Steering Clustering of Medical Data in a Self-Enforcing Network (SEN) with a Cue Validity Factor

Christina Klüver
University of Duisburg-Essen
Institute for Computer Science and Business Administration
Essen, Germany
christina.kluever@uni-due.de

***Abstract*—A Self-Enforcing Network (SEN), which is a self-organized neural network, is introduced to cluster medical data. In addition, a cue validity factor is defined, which affects the clustering of the data. The results show that a user can influence the clustering of data by SEN, thus allowing the analysis of the data depending on economical, medical or nursing interests. The described prototype includes concrete examples and shows the potential of such a network for the analysis of complex data.**

Keywords—self-enforcing network, cue validity factor, self-organized learning, neural networks, medical data

I. INTRODUCTION

Clustering of medical data means to deal with several problems at once. On the one side, a very large number of clustering algorithms are at disposal (e.g. [1-3]), on the other side, the medical data are increasing and the technical problems dealing with such big data are still a major challenge [e.g. 4].

Typically, medical data have not only a big volume; they are also very complex, containing different types of numerical data and text components, which have to be pre-prepared for a suited algorithm. The choice of a clustering algorithm is meanwhile difficult because of the increasing number of developed algorithms and techniques. Beside the classical cluster algorithms as 'k-means' [2] and further developments (e.g. 'Lazy Quantum Clustering' [5]), 'semi-supervised fuzzy clustering' algorithms [6], clustering selection by 'meta-learning systems' [7], 'hierarchical clustering methods' to name only a few, are developed in recent years to optimize different aspects of clustering, which are also used for medical data [8-11].

New technical developments in hardware- and programming-techniques provide support to handle the data stream in clustering algorithms [12-15], but the research is only just beginning.

All the techniques have advantages as well as disadvantages, but the greatest problem remains for users, who are not familiar with these different possibilities to analyze the data, as nursing staff or doctors without specific qualifications. In addition the interests of the analysis of the data are different, as a doctor is maybe interested in correlations between diseases and drugs, while a person from clinical administration is more interested in the costs incurred.

In consequence, it could be desirable to have an influence on the *building* of clusters. In last years additional algorithms were introduced e.g. for Self-Organized Maps (SOM), which are popular not only for clustering medical data. SOM belong to unsupervised learning neural networks and have a long tradition, but the results are not easy to interpret. As orientation support, an automatic "labeling" [16] or "colored marks" [17] were introduced for interpreting a trained map, because the features responsible for a cluster are not evident for a user. Additional recent developments such as a "weighted SOM" [18], introducing a user-specified instance-varying weight to improve the learning algorithm, combinations with techniques of Extreme Learning Machine (ELM) [19], to define the optimal structure for the clusters, or alternatives to SOM, e.g. Self-Adjusting Feature Maps (SAM) [20] using self-adjusting mechanisms to adapt the network size and keep a dynamical neighborhood topology, are only few examples for the efforts to optimize the learning processes and visualizations of the results.

In this article, a new clustering method is introduced, namely a Self-Enforcing Network (SEN), which is a self-organized learning neural network, using a "cue validity factor" to *steer* the clusters according to the interests with respect to the analysis. Subsequently the used medical data are presented and different results depending on the cue validity factor (cvf) are shown in a "map-visualization". In addition it is shown how specific sub-clusters can be generated that might be of special interest. In another step it is demonstrated how single cases can be associated to the according clusters and sub-clusters. Finally, additional possibilities of analysis with SEN are shortly discussed. Therefore, the aim of this study is twofold, namely to demonstrate how useful the combination of a SEN network with the usage of suited cvfs might be.

II. THE SELF-ENFORCING NETWORK (SEN)

The SEN is a new type of self-organized learning neural networks or unsupervised learning networks respectively, developed by our Research Group "Computer Based Analysis of Social Complexity" (CoBASC). “Self-organized learning” means that the network does not get any explicitly given learning goal but has to structure the input given to it according to an internal learning logic. The SEN is rather simple and comfortable to handle for a user; in addition its results are rather easy to understand for laymen not trained in neural networks. (cf. e.g. [21]).

The chief function of the SEN is the ordering or classifying respectively of data sets, i.e. objects with certain attributes. Hence each SEN operates on a database consisting of such objects and attributes. Usually these data are represented in a "semantical matrix": The rows of the matrix represent the objects and the columns the according attributes; the values of the matrix are the "degree of affiliation" of the attributes to the objects. In this case the values of the semantical matrix are the real data imported from xlsx- or csv-files, using the min-max normalization, accordingly adjusted for the SEN with the interval [-1.0 – 1,0]:

$$v_{norm} = \frac{v_{raw} - r_{min}}{r_{max} - r_{min}} * (n_{max} - n_{min}) + n_{min}. \tag{1}$$

The topology of SEN is dependent on the specific problem. It can be one-layered, if the semantical matrix is defined as a square matrix and if only objects that shall be classified are represented. The connections define if there is a feed-forward, feed-back, or, in the case that all objects are connected, a recurrent topology. All neurons might be input- and output neurons, depending on the external activation [21].

In the most cases, i.e. if the semantical matrix is defined by objects, attributes, and the connections between attributes and objects a SEN can be understood as a two-layered network by considering the attributes as input neurons and the objects as according output neurons. Again, depending on the distribution of the values in the semantical matrix, the network has a corresponding topology: If there are just connections between the attributes and the objects SEN has a feed-forward topology; if the objects are additionally connected with the attributes there is a feed-back topology, and if all neurons are connected then obviously SEN has a recurrent topology.

As in each neural network the dynamics of a SEN is generated by so-called activation functions. A user of a SEN can choose between different activation functions. In all cases a_j is the activation value of the receiving neuron j, a_i are the activation values of the sending neurons i, and w_{ij} as usual are the according weight values:

a) linear function

$$a_j = \sum w_{ij} * a_i, \tag{2}$$

b) tangens hyperbolicus (hyperbolic tangent)

$$a_j = \tanh(net_j) = \frac{e^{(net_j)} - e^{(-net_j)}}{e^{(net_j)} + e^{(-net_j)}}, \tag{3}$$

and

c) the logistic function

$$a_j = \frac{1}{1 + e^{-net_j}}. \tag{4}$$

In addition the user can select three functions, developed by our research group:

d) the linear-mean value function (LMF),

$$a_j = \sum \frac{w_{ij} * a_i}{k}, \tag{5}$$

with k = number of connections

e) the so-called logarithmic-linear function (LLF),

$$a_j = \sum \begin{cases} lg_3(a_i + 1) * w_{ij,} & if\ a_j \geq 0 \\ lg_3(|a_i + 1|) * -w_{ij}, & \text{else.} \end{cases} \tag{6}$$

One can interpret the use of the logarithm as a dampening factor that is "internal" to the function in contrast to "external" factors like, e.g., scale or decay in interactive networks. The basis 3 of the logarithm was chosen simply because basis 2 would generate too small activation values and basis 4 too large values. The function d) was also constructed for obtaining dampening effects. [22].

f) The enforcing activation function (EAF).

$$a_j = \sum_{i=1}^{n} \frac{w_{ij} * a_i}{1 + |w_{ij} * a_i|} \tag{7}$$

It depends on the specific problem, which activation function is best suited; in this study the linear function was used.

The operations of a SEN start by analyzing the values v_{sm} of the semantical matrix and by transforming the values of the semantical matrix into the weight matrix of the network. The weight matrix, hence, is generated from the semantical matrix. If an object o does not have the attribute a and hence the according semantical value $v_{oa} = 0$, then the weight value $w_{oa} = 0$ and remains so; in all other cases the weight value w_{oa} is

$$w_{oa} = c * v_{oa} \tag{8}$$

c is a constant usually defined as $0 \leq c \leq 1$. It has the same function as the well-known learning rate in standard neural networks.

The learning rule of a SEN that varies the values of the weight matrix according to the problem is:

$$\begin{gathered} w(t+1) = w(t) + \Delta w\ and \\ \Delta w = c * w_{oa}. \end{gathered} \tag{9}$$

If more learning steps are necessary, i.e. if SEN has not reached an attractor, then

$$\begin{gathered} w(t+1) = w(t) + c, \\ if\ w(t) \neq 0, \end{gathered} \tag{10}$$

if w(t) = 0 then w(t + 1) = 0 for all learning steps.

In most cases, according to numerous experiences, it is sufficient to use c = 0.1. To be sure, the weight values between two different objects a and b and two different attributes x and y usually are

$$w_{ab} = w_{xy} = 0. \tag{11}$$

A. The cue validity factor (cvf)

The cue validity is a measure how important certain attributes are for membership in a given category [23; 24].

For example, the attribute of four legs is not very typical for the category "dog" because a lot of other animals also have four legs. In contrast the attribute "barking" is very typical for a dog as no other animals make such noises.

By using cvf-values it is possible to distinguish between the importance of an attribute for the analysis. If the value of the cvf = 1 or higher, the attribute is most important; if cvf = 0 than the attribute is not considered for clustering.

Then Equation (8) becomes

$$\Delta w = c * v_{oa} * cvf_a. \tag{12}$$

To put it into a nutshell, a SEN learning process consists of a) the transformation of the semantical matrix into a weight matrix according to equation (9), and b) the learning runs, i.e. the enforcing of the weight values, according to (9) or (12) respectively. The learning process, i.e. the assignment of a new object, is finished when a point attractor has been reached. The result of this learning process is given by the end activation values of those neurons that represent the specific objects.

This assignment of a new object is done by the comparison of the attribute values of the new object with the end activation vectors of the objects that are already part of the network.

B. Visualizations

The results of a SEN system can be seen by a user in different ways that allow a fast interpretation:

a) Tabular results, consisting of the end activation values of the object neurons; in this case the meaning of the end activation values is that the higher the values are the more relevant are the objects, according to the purposes that are to be analyzed.

b) A map visualization, representing the approximated similarity between the objects.[1]

Because this visualization is most important in this article, it should be described in more detail: Two entities should attract each other when their distance in the coordinate system is larger than the Euclidean distance of the vectors these entities represent. Likewise two entities should repel each other when their distance is smaller than the distance of their vectors.

A so-called error vector $\vec{e_i}$ of the entity i is defined as follows:

$$\vec{e_i} = \sum_{k=0}^{n} (\vec{p_i} - \vec{p_k}) \times \frac{(|\vec{v_i} - \vec{v_k}| - |\vec{p_i} - \vec{p_k}|)}{|\vec{p_i} - \vec{p_k}|}, \tag{13}$$

where n is the number of entities in the coordinate system, p is the entity's position in the coordinate system and v is the actual vector the entity represents. As the fraction's denominator is 0 whenever $\vec{p_i}$ and $\vec{p_k}$ are equal, entities do not interact with themselves or other entities at the same coordinates.

In the second step all entities get shifted according to their error vector $\vec{e_i}$. At timestep t an entities position is defined as:

$$\vec{p_i}(t) = \vec{p_i}(t-1) + \frac{\vec{e_i}}{2+n}. \tag{14}$$

Whenever the current step led to a *stable* state, i.e. no entity was shifted more than a certain threshold[2] the number of entities in the coordinate system is doubled by removing n entities from the waiting queue, and positioning them in the coordinate system, where n again is the number of entities already presented there.

The algorithm ends as soon as it reaches a stable state and the waiting queue is empty.

After the learning process is finished, a user can insert a so-called *input vector* containing the different attributes, in the context of this article, new medical data.

The following SEN-visualizations are only activated if a user inserts a new input vector; these forms of representing the results are chiefly important for a (practical) user.

c) SEN-Visualization in form of a "center modus":

This modus operates the following way: When a user inserts a new case (e.g. counter ID) into the SEN system, the center modus places the inserted vector in the center of the visualization plane – hence the name of the modus. The objects, i.e. the trained counter IDs, are placed in the beginning at the periphery of the plane. Subsequently the objects are attracted to the center in dependency of the end activation values. In the end the user gets the visual information that those objects are the more probable candidates for the inserted IDs the nearer they are placed to the center. In other words, this visualization modus computes the geometrical distance of the different objects to the inserted vector as center. This is done by computing a distance r of an object to the input-vector:

$$r = \begin{cases} 1 - a_{rel}, & \text{if } a_{rel} \geq 0 \\ \dfrac{1 - tanh(a_{rel})}{2}, & \text{else.} \end{cases} \tag{15}$$

The relative activation a_{rel} is defined as:

$$a_{rel} = \frac{a}{a_{max}}. \tag{16}$$

One can interpret this modus as the transformation of semantical relations into geometrical ones.

d) Ranking and Distance after an input

The classification of the objects, in particular by comparison with one or several reference types, is measured twofold: by computing the distance between the final activation values of the respective attributes that characterize the objects, and measuring the Euclidian distance between each trained objects and the new object which shall be classified. In other

[1] The algorithm for the map visualization was developed by Björn Zurmaar, a PhD student of our research group.

[2] This threshold is currently 5% of the Euclidean Distance of the two most different vectors.

words, in the first case the final activation values are ordered as a vector and the according "distance" to other attribute vectors is measured by the difference of the *highest activation* values to those of the reference types, and in the second case the *smallest difference* is taken into account. The values of the end activations are additionally shown in form of beams whose lengths represent the size of the values.

III. Selected Data

The selected data are already analyzed with the 'multivariable logistic regression' by [25]. In this article the focus of the study was the importance of the measurement of HbA1c (Glycated hemoglobin) using 70.000 inpatient diabetes encounters. The dataset includes 10 years (1999-2008) of clinical care (1999-2008) at 130 US hospitals. The encounters contain information following the criteria a) of a hospital admission, b) diabetes, in any kind, was entered to the system, c) the length of stay was between 1 and 14 days, d) laboratory tests were performed and e) medications were administered.[3]

For the purposes of this article, only the first 1.000 of 101.767 data were chosen to analyze how the data could be in principal clustered with different cvf-values. The goal of this study was to see, if there can be any clusters observed, without knowing too much about the data; in addition the selection of this data-set is to enable a reproducible analysis.

Because the dataset has numeric and nominal values, the data have to be transposed only in numerical values. Table 1 shows the list of the 19 selected attributes from the 55 in the data set, and the encoding for the data.

TABLE I. List of attributes and their code

Attributes	Codes
Race	1 - 5
Gender	1 = Femal; 2 = Male
Age	Original data
Medical specialty	1 - 38
Admission type	1 - 6
Discharge disposition id	1 - 25
Admission source id	1 – 20
Time in hospital	Original data
Number of lab procedures	Original data
Number of procedures	Original data
Number medications	Original data
Diagnosis 1	Original data and coded (e.g. V = 23) ICD-9
Diagnosis 2	Original data and coded ICD-9
Diagnosis 3	Original data and coded ICD-9
Number diagnoses	Original data

Insulin dosage	No insulin = 0;
	Up = 1; Steady = 2; Down = 3

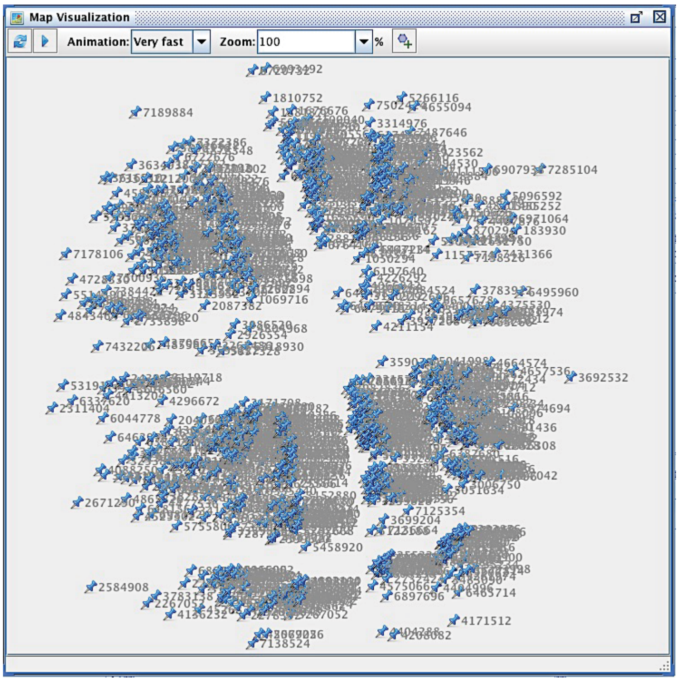
Diabetes Medications	0 = No; 1 = Yes
Readmitted	0 = No; 29 = readmitted in less than 30 days; 30 = readmitted in more than 30 days

HbA1c-Test	0 = Not measured;
	1 = >8 (the result was greater than 8% [25]);
	2 = >7 < 8;
	3 = normal

The objects are the encounter IDs; the patient number and other values were not taken into account for this analysis because it was only interesting to see how the data are clustered when all cvf-values are equal, in this case 1, and with different values accordingly. Because the most of the features are not relevant for the following study, the codes are not further defined. It is important to mention that the diseases in the data set are coded according to ICD-9.

IV. Results

In the first learning process all attributes have a cvf = 1; the settings for the activation function is linear and the learning rate 0.1; this remains constant for all learning processes. The result is shown in Fig. 1:



Map Visualization

Animation: Very fast

Zoom: 100 %

Fig. 1. The result of 1.000 data sets with a cvf-value 1.0 for all attributes

Fig. 1 shows that different clusters are being formed by SEN with some data belonging more or less to a cluster. Because the data are heterogeneous and complex they may have something in common like gender and / or race. Having a first approximation of the data, a user has different options at his disposal.

On a random basis one now has the opportunity to decide which features are of interest; in order to do this one can change the cvf-values accordingly. This is still done with the data from Fig. 1.

In consequence, in the next learning process, and to prove the influence of cvf-values, only the features for 'gender' have the cvf-value 1 (left side of Fig. 2) and subsequently 'gender', 'readmission', and 'medical specialty' (right side of Fig. 2), all the others always 0:

[3] http://archive.ics.uci.edu/ml/datasets/Diabetes+130US+hospitals+for+years+1999-2008 or http://archive.ics.uci.edu/ml/machine-learning-databases/00296/)

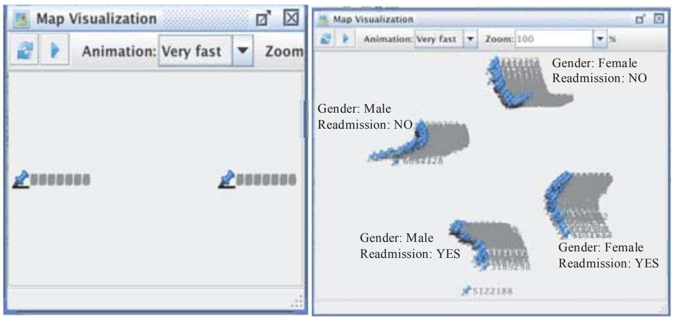


Fig. 2. Result of the SEN. On the left side the results with a cvf only for gender, on the right side for gender, medical specialty and readmission, else 0.

The results show that the influences of the cvf are clear. In the first case, no differentiation is given; the data are only clustered according to the gender. In the second case, only four clusters remain clear divided by gender and readmission according to the four combinatory possibilities, but there is a hint that the data are not homogeneous. Hence, it is obviously possible to steer the clustering according to the interesting question by using a suited cvf; the next examples demonstrate according steering effects.

The resulting question was now, if there are differences in readmissions in dependency of the medical specialty, diabetics medications and insulin. The cvf-values are accordingly: cvf-values = 1 for gender, medical specialty, insulin, diabetics medication, readmission, and HAb1c test; all others from Table I (= 13) have the value 0. The result is shown in Fig. 3:

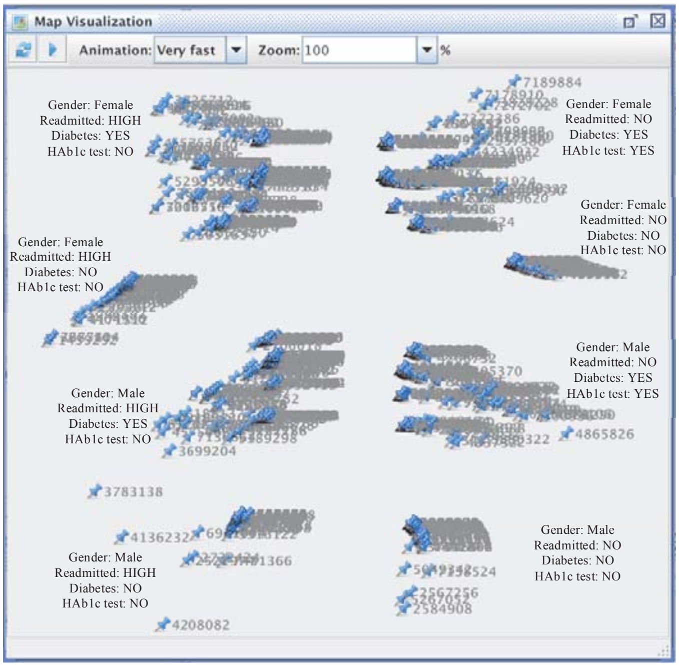


Fig. 3. The result of SEN

The result is more differentiated and there is an indication, that the number of readmitted patients with diabetics is higher than without diabetes and the reasons of the admissions in hospital are much more different.

A. Selection of sub-clusters

To prove this assumption, only one sub-cluster was selected, which is possible in the SEN tool, and the data within the cluster are exported as a csv-file. Fig. 4 and Table II show the selected cluster and the results in the csv-file:

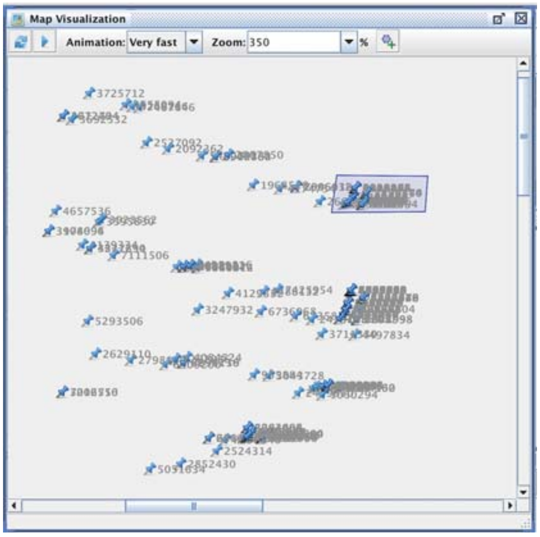
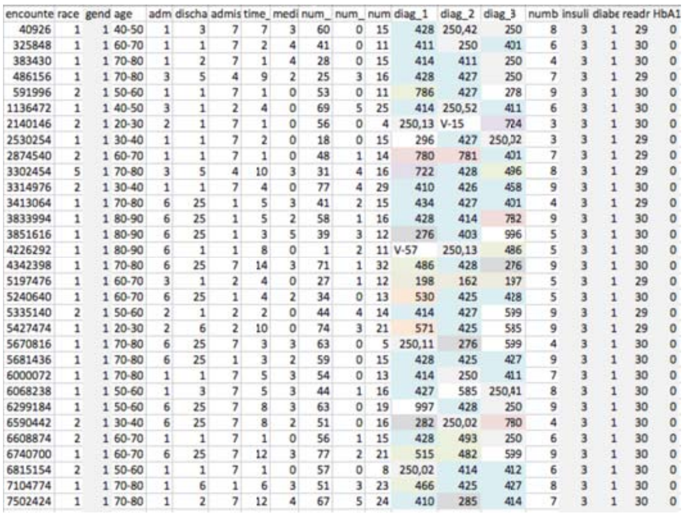


Fig. 4. The selected sub-cluster (zoom-option) of the cluster with the gender female, high readmitted and with diabetics.

The cvs-file contains the following information:

TABLE II. DATA IN SELECTED CLUSTER: 31 CASES IN TOTAL



encounte	race	gend	age	adm	discha	admis	time_	medi	num_	num_	num	diag_1	diag_2	diag_3	numb	insuli	diabe	readr	HbA1
40926	1	1	40-50	1	3	7	7	3	60	0	15	428	250,42	250	8	3	1	29	0
325848	1	1	60-70	1	1	7	2	4	41	0	11	411	250	401	6	3	1	30	0
383430	1	1	70-80	1	2	7	1	4	28	0	15	414	411	250	4	3	1	30	0
486156	1	1	70-80	3	5	4	9	2	25	3	16	428	427	250	7	3	1	29	0
591996	2	1	50-60	1	1	7	1	0	53	0	11	786	427	278	9	3	1	30	0
1136472	1	1	40-50	3	1	2	4	0	69	5	25	414	250,52	411	6	3	1	30	0
2140146	2	1	20-30	2	1	7	1	0	56	0	4	250,13	V-15	724	3	3	1	30	0
2530254	1	1	30-40	1	1	7	2	0	18	0	15	296	427	250,02	3	3	1	29	0
2874540	2	1	60-70	1	1	7	1	0	48	1	14	780	781	401	7	3	1	29	0
3302454	5	1	70-80	3	5	4	10	3	31	4	16	722	428	496	8	3	1	29	0
3314976	2	1	30-40	1	1	7	4	0	77	4	29	410	426	458	9	3	1	30	0
3413064	1	1	70-80	6	25	1	5	3	41	2	15	434	427	401	4	3	1	29	0
3833994	1	1	80-90	6	25	1	5	2	58	1	16	428	414	782	9	3	1	30	0
3851616	1	1	80-90	6	25	1	3	5	39	3	12	276	403	996	5	3	1	30	0
4226292	1	1	80-90	6	1	1	8	0	1	2	11	V-57	250,13	486	5	3	1	30	0
4342398	1	1	70-80	6	25	7	14	3	71	1	32	486	428	276	9	3	1	30	0
5197476	1	1	60-70	3	1	2	4	0	27	1	12	198	162	197	5	3	1	29	0
5240640	1	1	60-70	6	25	1	4	2	34	0	13	530	425	428	5	3	1	30	0
5335140	2	1	50-60	2	1	2	2	0	44	4	14	414	427	599	9	3	1	29	0
5427474	1	1	20-30	2	6	2	10	0	74	3	21	571	425	585	9	3	1	29	0
5670816	1	1	70-80	6	25	7	3	3	63	0	5	250,11	276	599	4	3	1	30	0
5681436	1	1	70-80	6	25	1	3	2	59	0	15	428	425	427	9	3	1	30	0
6000072	1	1	70-80	1	1	7	5	3	54	0	13	414	250	411	7	3	1	30	0
6068238	1	1	50-60	1	3	7	5	3	44	1	16	427	585	250,41	8	3	1	30	0
6299184	1	1	50-60	6	25	7	8	3	63	0	19	997	428	250	9	3	1	30	0
6590442	2	1	30-40	6	25	7	8	2	51	0	16	282	250,02	780	4	3	1	30	0
6608874	2	1	60-70	1	1	7	1	0	56	1	15	428	493	250	6	3	1	30	0
6740700	1	1	60-70	6	25	7	12	3	77	2	21	515	482	599	9	3	1	30	0
6815154	2	1	50-60	1	1	7	1	0	57	0	8	250,02	414	412	6	3	1	30	0
7104774	1	1	70-80	1	6	1	6	3	51	3	23	466	425	427	8	3	1	30	0
7502424	1	1	70-80	1	2	7	12	4	67	5	24	410	285	414	7	3	1	30	0

The in grey highlighted columns have indeed in common the gender (female), the dosage of insulin was decreased (3), all patients have a diabetes medication, a high readmission rate, and the fact that no HAb1c test was made. In addition they have often a disease of the circulatory system (ICD-9: 390–459, 785 [25] – blue color) or a disease of the respiratory system (ICD-9: 460–519, 786 – green color), to name only two.

The next cluster to be analyzed is that on the right side of Fig. 4, gender female, not readmitted, and no diabetes medications:

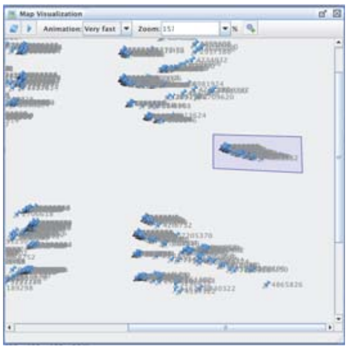
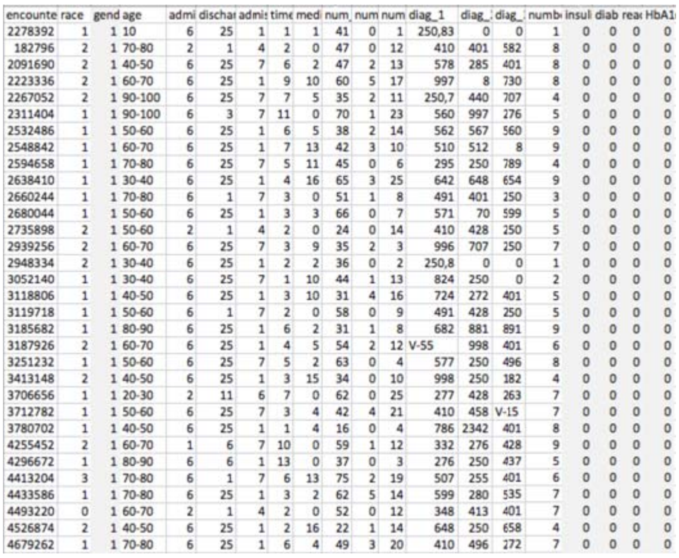


Fig. 5. Analysis of the cluster on the right side

The results are shown in Table III:

TABLE III. EXERPT OF THE DATA IN THE CLUSTER



encounte	race	gend	age	admi	dischar	admi	time	medi	num_	num	num	diag_1	diag_	diag_	numb	insul	diab	rea	HbA1
2278392	1	1	10	6	25	1	1	1	41	0	1	250,83	0	0	1	0	0	0	0
182796	2	1	70-80	2	1	4	2	0	47	0	12	410	401	582	8	0	0	0	0
2091690	2	1	40-50	6	25	7	6	2	47	2	13	578	285	401	8	0	0	0	0
2223336	2	1	60-70	6	25	1	9	10	60	5	17	997	8	730	8	0	0	0	0
2267052	2	1	90-100	6	25	7	7	5	35	2	11	250,7	440	707	4	0	0	0	0
2311404	1	1	90-100	6	3	7	11	0	70	1	23	560	997	276	5	0	0	0	0
2532486	1	1	50-60	6	25	1	6	5	38	2	14	562	567	560	9	0	0	0	0
2548842	1	1	60-70	6	25	1	7	13	42	3	10	510	512	8	9	0	0	0	0
2594658	1	1	70-80	6	25	7	5	11	45	0	6	295	250	789	4	0	0	0	0
2638410	1	1	30-40	6	25	1	4	16	65	3	25	642	648	654	9	0	0	0	0
2660244	1	1	70-80	6	1	7	3	0	51	1	8	491	401	250	3	0	0	0	0
2680044	1	1	50-60	6	25	1	3	3	66	0	7	571	70	599	5	0	0	0	0
2735898	2	1	50-60	2	1	4	2	0	24	0	14	410	428	250	5	0	0	0	0
2939256	2	1	60-70	6	25	7	3	9	35	2	3	996	707	250	7	0	0	0	0
2948334	2	1	30-40	6	25	1	2	2	36	0	2	250,8	0	0	1	0	0	0	0
3052140	1	1	30-40	6	25	7	1	10	44	1	13	824	250	0	2	0	0	0	0
3118806	1	1	40-50	6	25	1	3	10	31	4	16	724	272	401	5	0	0	0	0
3119718	1	1	50-60	6	1	7	2	0	58	0	9	491	428	250	5	0	0	0	0
3185682	1	1	80-90	6	25	1	6	2	31	1	8	682	881	891	9	0	0	0	0
3187926	2	1	60-70	6	25	1	4	5	54	2	12	V-55	998	401	6	0	0	0	0
3251232	1	1	50-60	6	25	7	5	2	63	0	4	577	250	496	8	0	0	0	0
3413148	2	1	40-50	6	25	1	3	15	34	0	10	998	250	182	4	0	0	0	0
3706656	1	1	20-30	2	11	6	7	0	62	0	25	277	428	263	7	0	0	0	0
3712782	1	1	50-60	6	25	7	3	4	42	4	21	410	458	V-15	7	0	0	0	0
3780702	1	1	40-50	6	25	1	1	4	16	0	4	786	2342	401	8	0	0	0	0
4255452	2	1	60-70	1	6	7	10	0	59	1	12	332	276	428	9	0	0	0	0
4296672	1	1	80-90	6	6	1	13	0	37	0	3	276	250	437	5	0	0	0	0
4413204	3	1	70-80	6	1	7	6	13	75	2	19	507	255	401	6	0	0	0	0
4433586	1	1	70-80	6	25	1	3	2	62	5	14	599	280	535	7	0	0	0	0
4493220	0	1	60-70	2	1	4	2	0	52	0	12	348	413	401	7	0	0	0	0
4526874	2	1	40-50	6	25	1	2	16	22	1	14	648	250	658	4	0	0	0	0
4679262	1	1	70-80	6	25	1	6	4	49	3	20	410	496	272	7	0	0	0	0

In this cluster there are 51 cases and all have in common the gender (female) and the values in the last four columns.

B. *Insertion of single cases*

In the next step all sub-clusters, which are not of interest, were excluded from the study. In this case all the data containing patients without medications of diabetes or insulin are removed from the semantic matrix.

In addition new data i.e. new patients can be inserted as input vectors and check if they are placed near to the clusters or as outliners, which means, that the data do not belong to the trained data.

Hence Fig. 6 only shows patients with diabetes and three new input vectors.

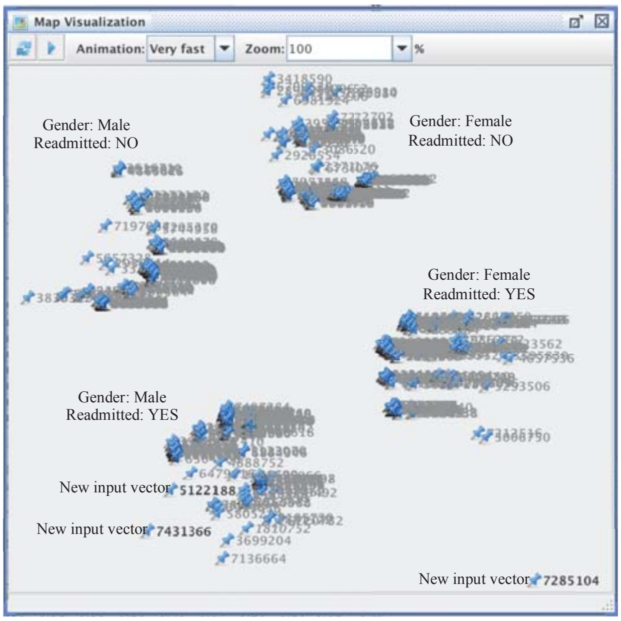


Fig. 6. Clustering of the IDs 5122188, 7431366 and 7285104 as input vectors.

In this case the gender male is on the left side; on the upper part are patients who are not readmitted, on the lower part with readmission.

The patients with the counter ID 5122188, and 743166 are inserted as a new input vector and are correct clustered to the patients who are high readmitted.

As an additional test, the counter ID 7285104 (female, high readmission, HAb1-test normal, no diabetes; appears as an outliner in the lower part on the right side of Fig. 6) was also inserted as a new input vector. In Fig. 7 the outliner is shown in a zoom-visualization.

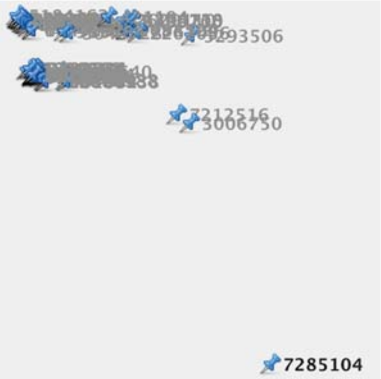
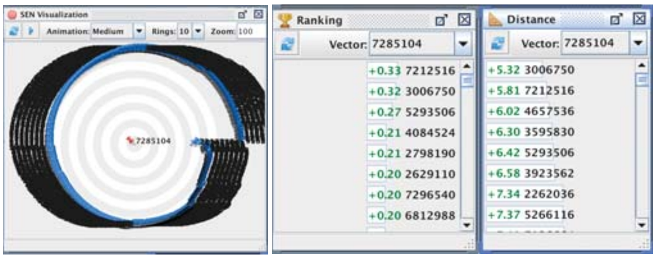


Fig. 7. Zoom-visualization of the placement

This counter ID as input vector allows the additional visualization in SEN, the computing of the ranking and distances, shown in Fig. 8:



Animation: Medium Rings: 10 Zoom: 100

Vector: 7285104

+0.33 7212516
+0.32 3006750
+0.27 5293506
+0.21 4084524
+0.21 2798190
+0.20 2629110
+0.20 7296540
+0.20 6812988

Vector: 7285104

+5.32 3006750
+5.81 7212516
+6.02 4657536
+6.30 3595830
+6.42 5293506
+6.58 3923562
+7.34 2262036
+7.37 5266116

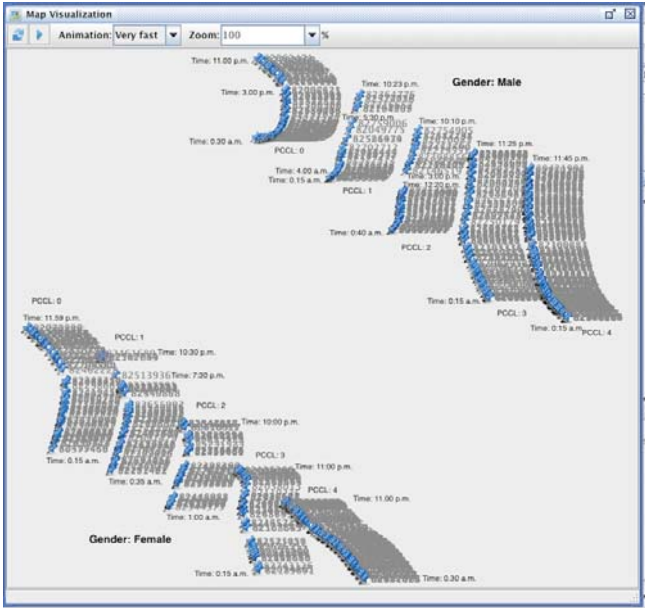
Fig. 8. Visualization of the result and the computed values in respect to the ranking and distance

The visualization algorithm shows that no one of the trained vectors are attracted to the center and the computed ranking, i.e. the strongest activated neuron, show a low activation level to the ID 7121516; the distance, meaning the lowest distance to a trained vector, has a high distance of 5.32 to the ID 3006750. Both patients are female, have a high readmission and the HAb1-test was normal; but both of them have diabetes and this is the most important difference to the new inserted vector.

V. Second Example

In this case other real clinical data, obtained from the University Hospital in Essen, are analyzed, which contain additional information as "Diagnosis-related group" (DRG), a system to classify hospital cases, "Patient Clinical Complexity Level" (PCCL) with the levels 0 – 4, describing no comorbidity and/or complication (0) to extremely severe comorbidity and/or complication (4), cost factors, and the record that the patients were falling during the hospital stay and at what time it happened. In total 800 records were analyzed.

The intuitive assumption was that patients are falling at night because of not wanting to disturb the nursing staff, or being disorientated in the strange environment. By analyzing the data, the SEN analysis shows a different picture with a cvf-value of 0.1 for gender, PCCL, and time of falling (Fig. 9):



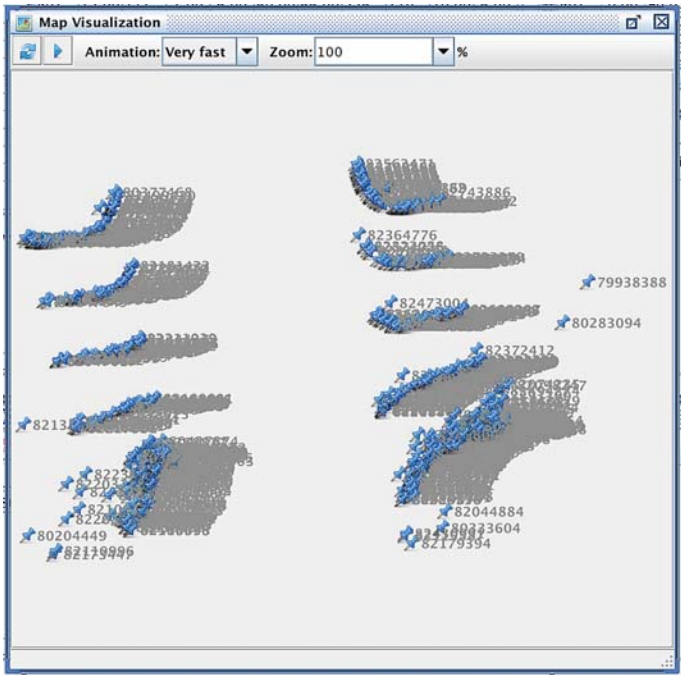
Animation: Very fast Zoom: 100 %

Fig. 9. Result with the cvf-values 0.1 for gender, PCCL, and falling time

The clusters are clear divided in female (left side) and male (right side), and according to the PCCL levels 0 - 4.

The results show that indeed, a lot of patients are falling during night; but especially patients with the PCCL level 4 are falling during the hole day, and, this is a little surprising, also patients with PCCL 0. The records for female patients with the level 1-3, and male patients with level 1-2 show some exceptions, which might be during breakfast or lunch time; in addition in the data set no records for falling are between 10.30 a.m. and 0.0 a.m. A closer study may show correlations between diseases, medication or other reasons for the falling cases; of particular importance are possible measures in order to prevent falling cases.

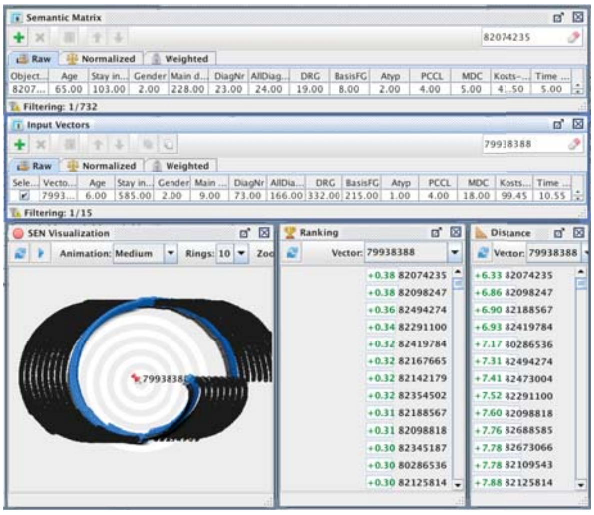
From an economic point of view, it is possible to add a cvf-value for the cost factor and see if there are differences in the clusters. The result shows some outliners (Fig. 10):



Animation: Very fast Zoom: 100 %

Fig. 10. Result with cvf-values of 0.1 for gender, PCCL, falling time and cost factors.

The differentiation according to the PCCL level remains; in this case outliners are of interest. The vectors representing the outliners can be selected and removed from the semantical matrix and transferred as "input vectors". This enables again the other visualization algorithms to compare the vectors. Fig. 11 shows the result:



Semantic Matrix

82074235

Object...	Age	Stay in...	Gender	Main d...	DiagNr	AllDiag...	DRG	BasisFG	Atyp	PCCL	MDC	Kosts-...	Time ...
8207...	65.00	103.00	2.00	228.00	23.00	24.00	19.00	8.00	2.00	4.00	5.00	41.50	5.00

Filtering: 1/732

Input Vectors

79938388

Sele...	Vecto...	Age	Stay in...	Gender	Main ...	DiagNr	AllDia...	DRG	BasisFG	Atyp	PCCL	MDC	Kosts...	Time ...
☑	7993...	6.00	585.00	2.00	9.00	73.00	166.00	332.00	215.00	1.00	4.00	18.00	99.45	10.55

Filtering: 1/15

SEN Visualization

Animation: Medium
Rings: 10

Ranking

Vector: 79938388

+0.38 82074235
+0.38 82098247
+0.36 82494274
+0.34 82291100
+0.32 82419784
+0.32 82167665
+0.32 82142179
+0.32 82354502
+0.31 82188567
+0.31 82098818
+0.30 82345187
+0.30 80286536
+0.30 82125814

Distance

Vector: 79938388

+6.33 82074235
+6.86 82098247
+6.90 82188567
+6.93 82419784
+7.17 80286536
+7.31 82494274
+7.41 82473004
+7.52 82291100
+7.60 82098818
+7.76 82688585
+7.78 82673066
+7.78 82109543
+7.88 82125814

Fig. 11. Comparison of the input vector ID 79938388 with the ID 82074235 in the semantic matrix, which appears as the first computed result in the ranking and distance.

The ranking and distance visualization have both as result the ID 82074235 with an activation of 0.38 (ranking) and distance of 6.33, meaning that the next similar vector has more differences than similarities; for example: the age 6 vs. 65, stay in hospital 585 days vs. 103, costs factors 99.45 vs. 41.50. The only similarities are the gender male and the PCCL level 4. The other outliner can be analyzed accordingly.

When SEN has generated specific clusters of interest, new patient data can be inserted as input vector and it can be analyzed, if they belong to a known cluster or not. A user can decide if the new data should be transferred into the semantic matrix for further analysis.

VI. CONCLUSION AND FURTHER WORK

This prototype already shows the potential of a SEN for clustering medical data. Using for all features the same cvf-factor a first approximation of different clusters can help to make a decision, which sub-clusters should be analyzed in more detail or which features are of more interest. The demonstrated combination of the SEN algorithm and the usage of suited cvfs shows the possibilities for the important task of clustering clinical data.

Especially the selection and export of the data in sub-clusters enable a pre-selection of the data for additional algorithms, e.g. for statistical ones.

Using a tool for this demonstration means that only a limited number of the data can be analyzed in an acceptable time. Because of the promising approach, the SEN should be reprogrammed according to the newest developments in fastening the processing data.

REFERENCES

[1] A.K. Jain, “Data clustering: 50 years beyond K-means, Pattern Recognition Letters”, Vol. 31, 2010, pp. 651-666.

[2] C.C. Aggarwal, C.K. Reddy, Data clustering: algorithms and applications, 2013, Boca Raton, London, CRC Press Taylor & Francis Group

[3] M. E. Celebi, Ed., Partitional Clustering Algorithms, 2015, Cham, Heidelberg, Springer International Publishing Switzerland.

[4] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S.S. Iyengar, “Computational Health Informatics in the Big Data Age: A Survey”, ACM Computing Surveys, Vol. 49: 1. 2016, pp. 1-36.

[5] Y. Cui, J. Shi and Z. Wang, “Lazy Quantum clustering induced radial basis function networks (LQC-RBFN) with effective centers selection and radii determination”, Neurocomputing 175 2016, pp. 797-807.

[6] P. H. Thong and L. H. Son, “An Overview of Semi-Supervised Fuzzy Clustering Algorithms”, International Journal of Engineering and Technology, Vol. 8: 4, 2016, pp. 301-306.

[7] D. G. Ferrari, L. Nunes de Castro, “Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods”, Information Sciences 3012015, pp. 181-194.

[8] T. Cerquitelli, S. Chiusano and X. Xiao, “Exploiting clustering algorithms in a multiple-level fashion: A comparative study in the medical care scenario, Expert Systems with Applications”, Vol. 55, 2016, pp. 297-312

[9] R. N. Kumar and M. A. Kumar, “Medical Data Mining Techniques for Health Care Systems”, International Journal on Computer Science and Engineering, Vol. 02, Nr. 02, 2010, pp. 250-255.

[10] L. Pamulaparty, C.V. Rao and M. S. Rao, Cluster Analysis of Medical Research Data using R, Global Journal of Computer Science and Technology: C Software & Data Engineering, Vol. 16 Issue 1 Version 1, 2016, pp.16-22.

[11] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain”, International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, 2016, pp. 53-60.

[12] S. Ding, F. Wu, J., Qian, H., Jia and F. Jin, “Research on data stream clustering algorithms”, Artificial Intelligence Review 43: 593, 2015.

[13] R. Ding, Q. Wang, Y. Dang, Q. Fu, H. Zhang and D. Zhang, “YADING: Fast Clustering of Large-Scale Time Series Data”, Proceedings of the VLDB Endowment, Vol. 8: 5 2015, pp. 473-484.

[14] T. Richardson and E. Winer, “Extending parallelization of the self-organizing map by combining data and network partitioned methods”, Advances in Engineering Software 88, 2015, pp. 1-7.

[15] E. Lughofer and M. Sayed-Mouchaweh, “Autonomous data stream clustering implementing split-and-merge concepts – Towards a plug-and-play approach”, Information Sciences 304, 2015, pp. 54-79.

[16] A. Rauber, “LabelSOM: On the Labeling of Self-Organizing Maps”, Neural Networks, IJCNN '99. International Joint Conference on, Vol. 5, 1999, pp. 3527-3532.

[17] C. Klüver, and J. Klüver, Self-Organization and Adaptation in Socio-Cognitive Systems: A Computational Model. In: Tianfield, H., 2009 (Ed.): International Transactions on Systems Science and Applications, Vol. 5: 4, 2009, pp. 357-368.

[18] P. Sarlin, “A weighted SOM for classifying data with instance-varying importance”, International Journal of Machine Learning & Cybernetics, 2013, DOI 10.1007/s13042-013-0175-3.

[19] D.-L. Lia, M. Prasadb, C.-T. Linc and J.-Y. Changd, “Self-Adjusting Feature Maps Network and its Applications”, Neurocomputing, 2016, DOI 10.1016/j.neucom.2016.03.067.

[20] Y. Michea, A. Akusokb, D. Veganzonesc, K.-M. Björkd, E. Séverinc, P. du Jardine, M. Termenong, and A. Lendasseb, “SOM-ELM—Self-Organized Clustering using ELM, Neurocomputing”, Vol. 165, 2015, pp. 238-254.

[21] C. Klüver and J. Klüver, Self-organized Learning by Self-Enforcing Networks. In: I. Rojas, G. Joya and J. Cabestany, editos, proceedings of the 12th international work-conference on artificial neural networks (IWANN 2012), Part I Lecture Notes in Computer Science, 2013, 7902, Springer, pp. 518–529.

[22] J.L. McClelland and D.E. Rummerhart, “An Interactive Activation Model of Context Effects in Letter Perception: Part 1. An Account of Basic Findings. Psychological Review”, Vol. 88: 5, 1981, pp. 375-407.

[23] E. Rosch, C.B. Mervis, “Family Resemblances: Studies in the Internal Structure of Categories”, Cognitive Psychology, Vol. **7**: 4, 1975, pp. 573-605

[24] J. Klüver and C. Klüver, Social Understanding, On Hermeneutics, Geometrical Models, and Artificial Intelligence. 2011, Dordrecht (NL): Springer.

[25] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, Impact of HbA1c “Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records”, 2014, BioMed Research International Vol. 2014, DOI 10.1155/2014/78167.