Cluster Analysis of Patients' Clinical Information for Medical Practitioners and Insurance Companies

https://doi.org/10.3991/ijoe.v16i04.13119

Qurban A. Memon (✉), Mohammad E. Hassan
UAE University, Al Ain, UAE
qurban.memon@uaeu.ac.ae

Abstract—A number of approaches have been proposed in literature to collect and classify patient related information for purpose of better clinical diagnosis for safer treatment and administration of related activities. This type of data collection and classification benefits doctors and the corresponding hospitals. However, no effort is made, as to our knowledge, to classify accumulated data within insurance company databases to facilitate doctors as well as insurance companies for better analysis and cost-effective treatment of patients suffering from chronic (and expensive to treat) diseases such as related to oncology. In this study, a customized self-organized data classification model is applied to an insurance company database to build clusters based on age, patient condition, tests done, etc. These clusters provide integrated analysis to doctors in providing patient-specific, disease-specific, etc., and cost-effective treatment. On the other side, it saves on costs to be incurred on repeated tests to be done on the patient. An experimental setup is developed to train such a network, and testing results are presented. The practical constraints are also discussed.

Keywords—Clustering; Clinical Information; Data Classification

1 Introduction

A number of improvements in hospitals have led to implementing automated ordering and dispensing systems [1] resulting in more time in patient care activities, performance improvement, savings, etc. But, data collected within hospitals remains focused typically on billing or administrative information. Several types of products have applications for in-patient pharmacy departments. Real-time information technology (IT) systems include (1) automated prescribed order-entry systems, (2) clinical decision-support system platforms including intelligent systems to guide treatment and check orders, (4) automated patient records integrated with data from other patient related departments [2-3]. From a business point-of-view, clinical data security is important to maintain edge over competing institutions, since organizations invest huge money to research and development units to develop new drugs, medical devices and medical treatment procedure [4]. The corresponding results are stored in clinical trial files and records.

Patients' records typically vary in a multitude of ways, some of which include diagnosis, severity of illness, medical complications, and the speed of recovery, resource consumption, lab tests, discharge destination, and social circumstances. Such data is classified for various purposes, mostly within hospital domain. For example, in [5], the authors study medical data classification approaches applied to heart disease cases, and employ decision tree algorithms to collect results. In a similar work, the authors [6] discuss an end-to-end dynamic neural network that examines medical records, updates previous medical history, and then infers illness states and predicts future medical states. In another research [7], the authors propose an approach that combines rule-based features and knowledge-guided learning models for effective disease classification. The classification of such data may also be used for the purpose of analysis, planning, decision making, etc., and this area is active research subject in many disciplines, such as neural networks. The most important and frequently used method in classification, where no information is available about clusters, is Self-Organizing Map (SOM).

The objectives under this study are:

i) To guide doctors on chronic patient cases, as well as insurance companies to improve on future professional links with hospitals and doctors
ii) To facilitate doctors in analysis on patient clinical information from different hospitals
iii) To benefit insurance companies on savings for cost-effective treatment.

Typically, insurance investigation of patient cases, with special reference to oncology involves intensive, long term and phase wise collection of patient data. This helps insurance companies to determine the type and cost of tests, diagnostics, and treatment conducted in the hospital per patient per physician per hospital. Indirectly, this may help insurance companies to protect their interests. The system reported in this study is based upon a neural network and is intended to be used by insurance companies as well doctors. Effectively, the solution classifies a set of standard data into a number of classes taken from an insurance company database. The paper is structured as follows. In section 2, we present proposed approach for data classification as applied to oncology patient database. Section 3 presents experimental setup, network training and testing results, followed by conclusions in section 4.

2 Proposed Approach

When input data is fed to neural network in unsupervised mode, the Euclidean distance or the straight-line distance between the nodes is computed. Unsupervised learning is a class of machine learning techniques to find the patterns in data. In unsupervised learning, as the actual data moves through neural network, the weights of the links between the nodes start to look more like data as iterations continues. The data given to unsupervised algorithm is not labelled, which means only the input variables (x) are provided with no corresponding output variables. The resulting output grid map does not have target vectors, since their purpose is to divide the input vectors into clusters

based on similarity. Generally, more the nodes in grid map, more detailed the clustering is but requires more time for training.

The neural network trains itself to see patterns in the data much the way a human see. In this unsupervised mode, each neuron is fully connected to all the source units in the input layer. The number of input nodes equals the dimension of input vector in the network. The number of output nodes, typically set by the user, determines the maximum number of classes to be found. Each neuron (node) in the output layer represents a cluster, or alternatively a set of common features. Nearby nodes represent similar clusters and the network is trying to associate input patterns with common features to the same (or nearby) output node. The node in the network that is most similar to the input data is called the best matching unit. The neurons become selectively tuned to various input patterns during learning. The locations of the neurons are so tuned that the winning neurons become ordered and a meaningful coordinate system for the input features is created on the lattice.

The most commonly used form of unsupervised learning is self-organizing map (SOM) or self-organizing feature map (SOFM), also known as Kohonen Network [8-9], as shown in Figure 1. The main advantage of using a SOM is that the data is easily interpreted and understood. The drop of dimensionality and grid clustering makes it easy to detect similarities in the data.

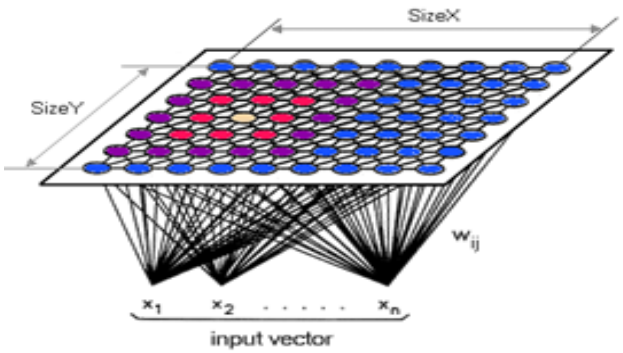


Fig. 1. SOM for input data and clusters with weights w_{ij}

Self-organizing maps differ from other artificial neural networks as they apply computationally convenient competitive learning approach using a neighborhood function, in order to preserve the topological properties of the input space. The nodes in the resulting map may be arranged on a hexagonal grid, and since a format (map ratio) is taken into account, the number of nodes in the actual map may be slightly different than specified. For the purpose of competitive learning, a well-known SOM-Ward distance measure may be used [10]:

$$d'_{xy} = \begin{bmatrix} d_{xy} & if\ clusters\ x\ and\ y\ are\ adjacent\ in\ the\ SOM \\ \infty & otherwise \end{bmatrix} \quad (1)$$

where

$$d_{xy} = \frac{n_x \,.\, n_y}{n_x + \, n_y} \,.\, \left\| \bar{x}_x - \bar{x}_y \right\|^2$$

where x and y denote two specific clusters, n_x and n_y denote the number of data points in the two clusters, $\bar{x}_x$ and $\bar{x}_y$ denote the centers of gravity of the clusters; and ||.|| is the Euclidean norm. Thus, the SOM-Ward distance observes the topological location of the clusters. In particular, two clusters that are not adjacent in the SOM are never considered to be merged.

3 Experimental Results

In this section, experimental setup and results are presented based on a typical database consisting of 800 case studies related to oncology in Al-Ain, UAE. Each case study specifies a data vector as an input (for example patient condition (ten different diseases), test type (total six different tests), patient origin (local or expat (from seven different countries)), doctor (local or expat (six from different countries)), patient age (four age ranges), patient insurance coverage (five different types), and hospital name (six different names)). The training set is a database consisting of these case studies. Each case study specifies an input vector.

3.1 Setup and training

A set of parameters are to be defined. The input vector consists of seven inputs:

1) X_1: Patient Condition
2) X_2: Test Type
3) X_3: Patient Origin
4) X_4: Doctor Origin
5) X_5: Patient Age
6) X_6: Patient Insurance
7) X_7: Hospital Name

Since the size of input vector is seven (7), a grid map of roughly 10x10 clusters seems sufficient to represent data. To enter the above inputs to any neural network and train it, the possibilities of each input must be represented by a continuous numeric value between 0.0 and 1.0, so that the input vector entered to the network at the end is consisting of only numerical values. For this purpose, the range between 0.0 and 1.0 is divided into equidistance values according to the number of possibilities of that input. The Table 1 shows the inputs with their different possibilities and the corresponding continuous numerical value for each possible input.

The topology is a 10x10 grid, so there are 100 neurons. Using Matlab, input vectors are randomly generated, and resulting topology is shown in Figure 2. The Figure 2 shows that the maximum number of hits associated with any neuron is 15. Thus, there

are 15 input vectors in that cluster. We set the learning parameters to 0.1 and initial number of epochs to 200. For training, 70% (560 samples) of data ware selected to train the network for 200 iteration. The first training result is the self-organizing clustering map, as shown in Figure 3a. The rows of the Figure 3b represent the clusters. The first column contains the cluster name; the second column displays the descriptions (if any) of the cluster; the third column displays the median; the fourth column displays the frequency, etc. Subsequent columns display the aggregated attribute values. The cell value is the aggregated value. The map shows that our input data are classified in four clusters, with each cluster containing a number of neurons.

Table 1. Input values

no.	Feature	Values	Numerical Rep.
1	patient_condition	disease 1	0.05
		disease 2	0.15
		disease 3	0.25
		disease 4	0.35
		disease 5	0.45
		disease 6	0.55
		disease 7	0.65
		disease 8	0.75
		disease 9	0.85
		disease 10	0.95
2	test_type	test 1	0.1
		test 2	0.26
		test 3	0.42
		test 4	0.58
		test 5	0.74
		test 6	0.9
3	patient_origin	local	0.12
		country 1	0.23
		country 2	0.34
		country 3	0.45
		country 4	0.56
		country 5	0.67
		country 6	0.78
		country 7	0.89
4	doctor_origin	local	0.02
		country 1	0.212
		country 2	0.404
		country 3	0.596
		country 4	0.788
		country 5	0.98
5	patient_age	range 1	0.15
		range 2	0.4
		range 3	0.65
		range 4	0.9
6	insurance_coverage	category 1	0.04
		category 2	0.26
		category 3	0.48
		category 4	0.7
		category 5	0.92
7	hospital	hospital 1	0.06
		hospital 2	0.24
		hospital 3	0.42
		hospital 4	0.6
		hospital 5	0.78
		hospital 6	0.96

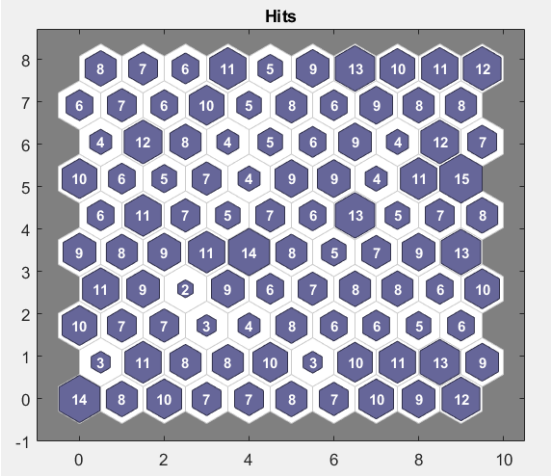
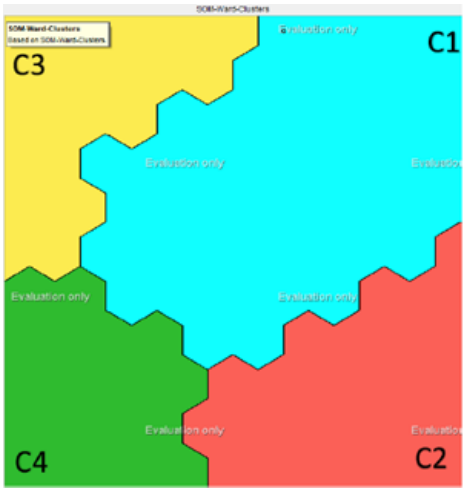


Fig. 2. Initial Clusters during setup

At the end of training, the centers of all 100 neurons are determined. To gain insight, some of the clusters are shown in Table 2, for first row arranged from top left corner with coordinates 0, 0 moving to right side 8, 0.



(a) First phase of Training

Cluster	Description	Abs. Profile Median	Frequency	Condition	H Name	Test	Age	D Origin	P Origin	Insurance
C 1		0.2764	38.00%	0.454	0.480	0.637	0.595	0.380	0.496	0.573
C 2		0.3104	23.00%	0.487	0.345	0.434	0.648	0.782	0.625	0.675
C 3		0.6133	19.50%	0.815	0.772	0.559	0.607	0.533	0.414	0.609
C 4		0.4960	19.50%	0.458	0.810	0.611	0.590	0.799	0.754	0.504

(b) Clustering Map and cluster centers

Fig. 3. First Phase Clusters

Table 2. Cluster centers

x	y	Condition	Test	P_Origin	D_Origin	Age	Insurance	H_Name
0	0	0.91490	0.86644	0.44364	0.28893	0.47628	0.41778	0.77721
1	0	0.85702	0.53566	0.62529	0.38278	0.46602	0.31325	0.74812
2	0	0.85000	0.29406	0.61798	0.41811	0.32827	0.57848	0.72761
3	0	0.80354	0.36176	0.35177	0.65337	0.29267	0.74996	0.72647
4	0	0.54495	0.38602	0.32654	0.60629	0.31267	0.82746	0.83581
5	0	0.27449	0.53071	0.48810	0.49065	0.40620	0.88294	0.82585
6	0	0.34257	0.66847	0.33562	0.41815	0.31287	0.87088	0.48367
7	0	0.40783	0.49608	0.29074	0.49309	0.30446	0.83557	0.24404
8	0	0.74744	0.64393	0.26758	0.42342	0.36694	0.80611	0.24295

3.2 Testing and results

To test the network, 30% (240) data was used to evaluate the network. Four vectors belonging to four different clusters were selected, and later extreme noise is added to the inputs. The Table 3 shows coordinates X, Y of centers of four clusters (C1, C2, C3, C4). Next, extreme noise (0.0) is added to condition input (Table 4). The network is still giving good response (75%) for all clusters except for C3 where the data vector is moved to another cluster. This response is expected since C3 is dependent on high values of "Condition" input and noise forced this input to be zero.

In another test, extreme noise (1.0) is added to condition input (as shown in Table 5). The network is still giving accurate response for all clusters 100%. It is observed from the testing that clustering error usually happens when the noise shifts the input far from its original value or far in a dimension away from the cluster centre in that dimension. To analyze the performance of the network in another way, a noisy data is replaced by its mean value. For example, if we replaced the noisy input in test 2 (Table 4) by its mean value, the network works perfectly (Table 6). In order to enhance the performance of neurons in the network, the number of neurons were increased from 100 to 1000. This resulted as shown in Figure 4. The clusters are increased to 5 instead of 4 and the data is now uniformly distributed among the clusters. Each cluster contains 1/5 of the data set. Reducing the learning rate will also enhance the network performance as the Figure 5 shows that the clusters are more uniform.

Table 3. Clean data vectors

Clean Data Vectors							Nearest Neuron		Cluster
Condition	*Test*	*P-origin*	*D-Origin*	*Age*	*Insurance*	*H_Name*	*X*	*Y*	
0.60	0.83	0.29	0.50	0.75	1.00	0.17	9	1	C1
0.20	0.50	0.86	0.67	0.25	0.60	0.17	8	8	C2
0.90	0.33	0.14	0.67	0.25	0.20	0.50	2	1	C3
0.70	0.17	1.00	0.50	1.00	0.80	0.83	0	7	C4

Table 4. Test of adding extreme noise (0.0) to "Condition" input

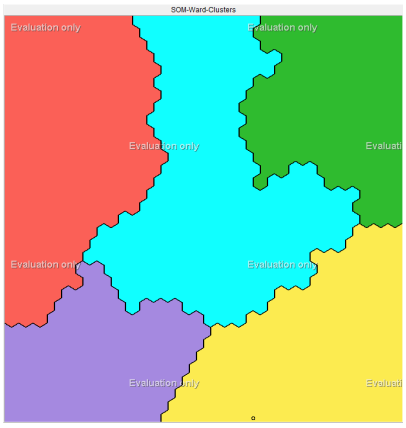
Noisy Data Vectors							Nearest Neuron		Cluster
Condition	***Test***	***P-origin***	***D-Origin***	***Age***	***Insurance***	***H_Name***	***X***	***Y***	
0.0	0.83	0.29	0.50	0.75	1.00	0.17	7	1	C1
0.0	0.50	0.86	0.67	0.25	0.60	0.17	8	8	C2
0.0	0.33	0.14	0.67	0.25	0.20	0.50	4	3	C1
0.0	0.17	1.00	0.50	1.00	0.80	0.83	0	8	C4

Table 5. Test of adding extreme noise (1.0) to "Condition" input

Noisy Data Vectors							Nearest Neuron		Cluster
Condition	***Test***	***P-origin***	***D-Origin***	***Age***	***Insurance***	***H_Name***	***X***	***Y***	
1.0	0.83	0.29	0.50	0.75	1.00	0.17	9	1	C1
1.0	0.50	0.86	0.67	0.25	0.60	0.17	9	5	C2
1.0	0.33	0.14	0.67	0.25	0.20	0.50	2	1	C3
1.0	0.17	1.00	0.50	1.00	0.80	0.83	0	7	C4

Table 6. Result of testing the network with the mean value of noisy input

Noisy Data Vectors							Nearest Neuron		Cluster
Condition	***Test***	***P-origin***	***D-Origin***	***Age***	***Insurance***	***H_Name***	***X***	***Y***	
0.53	0.83	0.29	0.50	0.75	1.00	0.17	9	1	C1
0.53	0.50	0.86	0.67	0.25	0.60	0.17	8	8	C2
0.53	0.33	0.14	0.67	0.25	0.20	0.50	3	1	C3
0.53	0.17	1.00	0.50	1.00	0.80	0.83	0	7	C4



Cluster	Description	Abs. Profil...	Frequency	Condition	P Origin	Age	Insurance	H Name	Test	D Origin
C 1		0.4020	22.63%	0.412	0.467	0.703	0.382	0.564	0.576	0.472
C 2		0.2717	24.75%	0.773	0.475	0.653	0.610	0.763	0.530	0.515
C 3		0.4057	20.75%	0.473	0.612	0.643	0.696	0.334	0.449	0.818
C 4		0.5426	17.50%	0.543	0.611	0.457	0.737	0.383	0.656	0.352
C 5		0.5355	14.38%	0.383	0.717	0.520	0.550	0.816	0.698	0.823

Fig. 4. Resulting clusters when number of neurons is increased to 1000

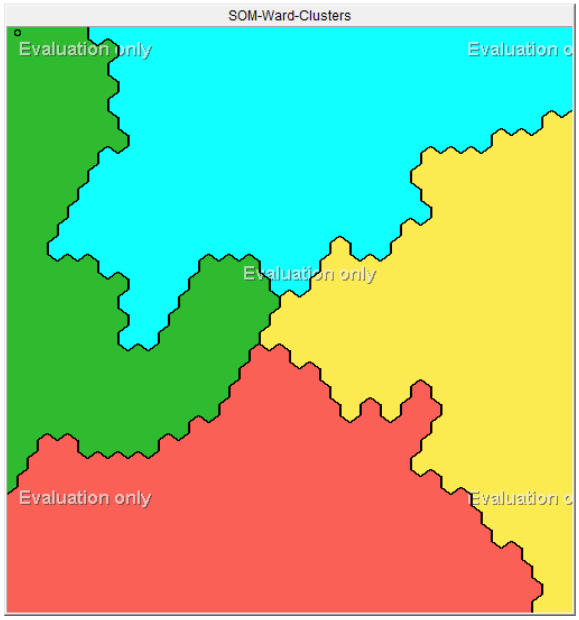


Fig. 5. Network map with 1000 neurons and lower learning rates

4 Conclusion

As a re-focus on objectives, an analysis platform to enable integrated and wider view on patients clinical information retrieved from insurance company database, was presented to facilitate doctors for in-depth medical analysis, and for insurance companies to benefit cost-effectiveness. The platform is not limited to oncology, though other clinical databases may be linked to the same or separate analysis for each database may be developed. Based on our experimental results, it is recommended that:

- The number of clusters (or neurons) should be increased to fine tune the clustering map
- The number of iterations should be at least 500 times number of neurons in the lattice
- Learning rate parameter be selected as a small constant, typically 0.01, and decreased during thousands of iterations, but never goes to zero
- Mostly used metric in SOM is the Euclidean distance, which is not the best to some problems, so a different metric be investigated to produce competitive results for data classification

Furthermore, as initial positions of neurons differ each time the SOM analysis is run, the eventual SOM map generated will also differ. This can be alleviated by initially setting the input vector to its maximum size and increasing the number of neurons before training for smoother classification.

This platform however poses certain legal challenges. Typically, the clinical information is only viewable within one hospital, and only written and authorized medical reports/records can be taken to other hospitals. For insurance company data to be inter-hospital use, a regulation is needed amongst hospitals (linked to same insurance company) and the government department(s) to protect patient data privacy.

5 References

[1] Laura T. Pizzi; et al., "Clinical Information Management Systems: An Emerging Data Technology for Inpatient Pharmacies," *American Journal of Health-System Pharmacy*, Vol. 61(1), 2004 https://doi.org/10.1093/ajhp/61.1.76

[2] Memon, Q., "Smarter Health-Care Collaborative Network," *Building Next-Generation Converged Networks: Theory and Practice*, 451-476, 2013. https://doi.org/10.1201/b14574-23

[3] Q. Memon, A. Mustafa, Exploring mobile health in a private online social network," *International Journal of Electronic Healthcare*, Vol. 8, No. 1, 2015, pp. 51-75.

[4] Md. Islam,Tahmina N., Yu-Chuan J., "Recent Advancement of Clinical Information Systems: Opportunities and Challenges," *Yearbook of Medical Informatics*, Vol. 27(1), 83–90, 2018

[5] Tzung-I T., et al. "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis," *IEMS*, Vol. 4, No. 1, pp. 102-108, June 2005

[6] T. Pham, T.Tran, D. Phung, S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *Journal of Biomedical Informatics*, Vol. 69, May 2017, pp. 218-229. https://doi.org/10.1016/j.jbi.2017.04.001

[7] L. Yao, C. Mao & Y. Luo, "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," *BMC Medical Informatics and Decision Making*, Vol.19, Article number: 71 (2019). https://doi.org/10.1186/s12911-019-0781-4

[8] Seng Poh L., and Habibollah H., "Applying Kohonen Network in Organising Unstructured Data for Talus Bone," 3rd *International Conference on Theoretical and Mathematical Foundations of Computer Science, Lecture Notes in Information Technology*, Vol.38, 2012

[9] Noda, S., Tim R., Whitney B. "Applications of artificial neural networks in health care organizational decision-making: A scoping review," *PLoS One*. 2019; 14(2): e0212356 https://doi.org/10.1371/journal.pone.0212356

[10] Nurettin Y., Ilker U., Halil A., "Using Self-Organizing Neural Network Map Combined with Ward's Clustering Algorithm for Visualization of Students' Cognitive Structural Models about Aliveness Concept," *Computational Intelligence and Neuroscience*, 2016, https://doi.org/10.1155/2016/2476256

6 Authors

Qurban A. Memon has contributed at levels of teaching, research, and community service in the area of electrical and computer engineering. He graduated from University of Central Florida, Orlando, US with PhD degree in 1996. Currently, he is working as Associate Professor at UAE University, College of Engineering, United Arab Emirates. He has authored/co-authored over ninety publications in his academic career. He

has executed research grants and development projects in the area of intelligent based systems; security and networks. He has served as a reviewer of many international journals and conferences; as well as session chair at various conferences.

Mohammad E. Hassan completed his Master's degree in Electrical Engineering in 2018 from UAE University, UAE. Later, he joined UAE University as Ph.D. student in Electrical Engineering. He plans to graduate by end of 2021. His research interests are in imaging and electrical systems.

Article submitted 2020-01-10. Resubmitted 2020-02-18. Final acceptance 2020-02-18. Final version published as submitted by the authors.