Multidisciplinary Cancer Investigation **Original Article**

April 2017, Volume 1, Issue 2

Breast Cancer Risk Assessment Using Adaptive Neuro-Fuzzy Inference System (ANFIS) and Subtractive Clustering Algorithm

Alireza Atashi [1], Najmeh Nazeri [1,*], Ebrahim Abbasi [1], Sara Dorri [1], Mohsen Alijani-Z [1]

[1] *Informatics Department, Breast Cancer Research Center, Motamed Cancer Institute, ACECR, Tehran, Iran*

* *Corresponding author: Najmeh Nazeri, Informatics Department, Breast Cancer Research Center, Motamed Cancer Institute, ACECR, Tehran, Iran. E-mail: Najme.Nazeri@gmail.com*

DOI: 10.21859/mci-01029

Submitted: 15 July 2016
Revised: 22 November 2016
Accepted: 29 December 2016
ePublished: 8 March 2017

Keywords:
Breast Neoplasms
Decision Support Techniques
Cancer

Abstract

Introduction: The Adaptive neuro-fuzzy inference system (ANFIS) is a soft computing model based on neural network precision and fuzzy decision-making advantages, which can highly facilitate diagnostic modeling. In this study, this model was used for breast cancer detection.

Methods: A set of 1508 records of cancerous and non-cancerous participants' risk factors was employed for this study. First, the risk factors were classified into three priorities according to their importance level, were then fuzzified and the subtractive clustering method was used for their input with the same order. Randomly, the dataset was divided into two groups of 70% and 30% of the total records, and used for training and testing the new model, respectively. After the training, the system was separately tested with the Wisconsin and real clinical data, and the results were reported.

Results: The desired fuzzy functions were defined for the variables, and the model was trained with the combined dataset. Testing was conducted first with 30% of that dataset, and then with the real data obtained from a real clinical (BCRC) data, while the model's precision for the above stages was 81% (sensitivity = 85.1%, specificity = 74.5%) and 84.5% (sensitivity = 89.3%, specificity = 79.9%) respectively.

Conclusions: A final ANFIS model was developed and tested for two standard and real datasets on breast cancer. The resulting model could be employed with high precision for the BCRC Clinic's database, as well as conducting similar studies and re-evaluating other databases.

INTRODUCTION

Risk assessment for different diseases is crucial in medical decision-making, especially when numerous factors affect the appearance of that disease or disorder. It is thus necessary to establish a model for assessing and expressing the risk of that disease or disorder. Breast cancer is one such disease with various aspects, which inflicts huge expenses on the individual and the society, and is identified in the United States as the most common cancer type among women, and second most frequent cause of cancer mortality [1, 2]. Generally, breast cancer occurs in one out of every eight women, and kills one out of 36 of them [3].

Several models have been proposed for breast cancer risk assessment, including Gail, Claus and IBIS, each of which utilize a number of effective factors for breast cancer appearance in order to assess its risk. These models use crisp data, or only "statistical data", for their calculations. Due to the uncertainty of data regarding risk factors and the verbal expression of their impact, however, one could put soft data to higher use for assessments under fuzzy theories as well.

On the other hand, neural networks are increasingly employed in medicine, and several models for various medical purposes (such as prediction and prognosis models, or different classification models for diagnostic systems) have been established on their basis. The Adaptive neuro-fuzzy inference system (ANFIS) is one such model, and the findings point to better results of machine training techniques (such as artificial neural networks) than advanced statistical methods, including

nomograms based on cox-regression [4]. The most important aspect to consider about an ANFIS model is the fact that this combined model may utilize the precision of neural networks and fuzzy inference concurrently. Furthermore, the synergy of fuzzy logic and neural network eliminates the uncertainty of statistical analysis [4, 5]. Some previous studies confirm this fact for diagnosing breast cancer or other diseases. For instance, a study in 2012 used a multi-factor fuzzy system and assessed breast cancer risk factors to propose a model to determine its relative occurrence probability as high or low. That model and the available statistical data could help calculate an insurance premium for breast cancer, proportionate with its occurrence risk, which would be both reasonable for the client and profitable for the insurance company [5].

Furthermore, in a research on using ANFIS for automatic diagnosis of breast cancer through the Wisconsin Database records, the system was trained with the ANFIS classification criteria. The proposed ANFIS model employed nine breast cancer symptoms as input for automatic diagnosis and identification of the cancer, while at the same time merging the potentials of adaptive neural networks with the fuzzy logic qualitative approach. The findings of this research claimed ANFIS to possess the requisite capacity for detecting breast cancer [6]. A similar study adopted a new approach with ANFIS to diagnose erythematous squamous, and known cases were used to train ANFIS classifier on how to identify new cases. The proposed model combined the potentials of adaptive neural networks with the fuzzy logic qualitative approach, and achieved results through ANFIS analysis related to the effect of variables on diagnosing erythematous squamous. According to this study, the model displayed favorable potentials for diagnosing erythematous squamous, and the proposed ANFIS model possessed a higher level of precision than the neural network model alone [7].

Finally, in another study, the ANFIS model was adopted in one research for classifying electroencephalogram (EEG) signals, where the model was developed in two stages: extracting features by wave-less transform, and then training ANFIS with back-propagation gradient descent method combined with the least squares method. The proposed ANFIS model possessed both the neural network adaptive capabilities and the fuzzy logic qualitative approach. The study confirmed that the proposed ANFIS model was potentially capable of classifying the EEG signals [8].

Adaptive neuro-fuzzy inference system is a system based on fuzzy if-then rules that cannot be explained by classic probability theories. Fuzzy logic aims at extracting precise results, using rules defined by specialized experts. On the one hand, neural networks can be trained, and are capable of use observed data to determine network parameters in a manner that the desired input would produce favorable output. On the other hand, neural networks are unable to employ human knowledge, or be deduced from verbal expressions as fuzzy systems do [9, 10]. In addition to possessing the learning capacity of neural networks and the deduction ability of fuzzy systems combined, ANFIS networks are able to find any non-linear graph or model, in order to precisely relate input (initial values) to output (predicted values) data. Figure 1 illustrates the structure for such networks.

Considering the importance of breast cancer risk assessment and its fuzzy nature, it is interesting to determine the level or power at which this model can detect breast cancer patients, using the prerequisite data. In addition, how well such a model may perform for real data or how much one can rely on an ANFIS model to be re-used for real data (patients) are also important questions to be answered. This study, hence, proposed an ANFIS model inspired by the learning capacity of neural networks with the aim of promoting the learning capacity, maximizing approximation precision and simplifying the structure of the machine in breast cancer diagnosis.

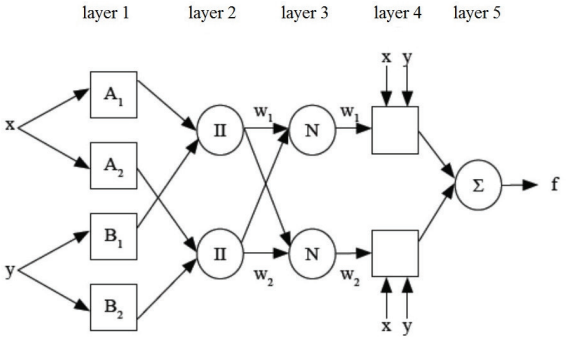

Figure 1: General Structure of Adaptive Neuro-Fuzzy Inference System (ANFIS) in Five Layer.
A combination of Neural Network with fuzzy system. Two inputs: x,y arriving the system will turn to an output function: f to help breast cancer diagnosis.

METHODS

Data and Preprocessing

For calculations and modeling purposes, this study used an aggregated dataset of our real database with Wisconsin University Database of breast cancer patients. The Wisconsin University Database includes information on 699 patients with 22 different characteristics. Since this study intended to develop a general model useful for all cases of breast cancer risk assessment, to have a suitable database including patients' and non-patients' information, the data pertaining to 809 non-patients referred to the Cancer Research Center of Shahid Beheshti University of Medical Sciences (CRC, SBMU) was added to the initial dataset. The result was 1508 re-

cords and 22 characteristics, employed as the raw data of patients and non-patients pre-processed for software analysis. For pre-processing the data, database columns irrelevant to cancer risk factors or containing patients' personal details (such as IDs) were first removed. It was then attempted to promote data mining validity by eliminating records with over 15% missing data, or those highly irrelevant to the research purpose, yet no such records were found in the combined dataset, according to the validation process conducted before inclusion. Finally, the missing data were supplied through the central index for the median of the 25 nearest neighborhood in SPSS21 software, which did not alter the total number of records as 1508.

Initially, the probable factors effective on cancer appearance needed to be determined. For importing the risk factors to the ANFIS model, we needed the priority or importance level for each one. Therefore, information was obtained from pertinent reference books and works published in reliable journals in this respect, and then expert opinions helped summarize major risk factors into three levels of severe, moderate and mild factors (with importance level of 1, 2 and 3 respectively) (Table 1) [5]. According to this study, the impact of these factors was then classified into low-risk and high-risk groups, which also meant that the fuzzification of input data was completed at this stage.

In the next stage, the neural network from ANFIS was chosen for modeling. To avoid the problem of inadequate efficiency in processing large numbers of network input (one shortcoming of this model occurring in some cases), the subtractive clustering technique was introduced into the model [11].

System Training and Testing

Seventy percent of the data (1055 records) was used for training, and 30 percent (453 records) for testing the ANFIS model. The intended models were thus designed according to this classification. Finally, the output was provided to the user in a numeric format between zero and one, which represented the occurrence probability of one of the trained model results.

Assessing the Adaptive System by Real Clinical Data

At this stage, the system's performance was assessed again, but using real breast cancer data from 2048 patients referred to the BCRC Clinic, and the precision degree of the adaptive system was recorded.

It must be mentioned that all stages of the study, including training and testing, were conducted by the Matlab-11R software.

RESULTS

Variable Addition Results

Table 1: Breast Cancer Risk Factors Organized by Their Priorities and Characteristics According to the Literature and Experts

Risk Factor	High Risk Group	Low Risk Group	Priority
First-Degree Relatives With BC (Mother, Sisters, Daughters)		0	1
Second-Degree Relatives With BC (Grandmothers, Aunts, Nieces, Cousins)		0	1
SNP Information			
Age, Y	> 50	< 45	1
Inheriting BRCA1, 2	Yes	No	1
Age at Menarche, Y	< 12	> 14	2
Age at Menopause, Y	> 55	< 45	2
Age at First Child Birthday, Y	> 30	< 30	2
Number of Pregnancies	<3	>4	2
Mammogram density	>50%	<5%	2
Biopsy Abnormalities	Yes	No	2
Exposure to Radiation, mrad	> 400	< 200	2
Oral Contraceptive Consumption Period, Y	> 4	< 2	3
Alcohol Consumption	> 2 drinks a day	< 2 drinks a day	3
Hormone Replacement Therapy Period, Y	> 4	< 2	3
First-degree relatives with other cancers		0	3
Second-degree relatives with other cancers Obesity	> 25	0 < 25	3
Vegetable and Fruit Consumption (serves a day)	<1	>2	3
Physical Exercises, Min	> 5	<15	3
Race	East European, European	Asian, African	3

After being extracted and verified by experts, the variables were displayed in one table (Table 1), according to their importance level. It is worth mentioning that race was considered as a one-way variable, and excluded from assessment computations in this study.

System Training

Defined Fuzzy Functions

The fuzzy membership functions selected after reviewing the related literature were sigmoid functions for first and second inputs, due to the continuity and proximity of the values achieved by these functions to normal levels. Figure 2 illustrates these functions.

Training the System and Resting the Results With Combined Data

Figure 3 portrays the results of assessing the trained model with test data, and in this case, 453 samples from the mentioned dataset were used for this purpose. The model trained with the dataset in the previous stage was tested here by the remaining 30% of the combined data. The software usually displays a general view of the training data, and this number was chosen by the researchers for higher image clarity.

The precision index of the developed model for this dataset at this stage was 81%. The sensitivity and specificity indices were 85.1% and 74.5%.

Testing Results With Real Data

The next stage entailed testing the developed model with the real data from the BCRC Clinic, using similar data belonging to 2048 patients, as well as 10 features similar to the first stage. The test results for the first 140 patients displayed by the software are illustrated in the figure below (Figure 4).

The researchers' objective was to examine whether a model trained with a standard dataset could be applied to a database of actual patients, different from the training data used for the model.

At this stage, the model was tested by all the previous assumptions, but using actual patients' data in this case. More precisely, 80% of the data was used for training, and the rest for testing the model. The proposed model registered a precision index of 84.5%, which exceeded the corresponding value achieved for the Wisconsin dataset. This result was repeated for sensitivity and specificity in value. These two indices were 89.3% and 79.9%, respectively in the second test.

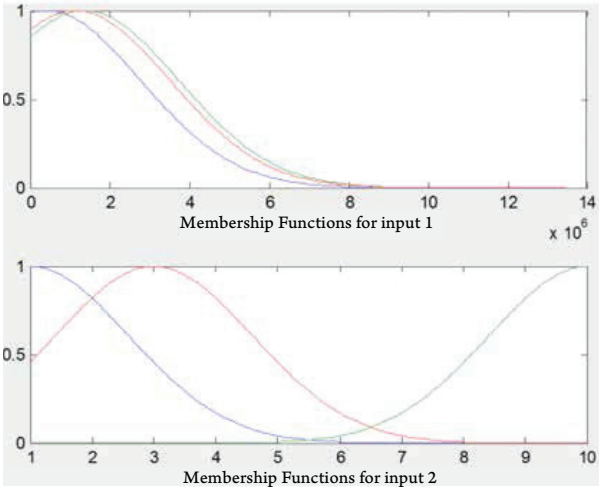

Figure 2: Membership Functions of First and Second Input Batches. They were defined due to the continuity and proximity of the values achieved by these functions to normal levels in the process of optimization.

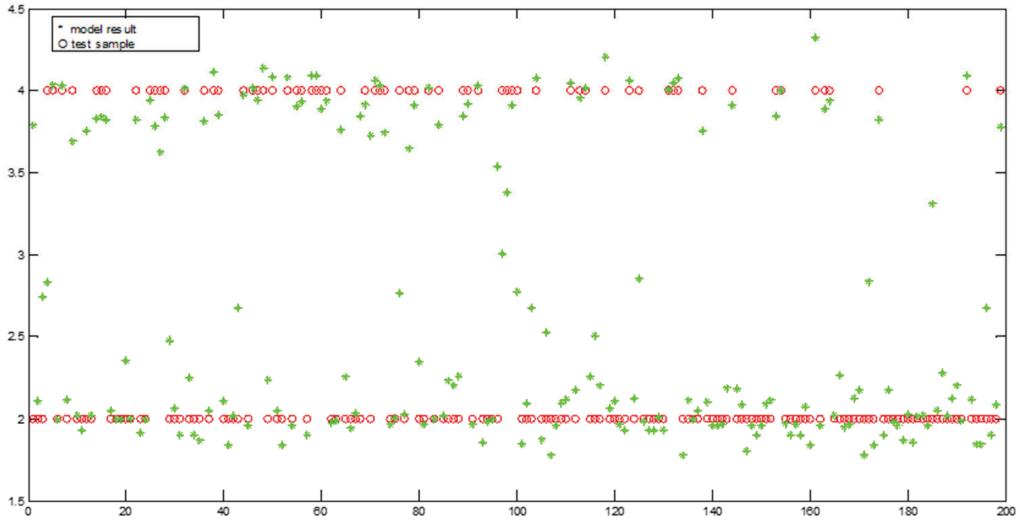

Figure 3: Results of Assessing the Trained Model With Test Data Pertaining to the First 200 Samples Displayed by the Software.
The Model lines are selected to model the data and training the ANFIS system. The system will use this model for diagnosis in the Test phase.

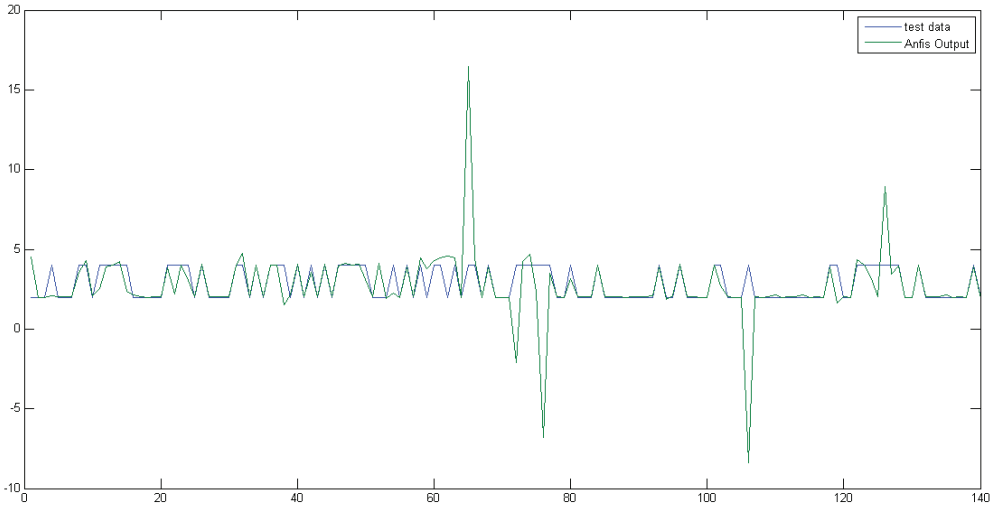

Figure 4: Results From Testing the Model With Real Data for 140 Test records from BCRC patients.

DISCUSSION

In this study, a final ANFIS model was developed from a standard dataset, and was tested with the same and real data on breast cancer. This model, established on the records of the Wisconsin breast cancer database, is a general model that can be applied to any dataset, such as the patients' data from BCRC Clinic, as a tool for breast cancer risk assessment. For that purpose, the factors effective on cancer occurrence were first prepared through the subtractive clustering algorithm, and then fed as input values into the ANFIS engine. The system, trained with the standard breast cancer dataset, would be tested as the next step, and the output would finally be expressed as a numeric value representing probability. It was accepted that the Wisconsin dataset possessed sufficient clinical relevance, comprehension and acceptability to assure the researcher(s) of initial applicability and the correct process for algorithm implementation. However, real datasets are usually entangled with problems regarding missing, probably irrelevant or unreliable data, which dictate the need to perform the customary pre-processing of data; this would in turn lead to an increased risk of systematic errors in the system on the one hand, and require further time and resources for the study on the other.

The most significant result of our study was the high accuracy calculated in the second test (by real data). According to the high number of real records (more than two thousand referees to a breast cancer Clinic) and after some iterations in deploying the algorithm, the authors concluded that the calculated precision would not differ significantly by adding more records. This relative reliability of the results should also be confirmed by re-testing the model, or applying it to real databases. As we know, all models would be applicable at least for the database they were implemented in. Contrary to this obvious fact, the results of the present study showed that a model designed based on a standard database could be used for other databases with an acceptable accuracy level after reliable tests on each database before use.

The ANFIS model proposed by the authors in this study is prominent in different respects. First, the model was prepared and tested with standard data, and then separately assessed with those pertaining to actual patients. Second, assessment results by real data showed conformity with the patients' present data by some 84.5%, which is relatively high and significant, albeit strange and unexpected in modeling terms, since a model trained with a particular dataset would solely be valid for that set [12]. The favorable results observed in this study could be attributed to the standard nature of the Wisconsin dataset, and the absence of missing data therein. Third, using subtractive clustering guided the studied features to influence the study commensurate to their actual impact, regardless of their designated weight, discreteness and variable type [13]. Thanks to this advantage, such clustering method could also be employed for other soft computing studies in clinical fields. The present model cannot be illustrated in the form of regressions, graphs or other visual means, yet the integrated fuzzy precision arising from human mentality could produce results akin to the physicians' decisions [14], which is definitely a claim requiring further clinical research.

This study could boldly claim to be the first (inside or outside the country) to have employed ANFIS for breast cancer risk assessment. Current studies with ANFIS mainly focus on classifying medical imaging [15,

16]. As for cancer diagnosis, Al-batah et al. conducted a research with a relatively high precision of 94%, pointing toward the higher quality and precision of the acceptable results of using the adaptive system, compared to those achieved by each assessment system separately [17]. Researchers are therefore comparing the precision and other crucial features of the adaptive model with other alternatives, yet the comparison is challenged and complicated by the diversity of variables and outputs, as well as variety of indices in different software being used.

Ubeyli used ANFIS for automatic detection of breast cancer. Although all system input variables did not fully conform in the two studies, the findings of this study could claim a higher quality than those of Übeyli, in terms of precision and inclusion of native data [6]. In their second study, Ubeyli et al. worked on neuro-fuzzy systems, which again was performed at a lower level than ANFIS. Neuro-fuzzy systems generally resemble adaptive ones, yet the former puts a higher emphasis on neural networks. Nevertheless, the nature of breast cancer data, and larger existence of fuzzy data are amongst factors that raise the expectations toward fuzzy-adaptive models for better efficacy [7]. The best possible result of the present study could be higher precision of native data from the BCRC Clinic, which verified not only the suitable native model being developed, but also the proper number of data records being chosen for such modeling purposes (whereas higher numbers of records would have jeopardized modeling precision) [4]. On the other hand, a model arising from native datasets would definitely have higher value and validity for that area or dataset. Generalizing these results to other datasets would require similar studies on corresponding databases of other health organizations.

It must be re-emphasized that the model developed with the Wisconsin Breast Cancer Database could be replicable for other datasets, yet further certainty would be achieved by evaluating the same database in another period, or with a higher number of clinical patients.

One advantage of the present study was predicting breast cancer risk for the first time in the country, which could be beneficial for relevant researchers, and useful for national screening purposes. Another technical innovation was coupling ANFIS with the simultaneous dimension reduction technique for the first time, which augments the proposed model's precision. More importantly, the model elicited precious results by assessing actual patients' data, which was relatively unprecedented in the field.

Among the limitations of this study, the first would be a high volume of missing data, which led to the elimination of unusable records. Despite the large number of suitable records remaining after such eliminations, the study results were still undeniably affected, even faintly, by their exclusion. Moreover, large data volumes and fostering large dimensions might also influence the results. Another limitation was using a relatively small fraction of the dataset for training. While 1508 appears to be a sufficiently high number of patients for this study, the variety of features requires larger populations in order to promote the model's precision.

As mentioned above, for all studies on artificial intelligence and soft computing, it should be kept in mind that the results of this study would only apply to the Wisconsin Database and the patients of the BCRC Clinic, only to be generalized if the present study was repeated for other centers as well. The results could be replicated for larger numbers of patients in the same database, in order to update or modify the model.

Further research in this respect would entail applying a multi-objective Genetic Algorithm (GA) to optimize a fuzzy system for breast cancer risk assessment, since such an algorithm would be capable of developing an optimal prediction model by formulating several parameters of a fuzzy system, including risk factor weights and inclusion functions in the system. On the other hand, researchers interested in advancing the topic in fuzzy directions could define the dependent variable of this study by other fuzzy functions as well. Moreover, a different number of defined functions could also initiate further studies recommended in this respect.

In this study, a final ANFIS model was developed and tested for two standard and real datasets on breast cancer. The resulting model could be employed with high precision for the BCRC database, as well as to conduct similar studies and re-evaluate other databases. This system and its similar counterparts can be recommended for screening purposes at the national level, as well as areas lacking sufficient numbers of specialized personnel.

ACKNOWLEDGEMENT

This study received no funding or support. The authors would like to express their gratitude to Mr. Vahid Zibakalam.

CONFLICT OF INTEREST

There is no conflict of interests.

ETHICS APPROVAL

The ethics committee of breast cancer research center of ACECR approved the study.

REFERENCES

1. Cancer-United States Cancer Statistics (USCS). Top Ten Cancers 2009 [2013 Sep 13]. Available from: http://apps.nccd.cdc.gov/uscs/toptencancers.aspx#text.
2. Jemal A, Siegel R, Ward E, Murray T, Xu J, Smigal C, et al. Cancer statistics, 2006. CA Cancer J Clin. 2006;56(2):106-30. PMID: 16514137
3. How many women get breast cancer? 2013 [cited 2013 Sep 13]. Available from: http://www.cancer.org/cancer/breastcancer/overviewguide/breast-cancer-overview-key-statistics.
4. Akl A, Ismail AM, Ghoneim M. Prediction of graft survival of

living-donor kidney transplantation: nomograms or artificial neural networks? Transplantation. 2008;86(10):1401-6. DOI: 10.1097/TP.0b013e31818b221f PMID: 19034010
5. Tatari F, Akbarzadeh TM, Sabahi A. Fuzzy-probabilistic multi agent system for breast cancer risk assessment and insurance premium assignment. J Biomed Inform. 2012;45(6):1021-34. DOI: 10.1016/j.jbi.2012.05.004 PMID: 22692028
6. Ubeyli ED. Adaptive neuro-fuzzy inference systems for automatic detection of breast cancer. J Med Syst. 2009;33(5):353-8. PMID: 19827261
7. Ubeyli ED, Guler I. Automatic detection of erthemato-squamous diseases using adaptive neuro- fuzzy inference systems. Comput Biol Med. 2005;35(5):421-33. PMID: 16136651
8. Guler I, Ubeyli ED. Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients. J Neurosci Methods. 2005;148(2):113-21. DOI: 10.1016/j.jneumeth.2005.04.013 PMID: 16054702
9. Chiu SL. Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst. 1994;2(3):267-78.
10. Yager RR, Filev DP. Generation of fuzzy rules by mountain clustering. J Intell Fuzzy Syst. 1994;2(3):209-19.
11. Moertini V. Introduction to five data clustering algorithm. Integral. 2002;7(2):87-96.
12. Larose DT. Data Mining: Methods and Models. USA: Wiley; 2005.
13. Kim DW, Lee KY, Lee D, Lee KH. A kernel-based subtractive clustering method. Pattern Recognit Lett. 2005;26(7):879-91.
14. LaBrunda M, LaBrunda A. Fuzzy logic in medicine. Inf Tech Res Advance Trends. 2009;1(1):27-33.
15. Hosseini MS, Zekri M. Review of Medical Image Classification using the Adaptive Neuro-Fuzzy Inference System. J Med Signals Sens. 2012;2(1):49-60. PMID: 23493054
16. Bhardwaj A, Siddhu KK. An Approach to Medical Image Classification Using Neuro Fuzzy Logic and ANFIS Classifier. Int J Comput Trends Tech. 2013;4(3):236-40.
17. Al-batah MS, Isa NA, Klaib MF, Al-Betar MA. Multiple adaptive neuro-fuzzy inference system with automatic features extraction algorithm for cervical cancer recognition. Comput Math Methods Med. 2014;2014:181245. DOI: 10.1155/2014/181245 PMID: 24707316