001), because the system is to learn by itself to extract the solution from input dataset without external aid. Thus the division must be provided automatically.

To illustrate the representative clustering problems a collection of test datasets was generated and arranged in fundamental clustering problems suite (FCPS). FCPS offers a variety of clustering problems any algorithm shall be able to handle when facing real world data. FCPS serves as an elementary benchmark for clustering algorithms.

FCPS consists of data sets with known a priori classifications (Morchen & et. al., 2005; Ultsch, 2005, a) that are to be reproduced by the algorithm. All data sets are intentionally created to be simple and might be visualized in two or three dimensions. Each dataset represents a certain problem that is solved by known clustering algorithms with varying success. This is done in order to reveal benefits and shortcomings of algorithms. Standard clustering methods, e.g. single-linkage, ward und *k*-means, are not able to solve all FCPS problems satisfactorily.

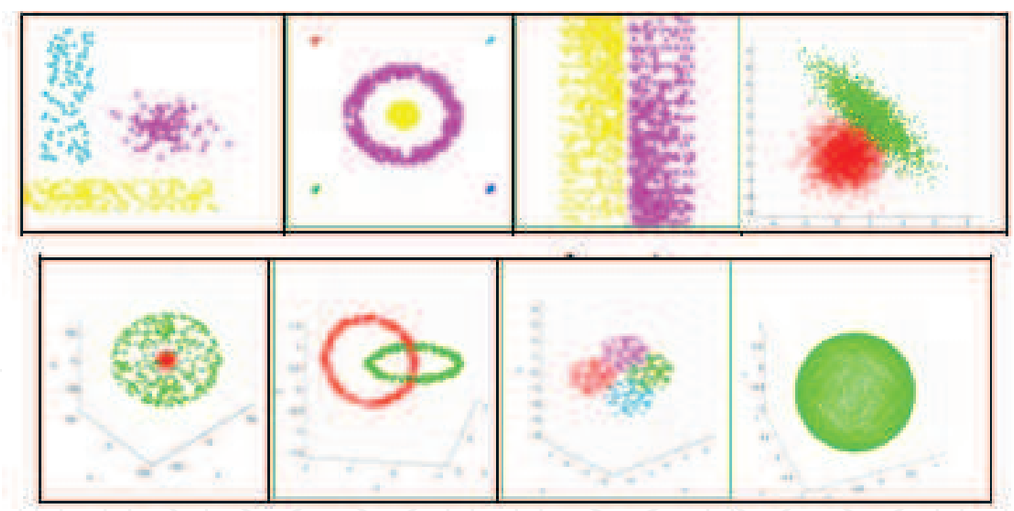

Fig. 1. Fundamental problems clustering suite: 2D and 3D clustering problems vary in density, mutual distance and compactness of clusters (Ultsch, 2005, b).

To solve clustering problem a lot of clustering techniques were developed to reveal most appropriate division of objects in the input dataset in terms of concrete measures of similarity (metrics). There are two types of metrics (Oliveira & Pedrycz, 2007; Han & Kamber, 2005): type 1 - similarity measure between objects within a cluster (euclidean, cityblock, Mahalanobis, Minkowski, cosine, Chebyshev, supremum norm); type 2 - similarity (or dissimilarity) measure between the clusters themselves (single linkage, complete linkage, median clustering, centroid clustering, Ward's method, statistical clustering).

The similarity measure depends greatly on mutual disposition of elements in the input dataset. If we have no a priori information about the type of groups (ellipsoidal, ball-shaped, compact, scattered due to some distribution or just chaotically, and this list is endless) then the probability of erroneous measure choice is very high (Han & Kamber, 2005; Kumar & et. al., 2006; Eidswick, 1973). If our hypothesis about the clusters interrelations or their form or their density does not fulfill then the application of clustering method to this dataset will perform erroneous results.

To overcome the data uncertainty expressed in unformalized variety of possible clusters interrelations usually an expert estimations are used to decide on the choice of clustering technique or interpret clusterization results. Without an expert each time application of a method to concrete dataset (when there is no a priori information available) is a roulette game. This is a serious obstacle on the way to automatic clustering.

To summarize there are three targets to be hit by one clustering technique: it should be fast in terms of calculations, independent to the information about number and topology of clusters, flexible to reveal inner structure of input dataset. So the main question is how to accomplish all this issues in one method.

4. Dynamic data mining

The most perspective direction is based on the attempts to model the work of human brain, which is a highly complex, nonlinear and parallel information-processing system. Complex cortex structure is modelled and formed by artificial neuron lattices, which are joined by great amount of interlinks. This global link of simple neurons provides their collective behaviour. Each neuron carries out the role of a processor. That's why neuron network structure is the most appropriate base for parallel computing – there is no need to prepare data (in neural network input data is already parallelized). For parallel computing to work correctly software should be able to partition its work and data it operates on over hundreds of processors. High speed and with the same time high quality solution of most various complicated problems can be received by means of microsystem's collective behaviour property. The main idea of self-organization is in distributed character of data processing, when one element dynamics means nothing, but at the same time group dynamics define macroscopic unique state of the whole system, that allows this system to reveal capabilities for adaptation, learning, data mining and as one of the results - high computation effectiveness.

Advances in experimental brain science give evidence to the hypothesis (Borisyuk & et. al., 1998; Borisyuk, R.M, & Borisyuk, G.N, 1997) that cognition, memory, attention processes are the results of cooperative chaotic dynamics of brain cortex elements (neurons). Thus the design of artificial dynamic neural networks on the base of neurobiological prototype seems to be the right direction of the search for innovative clustering techniques. Computer science development predetermined promising possibilities of computer modeling. It became possible to study complex nonlinear systems. Great evidence for rich behavior of artificial chaotic systems was accumulated and thus chaos theory came into being (Schweitzer, 1997; Mosekilde & et. al., 2002; Haken, 2004). Dynamics exponential unpredictability of chaotic systems, their extreme instability generates variety of system's possible states that can help us to describe all the multiformity of our planet.

It is assumed to be very advantageous to obtain clustering problem solution using effects produced by chaotic systems interaction. In this paper we try to make next step in the

development of universal clustering technique. Dynamic data mining combines modern data mining techniques with modern time-series analysis techniques.
To mimic nature highly unstable dynamics and distributed data processing were combined. Thus chaotic neural network (CNN) came into being originally in the form of Angelini's model (Angelini, 2003, Angelini & et. al., 2001). We modified the model greatly in order to generate clustering results of a better quality. As it is shown on Fig. 2 CNN is a recurrent neural network with one layer of n neurons. Each neuron corresponds to one point in the input dataset which in general case consists of n objects, each described by p features (p-dimensional image). CNN is a dynamic neural network, where each processing unit changes its state depending on the dynamics of all other neurons

$$y_i(t+1) = \frac{1}{C_i}\sum_{i\neq j}^{n} w_{ij} f(y_i(t)), \quad t = 1...T, \tag{1}$$

$$C_i = \sum_{i\neq j}^{n} w_{ij}, \; i, j = \overline{1,n} \tag{2}$$

$$W = \{w_{ij}\} = \exp\left(-d_{ij}^{2} / 2a\right), \quad i, j = \overline{1,n}, \tag{3}$$

$$f(y(t)) = 1 - 2y^2(t), \tag{4}$$

where n – number of neurons, w_{ij} - strength of linkage between elements i and j, d_{ij} - Euclidean distance between neurons i and j, a – local scale, depending on k-nearest neighbors, T – time interval. The a value is average number of neighbors, calculated via Delaunay triangulation. The initial state of neural network is described by random values in the range [-1, 1].

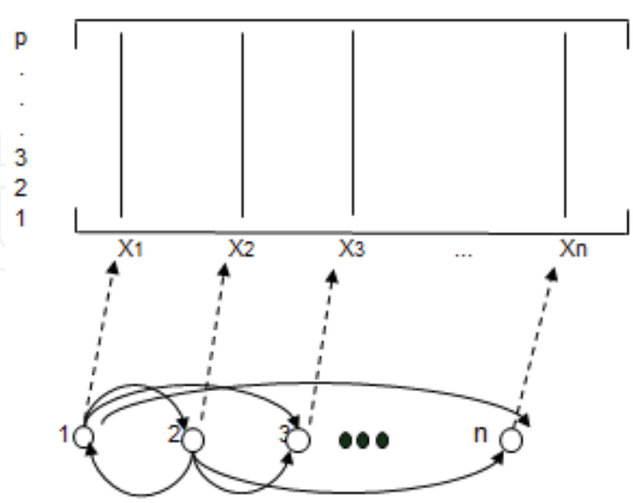

Fig. 2. The relation of input dataset and structure of chaotic neural network.

Key point of CNN functioning is the emergence of cooperative dynamics between neurons outputs via time. After some transition period they start to change states synchronously.

The synchronization extent depends greatly on mean field formed by linkage strengths. Primary results on modeling high dimensional chaotic map lattices were published by K. Kaneko (Kaneko, 1987). These works showed up the fact that globally coupled chaotic map lattices exhibit formation of ensembles synchronously oscillating elements. These ensembles Kaneko called clusters that serve as system's attractors. If there appear to be several clusters then the system is characterized by multistability, when several attractors coexist in the phase space at the same parameters values. He showed that in case of w_{ij}=0 chaotic map lattice dynamics converges to the ordered stage that corresponds to strange attractor (there are several groups of completely synchronized neurons); partly ordered stage (large oscillatory clusters coexist with a lot of small ones); turbulent phase (there are no big clusters only small ones evolve independently from each other).

In Angelini's model uncertainty about topology and number of clusters was replaced by the uncertainty about w_{ij} parameters that depend on a priori unknown value of k-nearest neighbors used in (3). Though the idea to make neural network an inhomogeneous one and set the linkage strengths via (3) is great.

To explore the chaotic neural dynamics (oscillatory clusters) we thoroughly use several visualization techniques: representation of total output dynamics and phase portraits. Exactly because of the deep analysis of the output dynamics we managed to discover new type of synchronization (fragmentary synchronization) and the way to control CNN dynamics.

This paper fully corresponds to the chaotic logic approach that comprises ideas from both humanitarian and natural sciences. Terminology of chaos theory is still a subject of heated discussions. That's why we would like to clear out the language. The key idea of the paper consists in the synthesis of investigations results in such fields as: topology (both structural and spatial), synchronization as a universal concept, control of many-dimensional nonlinear systems, times series analysis, neural networks application as a computing base, data processing and informatics. Due to this great variety of issues we use terminology from different scientific areas slightly interpreting or enriching the vocabulary.

The most crucial word of the paper is chaotic. The word chaos is naturally associated with extremely unpredictable systems dynamics, but not with the stable, and recurrent reproduction of the same results. And in the case of clustering problems we need to generate the only solution every time we use the same method. The chaotic dynamic of CNN is guaranteed by logistic map (4). We want to stress that chaos in CNN dynamics is important only to ensure the sufficient level of instability to make the emergence of self-organizing phenomenon possible. Though instant states of neurons remain to be chaotic mutual synchronization of elements due to the phenomenology of CNN is stable.

The phenomenology of CNN can be examined by the outputs dynamics analysis. The statistics on instant changes of CNN outputs via (1) is gathered after some transition period.

To analyse statistics complex time-series analysis should be accomplished. We applied various techniques and realized that none of them is adequate for time series generated by CNN. The reason consists in various types of synchronization that takes placeCNN may produce not only well-known synchronization but also such synchronization when instant output values in one cluster do not coincide neither by amplitude nor by phase and there is even no fixed synchronization lag. In spite of everything joint mutual synchronization exists within each cluster. This synchronization is characterized by individual oscillation cluster melodies, by some unique "music fragments" corresponding to each cluster. From this follows the name we give to this synchronization type - fragmentary synchronization.

Thus dynamic data mining is realized in the form of a shift from static output analysis to dynamic one. The input dataset is given to logistic map network by means of inhomogeneous weights assignment via (3). It means that input dataset is not given to neural network in the classical meaning. On the opposite the structure of chaotic neural network adapts every time to the input image. We want to stress that CNN can equally find clusters in the dataset of any dimension, because the compression via Euclidean metric evaluation is provided. As it is shown on Fig. 2 the number of objects in the input dataset coincides with the number of neurons in CNN. And we can say that each neuron represents its own point in the dataset. On Fig. 3 you can see the dynamics of CNN generated in response to different clustering problems.

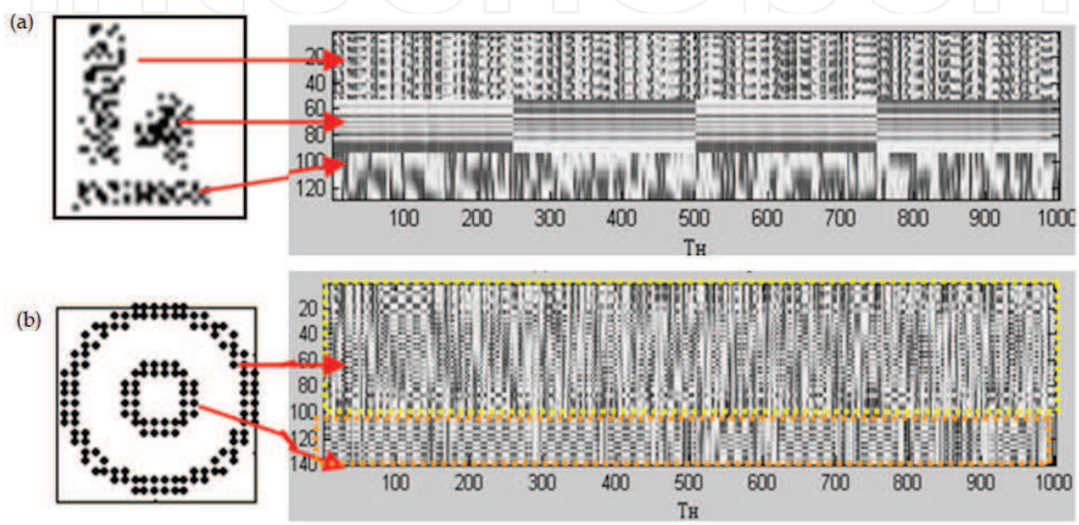

Fig. 3. Result of dynamic data mining consists in oscillatory clusters that in their turn are comprised by fragmentary synchronized outputs of chaotic neural network: (a) – three clusters found by CNN without any a-priori information in response to 2D clustering problem; (b) – two clusters found by CNN without any a-priori information in response to 2D clustering problem.

5. Synergy of bio-inspired methods

Separate class of dynamic neural networks - oscillatory networks - is successfully applied to solve segmentation problems. Thus, to solve highly complicated problems it is appropriate to combine achievements in nonlinear dynamics, self-organization theory and neural networks theory. In our project this integration is represented by investigation of new, complex recurrent neural networks type, that seems to be very perspective in future.

The proposed clustering technique possesses features almost every of bio-inspired methods mentioned above:

a. from *small–world networks* we take irregular, incomplete linkage between elements in (clusters are formed by nearest neighbours);
b. from *ant-based networks* we take parallel performance of elements (the solution is generated both by individual and collective dynamics of elements);
c. from *genetic algorithms* we take iterative improvement of intermediate solution by means of previous experience interchange (extensive search of best fit solution);

d. from *fuzzy logic* we take interval logic in post processing of clustering results (both vertical when we analyse fractal structure of system's output dynamics and horizontal when time-series analysis is conducted);
e. from *neural networks* we take processing element with complex transfer function (logistic map) and stress that in case of new technique its dynamics can be interpreted as learning process;
f. from *classical self-organizing map*s we take k-means metric.

Statistical methods, based on the idea of sample average, bring up to information losses. That's why the information processing must be similar to biological prototype. We solve the problem of CNN processing by accurate time-series analysis.

The clustering of all input datasets from Fig. 1 was successfully fulfilled by CNN thanks to refuse to the avoidance of metrics obtrusion. New method is based on determination of implicit law in data structure by means of self-organization. So CNN can be classified as oscillatory recurrent neural network, comprised by one layer inhomogeneous connected neurons, each with chaotic behaviour. The oscillatory nature of CNN makes it similar to brain dynamics and allows to make hypothesis on wide dynamic data mining abilities.

Further on we show application of proposed clustering technique to solve classification problems and multidimensional text categorization problem.

6. CNN classification and clustering

Classification problem is more simpler than clustering problem because there is information about classes representatives – centres of classes. In self-organizing maps (Kohonen's neural network is one to be easily compared with) learning process stops as soon as the centres of clusters stop changing during the search procedure (Kohonen, 1995, Liu & et. al., 1999). The answer to clustering problem is given by SOM after learning process in the following way: each point in the input dataset is shown to SOM and the nearest centre of cluster attracts the point – thus it is classified. We stress that classical SOM can't generate the answer "I do not know". CNN during learning process does not estimate centres of clusters. Though in this paper we propose to use CNN to solve classification problem as well.

To solve classification problem on the basis of CNN we need to solve previously clustering problem for concrete input dataset and then add new points (new neurons to neural network). For this added neurons we calculate outputs dynamics without recalculation of all other neurons outputs. Further on we analyse joint output dynamics.

We made different experiments on classification of additional points. Fascinating thing we discovered about CNN is that the network can classify at the same time not one or two but a lot more points. These new points, being in the neighborhood of a cluster join this cluster. The criteria of joining one of the clusters deals with the density of new points added to the original dataset. If density of the added points is equal or exceeds the density of the nearest cluster then these new points join the cluster. Otherwise CNN generates a new cluster – that is similar to the answer "I do not know". Thus CNN allows to solve both clustering and classification problems almost simultaneously that expands greatly its application domain.

7. Text categorization application of CNN

In our previous papers we tested CNN on 2D, 3D input datasets from FCSP. Here on we want to check whether CNN can cope with real data – sets of objects, each described by *p*-

dimensional number of features. We made our focus on Internet Search Engines and questionable results they generate as relevant in response to users «keywords sequence» requests. There are a lot of search optimization techniques that we do not take into account, focusing mainly on the quality of clustering results.

The aim of quantitative text analysis is the determination of some textual structure laws and contents laws, expressed in quantitative measure. These laws may concern some style features of an author, belonging to this or that genre, literary or scientific school, etc. Then these laws are used for text systematization by structure or by content.

Nowadays actual problems are such as determination of psychological condition of an author at the moment of text creation, authorship determination, and formal content determination.

To solve these problems it is important to define correlation between words in one document or in group of documents to find out formal structure of texts and as a result quantitative measure of their similarity.

To check the hypothesis whether CNN is applicable to solve text categorization problem we planned an experiment. We chose "neural networks" thematic from "artificial intelligence" domain. To model the situation as close as it is possible to the real information search problem we added texts to the input bunch of documents that belong to adjacent thematic – "fuzzy logic" and "machine learning" (ML). From the input set of documents CNN should find ones that are relevant to "neural networks" thematic. As a result we propose that CNN will divide all documents into clusters and one of them will comprise docs from "neural networks" thematic. In this interpretation of clustering problem the feature space consists of thematic's keywords. Input image for CNN consists of formal document representations – coded texts (the words in texts are replaced by zero if a keyword of the thematic occurs or one if not).

Text analysis methods are based on formal document presentation as some set of features that is previously formed by methods of textual preconditioning or by human being. If there are enough of input patterns (texts) and it is possible to classify each text to only one class in the previously formed set of classes then a solving rule can be used. This rule is received after learning with a teacher (probability classifiers, feed-forward network).

Clustering results reveal implicit law in text structure. Text clustering is characterized by such problems as great correlation between elements (texts), this cause big clusters intersections and absence of a priori information about clusters.

Hence "to measure" the text (its structure and its content as a result) it is proposed to use CNN. The order, created during functioning CNN – is the information that single neurons (or words, or documents) have about each other.

The objects to cluster are described by a set of features (feature vectors) and are represented as points in many-dimensional space. Consequently to verify the clustering results of different methods images with various complexity are applied. This approach allows not only to cluster the images (for example, analyze data taken from satellite image of an area) but produce visual demonstration of any clustering problem, because in this case it is interpreted as pattern recognition problem.

The decision result of the solving rule depends greatly on representation form of input data. In our case input data is the coded text. To form this code new knowledge in computer linguistics is applied: the idea of the text is highly correlated with distribution of clue words in text. To start text categorization we need to solve attendant tasks:

a. create database of keywords for various text categories;
b. choose text representation technique;
c. provide linguistic analysis (exclude prepositions, conjunctions, synonyms, etc.).

One of the reasons of unsatisfactory work of Information Processing Systems underlies in text representation methods drawbacks. There are several approaches to text representation: frequency, binary, compressive. The two first are applied more often. Frequency methods are based on calculations of clue words frequencies - times of appearance in the text. Binary methods fix absence ore presence of clue word in a document (the feature vector that describes a text will consist of nulls and ones: 0 – if the word is present, 1 – if absent. Frequency methods have some drawbacks due to frequency averaging-out of each of the clue words. This causes losses of meaningful information about text structure.

On the one hand it is undesirable when the same term is used several times in one sentence (it is better to apply synonyms). The result of this construction sentence rule is that the total term frequency appearance may be little. On the other hand very frequent usage of a term doesn't guarantee that the document's theme corresponds with the term.

The resent researches in computer linguistics show that the idea of the text is highly correlated with text structure – distributions of single clue words, pairs of clue words, combinations of three words (and so on). Similar in content and level of complexity texts must have similar structures – not only the average frequencies of clue words usage.

In this article a new approach to text representation is proposed. It is based on clue words distributions within texts. To do this a document is divided into M groups of words. N is defined by Sturges rule that tells us how much intervals to choose to construct a bar chart of distribution (for each document its own M is defined). Then usage frequencies of each clue word in each of group are calculated. Than the larger is the document than the larger is the group of word.

The advantage of distributed text frequency representation is in the decrease of information losses. Because with minimum computational spendings on the input of clustering system comes a text representation with little distortion in text specificity. Thus defining correlation in group words usage allows revealing the inner structure of the text and to a certain extent its content-idea.

The results of text clustering showed possibility of practical CNN application. Text collections, represented in D-dimensional feature space of clue words and belonging to three intersected themes were successfully clustered. The text representation method influence on clusterization quality was investigated and the results analysis permits to make a conclusion that proposed frequency distributed method gives better results over classical frequency method. The results of correct clustering are 3-5% more than in classical frequency method. The documents number increase makes better clustering quality.

To solve real clustering D-dimensional problem we can't visualize D-dimensional input objects, but we can preliminary solve by ourselves and compare the results. We created 3 collections of documents that belong to three thematics in different ratio: 30(NN)-30(FL)-30(ML), 50(NN)-20(FL)-20(ML), 20(NN)-40(FL)-40(ML). As a result dynamic data mining technique generated an answer that to the extent of 80% coincide with desired one, 6% of documents were not joined to any cluster as they were considered as noise. To increase the clustering quality we think it is important to expand number of documents in the input dataset.

8. Conclusion

We analyzed the multiformity of existing clustering techniques and concluded that considered CNN model can be related to bio-inspired ones. The ability to solve complex clustering problems in terms of oscillations clustering language in future research can be extended by dynamic inputs (at the moment the *p*-dimensional input dataset is fixed unchangeable during processing time). It was discovered that CNN can also be applied as a classifier. This fact can save a lot of time when working with invariant images (the space rotation of input image has no effect on CNN – the only thing that matters is mutual location of objects). The preliminary results on *p*-dimensional clustering of real data by means of CNN were positive in terms of clustering quality. Successful results were predetermined also by new text representation technique that gave 3-5% better results. But the time, needed to reveal oscillatory clusters prevents from industrial implementation of dynamic data mining technique. The direction of further research deals with the domain of modern time-series data mining – the alternative to algebraic approach, used at the moment.
We highly appreciate partial support of this research by St. Petersburg government, Science and Graduate Education Committee (diploma PSP №090190).

9. References

Aliev, R. A.; Aliev, R. R.; Guirimov, B. & Uyar, K (2008). Dynamic data mining technique for rules extraction in a process of battery charging, *Applied Soft Computing*, No. 8, pp. 1252–1258

Angelini, L. (2003). Antiferromagnetic effects in chaotic map lattices with a conservation law, *Physics Letters* A 307(1), pp.41–49

Angelini, L., Carlo, F., Marangi, C., Pellicoro, M., Nardullia, M. & Stramaglia, S. (2001). Clustering by inhomogeneous chaotic maps in landmine detection, *Phys. Rev. Lett.* N86, pp.89–132

Benderskaya, Elena N. & Zhukova, Sofya V. (2008). Clustering by chaotic neural networks with mean field calculated via Delaunay triangulation, *Proceedings of Third international workshop on Hybrid artificial intelligence systems*, pp. 408-416, ISBN 978-3-540-87655-7, Burgos, Spain, September 24-26, 2008, Lecture Notes in Computer Science, vol. 5271, Springer-Verlag, Berlin-Heidelberg

Benderskaya, Elena N. & Zhukova, Sofya V. (2009). Fragmentary synchronization in chaotic neural network and data mining, *Proceedings of 4th international conference on Hybrid artificial intelligence systems*, pp. 319-326, ISBN 978-3-642-02318-7, Salamanca, Spain, June 10-12, 2009, Lecture Notes in Computer Science, vol. 5572, Springer-Verlag, Berlin-Heidelberg

Blum, Ch. & Merkle, D. (2009). *Swarm Intelligence: Introduction and Applications*, ISBN 978-3642093432, Springer

Borisyuk, R. M., & Borisyuk, G. N. (1997). Information coding on the basis of synchronization of neuronal activity, *Bio Systems*, Vol. 40, No 1., pp. 3-10

Borisyuk, R. M.; Borisyuk, G. N. & Kazanovich, Y.B. (1998). The synchronization principle in modelling of binding and attention, *Membrane & cell biology*, Vol. 11, No. 6, pp. 753-761

Boryczka, U. (2009). Finding groups in data: Cluster analysis with ants, *Applied Soft Computing*, No. 9, pp. 61-70

Budayan, C.; Dikmen, I. & Birgonul M. T. (2009). Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping, *Expert Systems with Applications*, No. 36, pp. 11772–11781

Ceylan, R.; Ozbay, Y. & Karlik, B. (2009). A novel approach for classification of ECG arrhythmias: Type-2 fuzzy clustering neural network, *Expert Systems with Applications*, No. 36, pp. 6721–6726

Chee, B. & Schatz, B. (2007). Document clustering using small world communities, *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, International Conference on Digital Libraries, Canada, 2007, pp. 53-62

Crespo, A. & Weber, R. (2005). A methodology for dynamic data mining based on fuzzy clustering, *Fuzzy Sets and Systems*, No. 150, pp. 267–284

Dimitriadou, E., Weingessel, A., Hornik, K. & Voting-Merging (2001). An Ensemble Method for Clustering, *Lecture Notes in Computer Science*, Vol. 2130

Dressler, F. & Akan, O. B. (2010). A survey on bio-inspired networking, *Computer Networks*, No. 54, pp. 881–900

Eidswick, J.A. (1973). On some fundamental problems in cluster set theory, *Proceedings of the American mathematical society*, Vol.39, No.1, pp. 163-168

Georgieva, O. & Klawonn, F. (2008). Dynamic data assigning assessment clustering of streaming data, *Applied Soft Computing*, No. 8, pp. 1305–1313

Choi, Byung-In & Chung-Hoon Rhee, Frank (2009). Intervaltype-2fuzzy membership function generation methodsf or pattern recognition, *Information Sciences*, 179, pp. 2102–2122

Ghosh, A., Halder, A., Kothari, M. & Ghosh, S. (2008). Aggregation pheromone density based data clustering, *Information Sciences*, No. 178, Elsevier, pp. 2816–2831

Haken, H. (2004). Synergetics. Introduction and Advanced Topics, *Physics and Astronomy Online Library*, p. 758, Springer

Han, J. & Kamber, M. (2005). *Data Mining. Concepts and Techniques*, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann

Handl, J. & Meyer, B. (2007). Ant-based and swarm-based clustering, *Swarm Intelligence*, Vol. 1, No. 2. (16 December 2007), pp. 95-113

Herrmann, L. & Ultsch, A. (2008). The Architecture of Ant-Based Clustering to Improve Topographic Mapping, *Proceedings of the 6th international Conference on Ant Colony Optimization and Swarm intelligence*, pp.379-386, ISBN:978-3-540-87526-0, Brussels, Belgium, September 22 - 24, 2008, Lecture Notes In Computer Science, vol. 5217. Springer-Verlag, Berlin, Heidelberg

Herrmann, L. & Ultsch, A. (2009). Clustering with Swarm Algorithms Compared to Emergent SOM, *Proceedings of the 7th international Workshop on Advances in Self-Organizing Maps*, St. Augustine, FL, USA, June 08 - 10, 2009, pp.80-88, ISBN:978-3-642-02396-5, Lecture Notes In Computer Science, vol. 5629, Springer-Verlag, Berlin, Heidelberg

Jaimes, L. & Torra, V. (2010). On the Selection of Parameter m in Fuzzy c-Means: A Computational Approach, *Integrated Uncertainty Management and Applications*, Vol. 68, pp. 443–452

Jang, J-S. R. & Sun, Ch.-T. (1997). *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence,* ISBN 978-0132610667, Prentice Hall

Kaiser, M. (2003). Mean clustering coefficients – On clustering measures for small-world networks, *Chaos*, Vol. 13, No. 3

Kaiser, M. (2007). Brain architecture: a design for natural computation, *Philosophical Transactions of the Royal Society A,* 2007, Dec 15, 365(1861), pp.3033-3045

Kaiser, M.; Gorner, M. & Hilgetag, C. (2007). Criticality of spreading dynamics in hierarchical cluster networks without inhibition, *New Journal of Physics*, Vol. 8, May 2007, pp. 110

Kaneko K. (1987). Phenomenology of spatio-temporal chaos, *Directions in chaos,*World Scientific Publishing Co., pp. 272-353, Singapore

Kohonen T. (1995). *Self-Organizing Maps*, Springer Verlag, Berlin

Kumar, B. V.; Mahalanobis, A. & Juday, R.D. (2006). *Correlation Pattern Recognitio*n, Cambridge University Press

Li, Y. & Shen, Y. (2010). An automatic fuzzy c-means algorithm for image segmentation, *Soft Computing,* No. 14, pp.123–128

Liu, Q.; Rui, Y.; Huang, T. & Levinson, S. (1999). Video Sequence Learning and Recognition via Dynamic SOM, Proceedings. International Conference on Image Processing, ICIP-99, vol. 4, pp.93 - 97

Mendel, Jerry M. (2009). On answering the question ''Where do I start in order to solve a new problem involving interval type-2 fuzzy sets?", *Information Sciences,* 179, pp. 3418–3431

Mörchen, F., Ultsch, A., Nöcker, M. & Stamm, C. (2005). Databionic visualization of music collections according to perceptual distance, *Proceedings 6th International Conference on Music Information Retrieval (ISMIR 2005)*, London, UK, pp. 396-403

Mosekilde, E.; Maistrenko, Yu. & Postnov, D. (2002). *Chaotic synchronization*, World Scientific Series on Nonlinear Science, Series A, Vol. 42

Oliveira, V. J. & Pedrycz, W. (2007). *Advances in Fuzzy Clustering and its Applications*, Wiley

Pedrycz, W. & Weber, R. (2008). Special issue on soft computing for dynamic data mining, *Applied Soft Computing*, No. 8, pp. 1281–1282

Schweitzer, F. (1997). *Self-Organization of Complex Structures: From Individual to Collective Dynamics*, CRC Press

Sussillo, D. & Abbott, L. F. (2009). Generating Coherent Patterns of Activity from Chaotic *Neural Networks, Neuron*, Vol. 63, pp. 544–557

Ultsch, A. (2005, a). Density Estimation and Visualization for Data containing Clusters of unknown Structure, *Proceedings 28th Annual Conference of the German Classification Society (GfKl 2004)*, Dortmund, Germany, Springer, Heidelberg, pp. 232-239

Ultsch, A. (2005, b). Clustering with SOM: U*C, *In Proc. Workshop on Self-Organizing Maps*, Paris, France, pp. 75-82

Xu, R.; Damelin, S.; Nadler, B. & Wunsch, D. (2010). Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps, Artificial Intelligence in Medicine, No. 48, pp. 91–98

Knowledge-Oriented Applications in Data Mining
Edited by Prof. Kimito Funatsu

ISBN 978-953-307-154-1
Hard cover, 442 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by 'Data Mining' address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Elena N. Benderskaya and Sofya V. Zhukova (2011). Dynamic Data Mining: Synergy of Bio-Inspired Clustering Methods, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-154-1, InTech, Available from: http://www.intechopen.com/books/knowledge-oriented-applications-in-data-mining/dynamic-data-mining-synergy-of-bio-inspired-clustering-methods

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821