AUTOMATED CLASSIFICATION OF LAB BLOOD-TEST USING SELF-ORGANIZING MAPS

Basim J. Zafar
Electrical Engineering Department, Umm Al-Qura University, 21955 Makkah, Saudi Arabia
Corresponding author : bjzafar@uqu.edu.sa

ABSTRACT: *Automated classification and diagnosing of disease is very important for rapid proofing of patients which is used for all general epidemiological and many health management purposes. In addition to that, the adoption of the International Classification of Diseases (ICD) provided the basis for the compilation of national mortality and morbidity statistics, these records also enable the storage and retrieval of diagnostic information for clinical and epidemiological purposes. ICD is used to classify diseases and other health problems recorded on many areas of health and vital records including death certificates and hospital records. Achieving a correct ICD or concluding a correct patient diagnosis represents a challenge to the medical doctors because it demands sound experience that might not be possessed by many physician, especially in third-world countries. Classifying ICD according to blood-lab test profiles facilitates an accurate and fast patient diagnosis. In this study, a Self-Organizing Map (SOM) technique is used to classify ICD according to the blood-lab test profiles to a limited applicable set. SOM forms a non-linear mapping of data into a two-dimensional grid map that can be used as an exploratory analysis tool for generating hypotheses on the change as well as the characteristics of the profile of blood-lab tests. Similarity relationships within the ICD and blood-lab test are visualized and interpreted. The methodology and results of the analysis are presented in this paper.*

Keywords: ICD, self-organizing maps, lab blood-test, classification

1. INTRODUCTION

The International Classification of Diseases (ICD) has become the international standard of diagnostic classification for all general epidemiological and many health management purposes [1]. It is used to classify diseases and other health problems recorded on many areas of health and vital records including death certificates and hospital records. In addition to providing the basis for the compilation of national mortality and morbidity statistics, these records also enable the storage and retrieval of diagnostic information for clinical and epidemiological purposes [2]. Achieving a correct ICD represents a challenge to the medical practitioners. Moreover, concluding a correct patient diagnosis demands sound experience that might not be possessed by many physician, especially in third-world countries. While physicians are highly trained individuals, they rely on their educational background, training, accumulated knowledge base, experience and memory to make complex treatment decisions for the patients [3].

Computer based ICD classification according to blood-lab test profiles facilitates an accurate and fast patient diagnosis [3]. Computer based diagnostic tools are known as clinical decision support systems (CDSSs), these are computer programs with intelligent tools that accumulate and store human expertise and provide it back as second opinion or advice to physicians when making a clinical decisions [4]. Such solution will be very beneficial to assist an unexperienced physicians and provide a second opinion thought to them about the patient case and provide them with information they needed to know based on the diagnostic test results given to the system as inputs. Despite this simple definition given to CDSSs, they can vary greatly in function, type and targeted usage. Some CDSSs are designed for outpatient by given reminders and alerts while others are directed to assist physicians can be fully connected with the hospital/clinic medical record system.

There are also types of CDSSs that are used for inpatient care which can be connected with laboratory tests and provide alerts and suggestion to practitioners. However, it is important not to generalize the usage of CDSSs as their effectiveness is hindered by their design, functions and use case they were trained for [5].

In this paper we developed a CDSS system for disease classification based on lab blood-test. Specifically this system return the percentage of the diseases most properly meet the given blood-test profile. To achieve this, a Self-Organizing Map (SOM) technique is used to produce the ICD according to the lab blood-test profiles of a limited applicable set of ICDs. SOM forms a non-linear mapping of data into a two-dimensional grid map that can be used as an exploratory analysis tool for generating hypotheses on the change as well as the characteristics of the profile of blood-tests [6]. Similarity relationships within the ICD and blood-test are visualized and interpreted.

Figure 1 shows flow diagram for the proposed disease classification method. Initially the blood test data are preprocessed to remove noise and normalize it so that different tests are in the same scale. These data are fed to the self-organizing map network which is a type of unsupervised neural network that perform clustering on the given data to achieve a unique ICD value [7]

The remaining of this paper is organized as follows; section

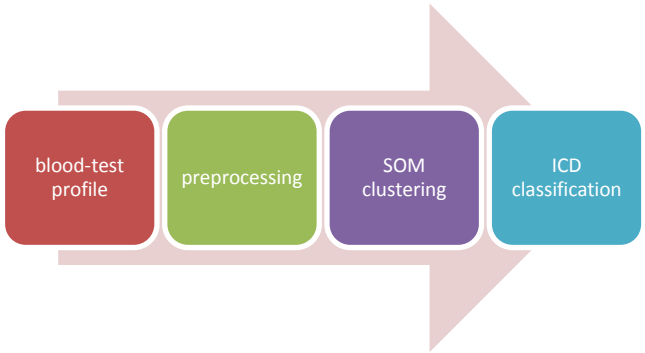

Figure. 1. Flow diagram for ICD predication using self-organizing maps (SOM) clustering method

(2) present some of the related works on CDSS and their applications in the medical field. Section (3) presents the self-organizing maps algorithm and the main types of SOM. Section (4) presents the medical data preprocessing which includes noise removal and data normalization. Section (5) present the classification results of blood-test using SOM. Finally section (6) concludes the paper by listing the main finding and future recommendations

2. RELATED WORKS

The growing interest in clinical decision support system (CDSS) for healthcare has led to many development different computer based method to assist physician for fast and accurate diagnosis. Berlin et al. [4] presented a taxonomy for computer-based clinical support system; they listed 74 decision support system published between 1997 and 2003 which spanned various clinical issues. There are two types of CDSS; patient directed DCS that provide preventive healthcare service to the general public using mail or phones which represented 38% of the surveyed system. The second category are inpatient CDSS which are used by physician and healthcare staff that provide recommendation to the physician about possible disease that the patient might have based on the test results of patient history using information priory encoded in the database.

Yan et al. [8] developed CDSS for heart disease diagnosis using multiple layer perceptron neural networks. They input basic patient information, patient's symptoms, inducement and disease history as well as physical signs and assisted examinations to the MLP neural network which then predict one of five possible heart diseases (hypertension, coronary heart disease, rheumatic valvular heart disease, chronic cor pulmonale and congenital heart disease) and give recommendations to the physicians to help them makes accurate clinical decisions. The authors claimed it achieve more than 90% diagnosing accuracy using a small dataset of 352 cases. Roshanov et al. [9] investigated how CDSS can improve practitioner's diagnostic-test ordering behaviors. They have argued that over/under-use of diagnostic testing have downstream implication on the patient outcomes and can lead to wrong interpretation of test results in addition to the financial implications of these tests. In this study it was found that CDSSs has improved the test ordering behavior of practitioner's in 55% of the case.

Anooj [10] used a weighted fuzzy rules based CDSS to diagnose heart diseases; they initially generated weighted fuzzy rules automatically which is used construct a fuzzy decision support systems. The experimental results has shown that the fuzzy logic system has higher prediction accuracy and sensitivity than neural network based prediction. Mago *et al.* [11] developed a fuzzy logic based CDSS for detail treatment and it was found that the fuzzy based CDSS showed more consistent treatment plans when compared with three professional dentists who has been given the same case independently. They recommended that this system to be used as expert second opinion to help dentist during the decision making process of treatment plan for a crack or broken tooth.

3. SELF-ORGANIZING MAPS

The Self-Organizing Map is an unsupervised Neural Network learning process, which forms a non-linear mapping of data to a two-dimensional map grid that can be used as an exploratory analysis tool [12]. SOM consists of neurons organized on a regular low-dimensional grid. Each neuron is represented by d-dimensional weight vector, m= [m_1 ….. m_d], where d equals to the dimension of the input vectors. The neurons are connected to adjacent neurons by a neighborhood relation, which dictates the structure of the map. SOM training algorithm updates the best matching unit and its topological neighbors on the map. The region around the best matching unit is pulled towards the presented training sample. At the end, the neurons will become orders on the output grid. SOM technique can be used for finding groups of similar data structures. The batch and sequential algorithm are the two main algorithm used in the training of the SOM. The batch algorithm is much faster than the sequential algorithm, and the results of the batch algorithm are typically just as good as or even better than the sequential algorithm [13].

3.1 Sequential SOM Algorithm

A sequential SOM algorithm is trained iteratively where the input data vectors are presented to the algorithm one at a time in a random order. For each input vector x, the BMU is determined as

$$\|x - m_b\| = \min_i\{\|x - m_i\|\} \quad (1)$$

Usually, Euclidean distance is used as a measure of the similarity. To calculate the distance between vectors with missing data, the algorithm excludes these values from the distance calculation. For that the missing value contribution to the calculation of the distance is zero. Because the same data are ignored in each distance calculation, this is a valid solution [14]. The distance measure after these changes can be written as

$$\|x - m_b\|^2 = \sum_{k \in K} (x_k - m_k)^2 \quad (2)$$

Where k is the set of known (not missing) variables of sample vector x. x_k and m_k are k_{th} components of the sample and weight vectors.

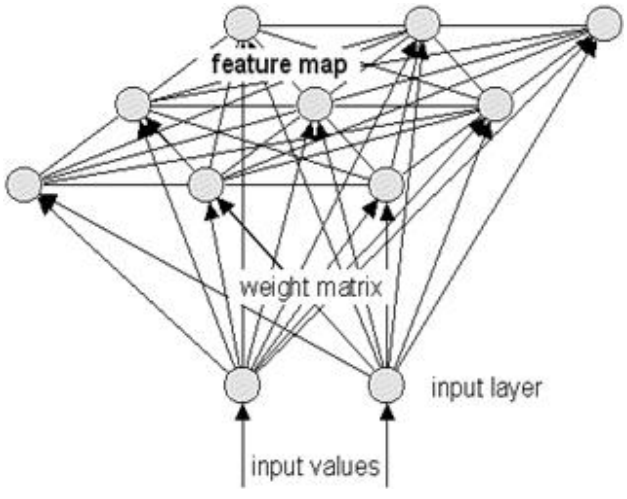

Figure 2: Illustration of self-organizing map showing input layer, the feature maps and their associated weight matrix[15]

In the next step, the prototype vectors are updated. The BMU and its topological neighbors are moved closer to the input vector in the input space. The update rule for the prototype vector of unit is

$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(t)[x - m_i(t)] \quad (3)$$

Where t denotes time, x(t) is an input vector drawn from the input data set at time t. $h_{ci}(t)$ is the neighborhood kernel around the winner unit c and α(t) is the learning rate. The Gaussian neighborhood function is mostly used in this type of problem. The Gaussian neighborhood function can be written as

$$h_{ci}(t) = \exp\left(-\frac{\|r_i - r_c\|^2}{2\sigma(t)^2}\right) \quad (4)$$

The SOM training is usually performed in two phases. In the first phase, a relatively large initial learning rate and neighborhood radius are used. In the second phase, both learning rate and neighborhood radius are small right from the beginning. This procedure first tunes the SOM approximately to the same space as the input data and then fine-tunes the map. The error function of the SOM [14] in the case of a discrete data set and fixed neighborhood function can be shown to be

$$E = \sum_{i=1}^{N}\sum_{j=1}^{M} h_{cj}\|x_i - m_j\|^2 \quad (5)$$

Where N is number of training samples, and M is the number of map units

3.2 Batch SOM Algorithm

At each step of the batch algorithm, all the input data vectors x are simultaneously used to update all the model vectors. In each training step, the data set is partitioned according to the Voronoi regions of the map weight vectors. The Voronoi region of a particular point P is that region on the plane where P is the closest point of all points in S, where S is a set of points in that plane. For points on a plane, the Voronoi regions will always be bounded by line segments positioned halfway between pairs of points [16]. The new weight vectors of the SOM are calculated after, then, as

$$m(t+1) = \frac{\sum_{j=1}^{N} h_{jc}(t)x_j}{\sum_{j=1}^{N} h_{jc}(t)} \quad (6)$$

Where c is the index of the BMU of the data sample of x_j. The new weight vector is a weighted average of the data samples, where the weight of each data sample is the neighborhood function value $h_{ic}(t)$ at its BMU. Missing values are simply ignored in calculating the weighted average. In the batch algorithm, the learning rate function used in the sequential algorithm is no longer used, but the width of the neighborhood is monotonically decreased during the learning.

4. DATA PROCESSING

The data used in this research consist of 50,000 profiles of patients medical records that consist of their blood-lab test results and their ICD. These data are provided by al-Noor specialist hospital in Makkah. Each profile consists of 15 lab blood-lab tests. The list of these lab blood-lab tests are represented in Table (1). Those blood-lab tests have been chosen because of their frequently use by medical doctors. The index of International Classification of Diseases is used to study the result of the SOM classification. Dealing with classification of blood-lab test results requires large amount of background work, namely: problem specification, preparation of the data, and interpretation of the results with a priori knowledge. Some remarks concerning practical aspects of medical data discussed below:

1) Missing data: It is not uncommon that the medical doctor does not order all the specified blood-lab test set and the values of this test will be considered as missing value. These missing values need to be avoided in any attempt to classify the data set. The simplest solution to overcome this problem, using SOM, is to use the available data vector value to find the best matching unit for each profile and exclude the missing values from the distance calculation. This has been demonstrated as a valid solution [13].

2) Noise: Some of the blood-lab test results data are corrupted by a measurement error. Since each model vector of the SOM is a weighted average of the data vectors in the map unit and the neighboring map units, error is reduced and does not usually constitute a major problem in the results of SOM clustering.

3) Normalization: The blood-test results have different ranges; each samples are normalized to range from [0 – 1] so that in input data have the same bases when it is fed to the SOM algorithm.

Table 1: Index of the blood-lab test

Code no.	Test
1	ALBUMIN
2	BILIRUBIN, DIRECT
3	BILIRUBIN, TOTAL
4	CALCIUM
5	CK/CPK
6	CREATININE
7	GLUCOSE (RBS)
8	LDH
9	MAGNESIUM
10	PHOSPHORUS
11	POTASSIUM—K
12	SGOT/AST
13	SGPT/ALT
14	SODIUM—NA
15	UREA

5. RESULTS AND ANALYSIS

It is well know that diagnosis of the patient is strongly related to his blood-lab test result. The result on this study show the percentage of the related ICD for each group of blood-lab test result classified by SOM. In this research a large data set of 50,000 medical records, each of which includes information on blood-lab test results and the diagnosis is used. After classifying the blood-lab test results data a 9 major groups are obtained. The data has been divided into 9 groups where Figure 3 shows the mean vector of each group of blood-lab test data. Each group has similar characteristics, so from these groups we can have statistical studies. Figure 3 shows the plot of the central vector of each group of the classification of SOM to all data set of blood-lab test results. The central vector shown in this figure is normalized in order to remove large variation between different tests. Classification to 9 groups has been shown here for illustration, nevertheless different number of clusters has been studied as it will be shown in later sections of this paper.

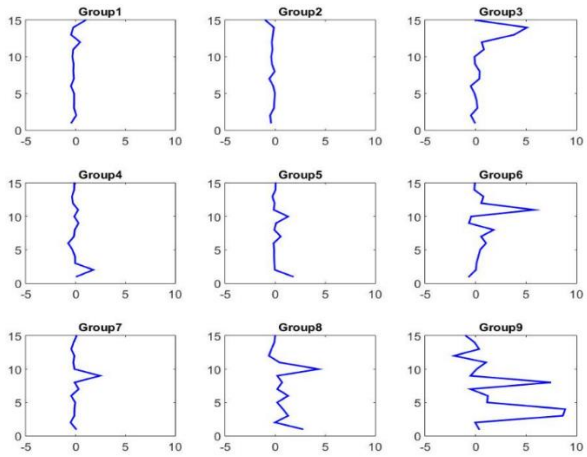

Figure 3. Plot of central vector of each vector after normalization

Figure 4 shows the percentage of input data in each one of the nine groups (clusters) obtained by the SOM algorithm. Some groups has very large percentage such as (Group1) which represent a majority in the test data as many samples falls in this category while other only represented 2.6% and 3% for Group8 and Group9 respectively. When taking larger number of clusters there is a possible some might be empty clusters.

Each one of the nine groups is associated with a list of possible disease with different weight factor for each disease (here is represented with ICD). Figure 5 shows percentage of ICDs found in group 8 of the clustered data and the weight factor of each ICD. This group is mostly associated with renal diseases which represented more than 60% of this group. The optimal approach is to take larger number of cluster so that each cluster has on disease of fewer disease with some common profile. However taking large number of clusters requires much large blood-test dataset and it might have larger clustering errors.

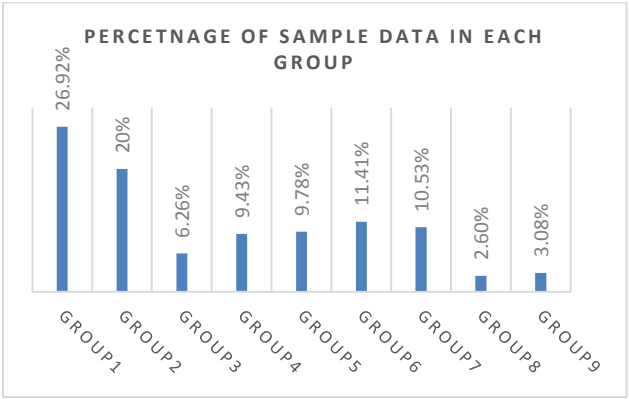

Figure 4. Distribution of samples among the selected groups (clusters)

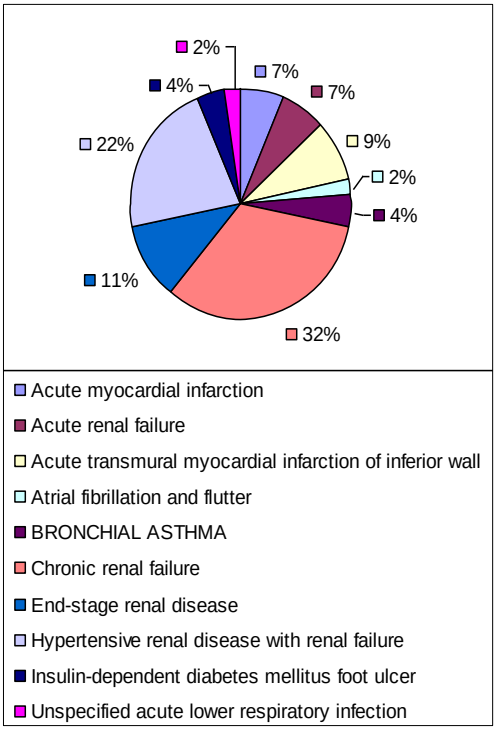

Figure 5. The percentage of ICD-10 in Group 8

Figure 6 shows percentage of ICDs in group 6 which is mostly related to newborn diseases; 20% is about respiratory diseases, 12% is low birth weight as well as transient tachypnea and 17% are other diseases.

Figure 7 shows the presence and distribution of the Gastroenteritis NOS in all clusters; this disease is mostly present in group 4 and group 9 (each weighted 35%) and group 6 only 13% has this ICD and group 1 has only 9%. In group 2 and group 5 this ICD represented only 4% while it is not present in other groups. It is possible to classify the blood-lab test result data into larger number of groups and study their ICDs may results in discovering new relation between the blood-lab test result and some associated ICDs.

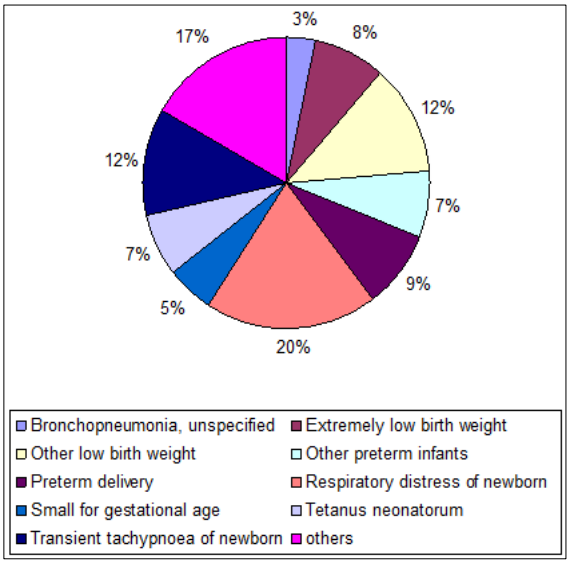

Figure 6. The percentage of ICD-10 in Group 6

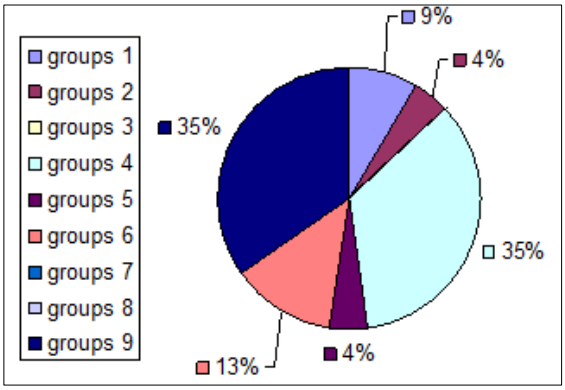

Figure 7. Percentage of acute renal gastroenteritis NOS in all groups

Similarly, Figure 8 show the distribution of Chronic ischaemic heart disease in all groups. 50% of this ICD is in group one 33% in group 7 and 17 percent in group 2.

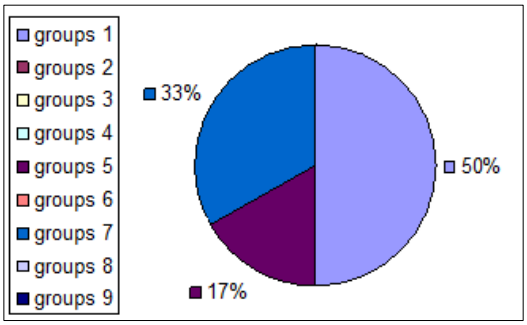

Figure 8. Chronic ischaemic heart disease

We have also studied the effect of number of clusters in the clustering accuracy by evaluating how accurately and independent test data can be clustered within using various number of cluster sets. To do this the data has been divided into two groups randomly; 70% has been used for clustering while the remaining 30% were used for testing and validation. The validation hypothesis assumes that a clustering is correct if the ICD of the sample is listed within the selected cluster (group) otherwise it is considered as incorrect clustering. This is quite a fair assumption given the fact that the number of clusters will always be less than the number of ICDs (diseases) recorded in the clustering data. Even though each cluster will have more than one ICD, we expect that relevant disease are grouped together because they are very likely to have similar blood-test profile. However it is also interesting to note that different diseases may have similar blood-test profile which can enlarge the physician understanding about the patient possible sickness to be consider when assessing their case and relevant additional testing can be recommended to disambiguate their status.

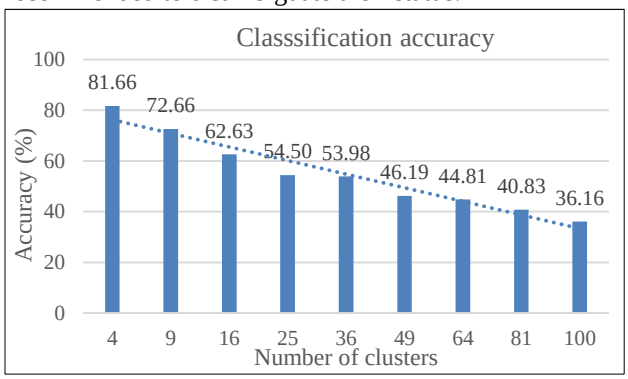

Figure 9. Classification accuracy for different number of clusters.

Figure 9 shows classification accuracy for different number of clusters; it is well understood that the accuracy is reduced with increasing number of clusters because the difference margin between these clusters is reduced when increasing the number of clusters. However the recoded accuracy for small number of clusters is high. At 4 clusters it scores almost 82% while at 9 clusters it score 73%. Even at 100 clusters where very small margin of difference is present, the clustering accuracy is more than 36%.

6. CONCLUSION

Computerized clinical decision support system can help practitioners to take correct decision when assessing patients by enlarging their knowledge and providing second thought and recommendations. This research have demonstrated the capability of self-organizing maps natural networks to classify the blood-test results in coherent groups of diseases with similar profiles and giving the percentage of their occurrences. This help the physician to make correct decision when diagnosing a patient based on these tests. Self-organizing maps neural networks can learn from examples and are flexible and powerful clustering tool. They are also very well suited for real time systems because of their fast response and computational times which are due to their parallel architecture. Using the classification of blood-lab test using SOM, we have determined percentage of occurrences of each ICDs. Each of these quantities is useful to large-scale studies of Classification of Diseases. The percentage of each ICDs for each blood-lab test results group can help the medical doctor to have a wider picture of the common

diagnoses. Certain diagnose is spread in only a few blood-lab test group.

7. ACKNOWLEDGEMENT

This research is supported by Al-Noor Specialist hospital in Makkah, Saudi Arabia

8. REFERENCES

[1] J. Stausberg, H. Lang, U. Obertacke, and F. Rauhut, *Journal of American Medical Informatics Association*, (2001).

[2] S Wu and T. Chow, "Clustering of the self-organizing maps using a cluster validity index based on inter-cluster and intra-cluster density", *Pattern Recognition* **37**(2), pp. 175-188 (2004).

[3] M. M. Haq, "Computer system for assisting a physician," *United States Patent 20020095313.*, (2002).

[4] A. Berlin, M. Sorani, and I. Sim, "A taxonomic description of computer-based clinical decision support systems," Journal of Biomedical Informatics., **39**(6), 656–667 (2006).

[5] I. Sim and A. Berlin, A framework for classifying decision support systems. Washington, DC: *Hanley and Belfus*, (2003).

[6] B. Zafar and V. Chandaraskar, "Adjustment of cross-track dependence of TRMM Precipitation Radar observation," *in Geoscience and Remote Sensing Symposium*, 3907–3909 (2004).

[7] World Health Organization, "The International Classification of Diseases (ICD), url: http://www.who.int/classifications/icd/en/ (last visited April 2016).

[8] H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, "A multiplayer perceptron-based medical decision support system for heart disease diagnosis," *Expert Systems Applications*, **30**, 272–281 (2006).

[9] P. S. Roshanov, J. J. You, J. Dhaliwal, D. Koff, J. a Mackay, L. Weise-Kelly, T. Navarro, N. L. Wilczynski, and R. Brian Haynes, "Can computerized clinical decision support systems improve practitioners' diagnostic test ordering behavior? A decision-maker-researcher partnership systematic review," *Implement. Sci.*, **6**(1), 88-93 (2011).

[10] P. Anooj, "Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules and decision tree rules," *Journal of King Saud University – Computer. Information. Sciences*, **24**, 27–40 (2012).

[11] V. K. Mago, N. Bhatia, A. Bhatia, and A. Mago, "Clinical decision support system for dental treatment," *Journal of Computer Science*, **3**(5), 254–261 (2012).

[12] T. Kohonen, Self-Organizing Maps, 3rd ed. *Springer*, (1997).

[13] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, **11**(3), 586–600 (2000).

[14] T. Samad and S. A. Harp, Self-Organization with Partial Data Network. *IOP Publishing*, (1992).

[15] T. Kohonen, "Self-organizing maps: Optimization approaches," *Proceedings of International Conference on Artificial neural Networks*, 981–990 (1991).

[16] T. Furukawa, S. Sonoh, K. Horio, and T. Yamakawa, "Batch learning algorithm of SOM with attractive and repulsive data," *in Proceedings of 5th Workshop on Self-Organizing Maps*, (2005).