International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 614 - 617, 2013, ISSN: 2319–8656

Detect Breast Cancer using Fuzzy C means Techniques in Wisconsin Prognostic Breast Cancer (WPBC) Data Sets

Tintu P B
CMS College of Science and Commerce
Coimbatore, Tamil Nadu, India

Paulin. R
CMS College of Science and Commerce
Coimbatore, Tamil Nadu, India

Abstract- Medical data mining is very much valuable to medical experts. The main task of data mining is diagnosing the patient's disease Classification. Breast cancer is a severe and life threatening disease very commonly found in woman. An unusual growth of cells in breast is the main source of breast cancer those cells can be of two types malignant (Cancerous) and benign (Non-Cancerous) these types must be diagnosed taking proper meditation and for proper treatment. Modern medical diagnosis scheme is totally based on data taken through clinical and/or other test; most of the decision related to classification of a disease is a very crucial and challenging job. In this research work, using intelligent techniques of data mining is Fuzzy C Means; we have focused on breast cancer diagnosis by fuzzy systems. Fuzzy rules are desirable because of their interpretability by human experts. It has been applied to classify data related to breast cancer from UCI repository site. Experimental works were done using MATLAB in order to reduce dimensionality of breast cancer data set a ranking based feature selection technique. Results on breast cancer diagnosis data set from UCI machine learning repository show that this approach would be capable of classifying cancer cases with high accuracy rate in addition to adequate interpretability of extracted rules.

Keyword— Classification, Clustering, Fuzzy C Means, Breast Cancer, Wisconsin Prognostic Breast Cancer (WPBC).

1. INTRODUCTION

Proper diagnosis of any human disease precisely and powerfully is a challenging task for the people involved in health care organization and offers a strong base for further treatment and medication. Breast cancer comes fourth in cancer diagnosis in women between 20 to 29 years. Breast cancer is most common type of cancer in women, with more than one million instances and nearly half million of deaths occurring worldwide annually [1]. In 2010, there were reported approximately 207090 newly diagnosed cases and 30840 deaths in the United States, and total of 1,638,910 new cancer cases is projected to occur in 2012 [2]. A breast cancer victim's chances for long-term survival are improved by early detection of the disease, and early detection is in turn enhanced by an accurate diagnosis. In addition, the National Cancer Institute of U.S. estimates that 16.4 percent of women born today and live with a breast cancer diagnosis [3]. For the diagnosis of the breast cancer cases as well as for the prognosis of the disease many methods have been discussed [4], [5], [6], [7], [8], [9], [10] and [11]. All machine leaning techniques to provide the same levels of exactness, without the negative sides of surgical biopsy.

The biggest problem in medical science includes the diagnosis of disease since the reason of breast cancer is unknown, although scientists know some of the risk factors like ageing, genetic risk factors, menstrual periods, family history, not having children, alcohol, overweight, obesity, etc. [12]. Symptoms of cancer include a lump in the breast or underarm that persists after menstrual cycle, swelling in the armpit, pain or tenderness in the breast, any change in the size, texture, contour, or temperature of the breast, a marble-like area under the skin. Many cancer diseases take place within the pale of the same family and the immediate relatives (siblings, parents, and children) of patients with cancers often have an increased risk of cancer.

This paper deals with the breast cancer diagnosis problem using the Wisconsin Diagnostic Breast Cancer (WDBC) as well as the Wisconsin Prognostic Breast Cancer (WPBC) data sets, which are publicly available by anonymous ftp [13]. These data sets involve measurements taken permitting the Fine Needle Aspirate (FNA) test. In case that a patient is diagnosed with breast cancer, the malignant mass must be excised. After this or a different post-operative procedure, a prediction of the expected course of the disease must be determined. However, prognostic prediction does not belong either on the classic learning paradigms of function approximation or classification. This is due to a patient can be classified as a "recur" case (instance) if the disease is observed, while there is no a threshold point at which the patient can be considered as a "non-recur" case. The data are therefore censored since a time to recur for only a subset of patients is known. For the others patients, the length of time after treatment during which malignant masses are not found is known. This time interval is the disease free survival (DFS) time, which can be reported for an individual patient or for a study population. In particular, the right endpoints of the recurrence time intervals are right censored, as some patients will inevitably change hospital, doctors or die of other unrelated with the cancer causes. Therefore, the training dataset for the learning phase is not well-defined. Several groups have approached prognosis as a separation problem using different learning architectures such as back propagation artificial neural networks [14], entropy maximization networks [15], [1] decision trees [17] and fuzzy-based measurements [18].

In this paper is proposed a innovative approach to automatically detect the breast cancer using Fuzzy C-Means data mining techniques. This approach utilizes fuzzy c-means clustering for classification of the data from the WBCD dataset. The rest of paper is organized as follows. Section 2 describes data set of breast cancer disease. Section 3 presents of fuzzy c-means algorithm for classification. Section 4 presents shows results and finally sections 5 conclude the paper.

2. DATA SET OF BREAST CANCER DISEASE

The WDBC and WPBC datasets are made at the University of Wisconsin Hospital for the diagnosis and prognosis of breast tumours solely based on FNA test. This test involves fluid mining from a breast mass using a small-gauge needle and then visual checkup of the fluid under a microscope. This dataset is created by Dr. William H. Wolberg from University of Wisconsin Hospitals.

Table1. Attributes and values of cancer clinical instances

No	Attribute	Values
1	Sample code number	Id number
2	Clump Thickness	1-10
3	Uniformity of Cell Size	1-10
4	Uniformity of Cell Shape	1-10
5	Marginal Adhesion	1-10
6	Single Epithelial Cell Size	1-10
7	Bare Nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoli	1-10
10	Mitoses	1-10
11	Class	2 and 4

This dataset contains a total of 699 clinical cases, with 458 benign and 241 malignant cases. Each and every clinical case has 9 attributes with assigned integer values varying from 1 to 10 and one class output with a binary value of either 2 or 4, showing benign and malignant breast cancer diagnoses, separately. The table 1 showed the physical meaning of the nine attributes. Among the 699 clinical cases, 16 instances are each missing one of the nine attributes. The dataset consists of 683 cases, with each entry indicating the classification for a certain ensemble of measured values. For a consistently of high accuracy, the 16 cases each have missing one attribute are removed from this dataset. The resulting dataset has 683 clinical cases, with 444 (65.01%) benign and 239 (34.99%) malignant diagnoses. The evolutionary experiments executed fall into two categories, according with the data repartitioning into two dissimilar sets: training set and test set. The experimental categories classified are: (1) data set which contains all 683 cases of the WBCD dataset (2) training set that contains 70 cases

3. PROPOSED FUZZY C MEANS METHOD

Complexity of medical diagnosis problems has showed that using traditional methods in solving these issues is not appropriate. In medicine, the absence of information, and its roughness, and contradictory nature is common facts. Fuzzy logic plays an important part of diagnosis the medical disease. Some examples showing that fuzzy logic involving many disease groups are the following [19]:

(1) To analyze diabetic neuropathy
(2) To determine appropriate lithium dosage
(3) To calculate volumes of brain tissue from magnetic resonance imaging
(4) To characterize stroke subtypes and coexisting causes of ischemic stroke.
(5) To improve decision-making in radiation therapy
(6) To control hypertension during anesthesia
(7) To determine flexor-tendon repair techniques
(8) To detect breast cancer, lung cancer
(9) To assist the diagnosis of central nervous systems tumors (Astrocytes tumors)
(10) To discriminate benign skin lesions from malignant melanomas
(11) To visualize nerve fibers in the human brain
(12) To represent quantitative estimates of drug use
(13) To study the auditory P50 component in schizophrenia

3.1 Fuzzy c-means algorithm

Clustering is a process of grouping data in clusters, where data placed in one cluster are more similar to each other than those in other clusters. Fuzzy c-means algorithm uses the reciprocal of distances to decide the cluster centers. Fuzzy logic was introduced by Zadeh during 1960s for handling the uncertain and imprecise knowledge in real world applications [20]. Fuzzy C Means centroid of a cluster is calculate as mean of all points' value, weighted by their degree of belonging to the cluster. Advantages of this algorithm are that this method gives better results than k-means algorithm. Furthermore the greatest advantage of using fuzzy logic lies in the fact that scientists can model complex systems by implementing human experience, knowledge, non-linear, and imprecise and practice as a set of inference (or fuzzy) rules that use linguistic or fuzzy variables [21]. The FCM algorithm is to improve the accuracy of clustering under noise. FCM method is created on minimization of the function:

$$J_{m=\sum_{i=1}^{N}\sum_{j=1uij}^{C}{}^{m}\|Xi-Cj\| \quad {}^{2},1\le m\le\infty}$$

(1)

Where u_{ij} represents degree of membership element x_i in the cluster *j*. The squared element is the Euclidian distance between i^th data and j^th center of cluster. Completed every iteration update of the membership function and center of clusters c_j is calculated as following:

$$uij = \frac{1}{\sum_{k=1}^{C}\frac{(\|Xi-Cj\|)}{\|Xi-Ck\|}^{\frac{2}{m-1}}} \qquad (2)$$

Where centers of the clusters can be calculated as follows:

$$Cj = \frac{\sum_{i=1}^{N} uij^{m}}{\sum_{i=1}^{N} Uij^{m}} \qquad (3)$$

Algorithm performs calculation as follows:

1. Initialization of *U=[uij], U(0)*
2. Calculation of centers of the vectors *C(k)=[cj] and U(k)*
3. Update of *U(k) to U(k+1)*
4. Comparison, if absolute value ‖U(k+1)-U(k) ‖ < ε, where ε represents predefined criteria, then ST In this study, 683 clinical instances in the Breast Cancer

4. RESULT

Wisconsin (Original) Dataset were used for this model. Even original data from has 699 clinical cases 16 cases were removed because of missing of one or more attributes. In rest 683 clinical cases there are 444 benign which represent 65% and 239 malignant breast cancer cases which represent 35%. The class output of an original binary value was 2 or 4 indicating benign and malignant breast cancer, one to one. Estimation of error of this model is done using two approaches such as training set and test set.

In Table1 are shown results using this method. In Fig.1 is shown scheme of the model and steps performed during the evaluation of the results.

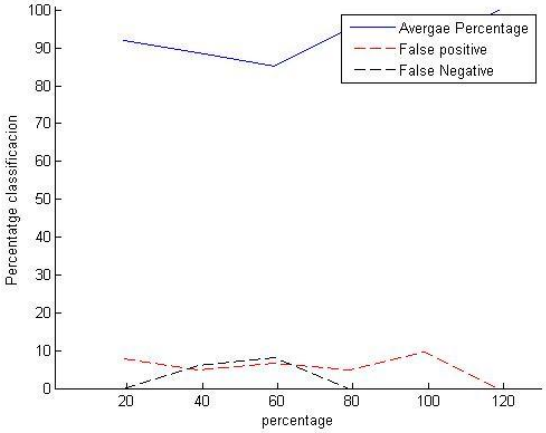

Fig 1. Evaluation Results of WPBC Dataset

Table 1: WPBC Data set result

Data	FCM Classification	Disease Diagnosis
True positive	100 %	100 %
True negative	87%	80.5%
False positive	0 %	0 %
False negative	13%	19.5%

In Table 2 are shown results using this method. Fuzzy c-means is done with initial data set as well as training data.

In this approach, WBCD data set divided to 10 parties, nine parties for train set and one party for test set. This process runs for more than ten times. Results of these run is indexed in table 2. Average of train set accuracy and test set accuracy is 98.91 and 97.99, in sequence.

Proposed approach is compared with some algorithms, such as Naïve Bayes, SVM, MLP. Result of comparison is indexed in table 3. According to Table 3, train set accuracy of FUZZY-C means algorithm (proposed algorithm) better than other algorithms and also test set accuracy of FUZZY-ACO algorithm better than other algorithms. However, the main advantage of proposed algorithm is high interpretability.

Table 2. Results of proposed approach

Run	1	2	3	4	5	6	7	8	9	10
Train	614 /629	614 /629	619 /629	613 /629	622 /629	624 /629	615 /629	619 /629	620 /629	616/630
Test	68 /70	69 /70	67 /70	69 /70	67 /70	69 /70	68 /70	68 /70	67 /70	66/69
Rules	20	20	20	20	20	20	20	20	20	20
Length	1.3	1.4	1.3	1.1	1.55	1.05	1.2	1.4	1.3	0.75

Table 3. Results Comparison of some algorithms

Algorithms	*MLP*	SVM	Naïve Bayes	Fuzzy C Means
Test Set	94.56	96.99	96.56	97.13
Train Set	97.29	98.09	98.41	98.62

5. CONCLUSION

Breast cancer has become the leading reasons of death in women in most of countries. We need most effective techniques to reduce breast cancer a death is detect it earlier. Early diagnosis requires a reliable and accurate diagnosis procedure that allows physicians to decide benign breast tumors from malignant ones without going for surgical biopsy. In this paper, is proposed a new alternative approach for breast cancer disease diagnosis and classifying benign and malignant breast cancer using fuzzy c-means. This proposed approach was based classification of input data, training data and test data. Results on breast cancer diagnosis data set from UCI machine learning repository show that the proposed FUZZY C Means would be capable of classifying cancer cases with high accuracy rate in addition to adequate interpretability of extracted rules.

7. REFERENCES

[1] G. Salama, M.B. Abdelhalim, and Magdy Abd-elghany Zeid. Son, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-classifiers" *International Journal of Computer and Information Technology*, vol. 01, pp. 36-43, September 2012.

[2] George J.Miao, Kathleen H.Miao, Julia H.Miao, "Neural pattern Recognition Model for Breast Cancer Diagnosis" *Journal of selected areas in Bioinformatics, August edition,2012*,pp. 1-8.

[3] U.S. National Institutes of Health, National Cancer Institute, http://cancernet.nci.nih.gov/

[4] Wang, T.C., Karayiannis N.B., Detection of microcalcifications in digital mammograms using wavelets, IEEE Transactions on Medical Imaging, Vol. 17, Issue. 4, Aug. 1998, pp. 498 – 509.

[5] Huo, Z., Giger, M., Vyborny, C., Wolverton, D., Schmidt, R., Doi, K., Automated computerized classification of malignant and benign mass lesions on digital mammograms, Acad. Radiol. 5, 155–168, 1998.

[6] Cheng Heng-Da, Lui Yui Man, Freimanis R.I., IEEE Transactions on Medical Imaging, Vol. 17, Issue. 3, June 1998, pp. 442 – 450.

[7] Pendharkar P.C., Rodger J.A., Yaverbaum G.J., Herman N. and Benner M., Association, statistical, mathematical and neural approaches for mining breast cancer patterns, Expert Systems with Applications, 17:223–232, 1999.

[8] Setiono R., Generating concise and accurate classification rules for breast cancer diagnosis, Artificial Intelligence in Medicine, 18:205–219, 2000.

[9] Chen D., Chang R.F., Huang Y.L., Breast cancer diagnosis using self-organizing map for sonography, Ultrasound in Medical Biology 2000, Vol. 26, pp. 405–11.

[10] Giger M., Huo Z., Kupinski M., Vyborny C., Computer-aided diagnosis in mammography. In Handbook of Medical Imaging, (Eds.) Sonka, M., Fitzpatrick, J., Medical Image Processing and Analysis, Vol. 2. SPIE Press, pp. 917–986, 2000.

[11] Tourassi G.D., Markey M.K., Lo J.Y., Floyd Jr. C.E., A neural network approach to breast cancer diagnosis as a constraint satisfaction problem, Med. Phys. Vol.28, pp. 804–811, 2001.

[12] S.Saheb Basha, Satya Prasad, "Automatic detection of breast cancer mass in mammograms using morphological operators and fuzzy c-means clustering" *Journal of theoretical and applied information technology*, pp. 704-709.

[13] http://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/, Wisconsin Diagnostic Breast Cancer (WDBC) Dataset and Wisconsin Prognostic Breast Cancer (WPBC) Dataset.

[14] Burke H. B., Goodman P.H., et al, Artificial neural networks improve the accuracy of cancer survival prediction, Cancer, Vol. 79, pp. 857-862, 1997.

[15] Choong P.L, deSilva C.J.S et al., Entropy maximization networks, An application to breast cancer prognosis, IEEE Transactions on Neural Networks, 1996, 7(3):568-577.

[16] Choong P.L., deSilva C.J.S, Maximum entropy estimation vs. multivariate logistic regression: which should be used for the analysis of small binary outcome data sets?, Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol.3, pp:1602 – 1605, 1998.

[17] Wolberg W.H., Street W.N., Heisey D.M., and Mangasarian O.L., Computer-derived nuclear features distinguish malignant from benign breast cytology, Human Pathology, 26:792--796, 1995.

[18] Seker H., Odetayo M., Petrovic D., Naguib R.N.G., Bartoli C., Alasio L., Lakshmi M.S., Sherbet G.V., A fuzzy measurement-based assessment of breast cancer prognostic markers, Proceedings of the 2000 IEEE EMBS International Conference on Information Technology Applications in Biomedicine, 9-10 Nov. 2000, pp.174 – 178.

[19] Angela Torres, Juan Nieto,"Fuzzy ogic in bioinformatics and medicine", Journal of Biomedicine and Biotechnology, Vol. 2006, pp. 1–7.

[20] Timothy J.Ross, "Fuzzy Logic with engineering applications", Third edition, Wiley, 2010.

[21] Victor Balanica, Ioan Dumitrache, "Evolution of breast cancer risk by using fuzzy logic" U.P.B.Sci.Bull, Vol. 73, 2011 pp. 54-64