Fuzzy C-Means Algorithm to Diagnose Breast Cancer

Dr. W. Abdul Hameed, Dr. Shaik Sharief Basha

School of Advanced Sciences, VIT University, Vellore-632 014, Tamil Nadu, India.

e-mail: abdulhameed@vit.ac.in

Abstract:- The automatic diagnosis of breast cancer is an important, real-world medical problem. A major class of problems in Medical Science involves the diagnosis of disease, based upon various tests performed upon the patient. When several tests are involved, the ultimate diagnosis may be difficult to obtain, even for a medical expert. This has given rise, over the past few decades, to computerized diagnostic tools, intended to aid the Physician in making sense out of the confusion of data. This Paper carried out to generate and evaluate fuzzy model to predict malignancy of breast tumor, using Wisconsin Diagnosis Breast Cancer Database (WDBC). Our objectives in this Paper are: (i) to find the diagnostic performance of fuzzy model in distinction between malignance and benign patterns, (ii) to reduce the number of benign cases sent for biopsy using this model as a supportive tool, and (iii) to validate the capability of this model to recognize new cases.

Keywords: *Breast cancer, fuzzy logic, fuzzy c-means algorithm*

*****

I. INTRODUCTION

Cancer, in all its dreaded forms, causes about 12 per cent of deaths throughout the world. In the developed countries, cancer is the second major cause of death, accounting for 21 per cent of all mortality. In the developing countries, cancer ranks third major cause of death accounting for 9.5 per cent of all deaths (ICMR, 2002). Cancer has become one of the ten main causes of death in India. As per the statistics, there are nearly 1.5 to 2 million cancer cases in the country at any given point of time. Over 7 lakh new cases of cancer and 3 lakh deaths occur annually due to cancer. Nearly 15 lakh patients require facilities for diagnosis, treatment and follow-up procedure (ICMR, 2001). Despite a great deal of public awareness and scientific research, breast cancer continues to be the most common cancer and the second largest cause of cancer deaths among women (Marshall E, 1993). In the last decade, several approaches to classification had been utilized in health care applications. A woman's chances for long term survival are improved by early detection of the cancer, and early detection is enhanced by accurate diagnosis techniques. Most breast cancers are detected by the patient or by screening as a lump in the breast. The majority of breast lumps are benign. And therefore, it is binding to diagnose breast cancer, that is, to distinguish benign lumps from malignant ones. In order to diagnose whether the lump is benign or malignant, the Physician may use mammography, fine needle aspirate (FNA) with visual interpretation or surgical biopsy. The reported ability for accurate diagnosis of cancer when the disease is prevalent is between 68% - 79%, in case of mammography (Fletcher SW *et al.*, 1993); 65% - 98% in case FNA technique is adopted (Giard RWM *et al.*, 1992), and close to 100% if a surgical biopsy is undertaken. And from this it is clear that mammography lacks sensitivity; FNA sensitivity varies widely and surgical biopsy although accurate is invasive, time consuming, and expensive. The goal of the diagnostic aspect of this Paper is to develop a relatively objective system to diagnose FNA with an accuracy that is best achieved visually. In this Paper, fuzzy model is used to diagnose whether the lump in the breast is cancerous or not, in the light of the patient's medical data.

II. MATERIAL

This Paper makes use of the Wisconsin Diagnosis Breast Cancer Database (WDBC) made publicly available at *http://ftp.ics.uci.edu /pub /machine-learning- database/breastcancer-wisconsin/*. This data set is the result of efforts made at the University of Wisconsin Hospital for the diagnosis of breast tumor, solely based on the Fine Needle Aspirate (FNA) test. This test involves fluid extraction from a breast mass using a small gauge needle and then a visual inspection of the fluid under a microscope. The WDBC dataset consists of 699 samples. Each sample consists of visually assessed nuclear features of FNA taken from patient's breast. Each sample has eleven attributes and each attribute has been assigned a 9-dimensional vector and is in the interval, 1 to 10 with value '1' corresponding to a normal state and '10' to the most abnormal state. Attribute '1' is sample number and attribute '11' designates whether the sample is benign or malignant. The attributes 2 to 10 are: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, blend chromatin, normal nucleoli and mitosis. There had 16 samples that contained a single missing (i.e., unavailable) attribute value and had been removed from the database, setting apart the remaining 683 samples. Each of these 683 samples has one of two possible classes, namely benign or malignant. Out of 683 samples, 444 have been benign and the remaining 239 have been malignant, as given by WDBC dataset.

III. FUZZY MODEL

Fuzzy System is an alternative to traditional notions of set membership and logic that had its origin in ancient Greek philosophy. It was Plato who laid the foundation for what would become fuzzy logic, indicating that there was a third region (beyond True and False) named as Utopia, where these opposites “tumbled about”. An iota of uncertainty is present in almost every sphere of our daily routine. Traditional mathematical tools are not enough to find solution to all the practical problems in fields such as Medical Science, Social Science, Engineering, Economics, etc., which involve uncertainty of various types. Zadeh (Zadeh L A, 1965) in 1965 was the first to come up with his remarkable theory of fuzzy set for dealing these types of uncertainties where conventional tools fail.

Fuzzy logic and neural networks have rooted firmly in multifarious application areas. Although these methodologies seem to be different, they have plenty of common features such as the use of basic functions (fuzzy logic has membership functions and neural network has activation functions) and the aim to estimate functions from sample data or heuristics. These methods are generally non-linear; they have an ability to deal with non-linearity; they follow more human-like reasoning paths than classical methods and are self-learning (Chennakesava R. Alavala, 2008).

Fuzzy C-Means Algorithm

In a mathematical or statistical environment a value may or may not belong to one class. In medicine, there are usually imprecise conditions and therefore fuzzy methods seem to be more suitable than crisp ones. Fuzzy c-means clustering is an easy and a well improved tool which has its application in several medical fields. However, in c-means algorithms, like in all other optimization procedures, which look for the global minimum of a function, there is the risk of coming down to local minima. Therefore, the result of such a classification has to be regarded as an optimum solution with a determined degree of accuracy (Bezdek J, 1974).Classes in which each member has full membership are called discontinuous or discrete classes. On the other-hand, classes in which each member belongs to some extent to every cluster or partition are called continuous classes (McBratney, AB., and JJ. de Gruijter, 1992). Continuous classes are a generalization of discontinuous classes when the indicator function of conventional sets theory, with values **0** or **1,** is replaced by the membership function of fuzzy sets theory, with values in the range of **0** to **1**. Let us take the partition of a set of n-individuals in c-discontinuous classes.

According to such partition, each individual is a member of exactly one class. This can be numerically represented by a (n x c) membership matrix U = (μ_{ij}) where μ_{ij} = 1 if the individual belongs to class **j** and μ_{ij} = 0 otherwise. In order to ensure that the classes are mutually exclusive, jointly exhaustive and non-empty, the following conditions are to be applied to **U**:

$$\sum_{j=1}^{c} \mu_{ij} = 1, \quad i = 1, 2, \ldots, n \qquad (1)$$

$$\sum_{i=1}^{n} \mu_{ij} = 1, \quad j = 1, 2, \ldots, c \qquad (2)$$

$$\mu_{ij}\ \varepsilon\{0, 1\} \quad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, c \qquad (3)$$

Condition (3) corresponds to all-or-nothing status of the membership in discrete classes. According to the fuzzy sets theory, this condition is relaxed in such a way that partial memberships are allowed i.e., to take any value between and including **0** and **1**. Thus, condition (3) for continuous classes becomes:

$$\mu_{ij}\ \varepsilon[0, 1] \quad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, c \qquad (4)$$

Any (n x c) matrix **U** satisfying (1), (2), and (4) represents a so-called fuzzy partition of the n-individuals into c-classes. Partitions satisfying 15), (2), and (3) are referred to as hard partitions. As condition (4) implies (3), hard partitions are special cases of fuzzy partitions. A special case of Fuzzy C-Means Algorithm (FCM) was first reported by Dunn in 1973. His algorithm was later generalized by Bezdek (1980), Bezdek *et al.*, (1984) and Kent *et al.*, (1988). These methods use an (n x p) data matrix X = (x_i) as input, where **p** denotes the number of variables and x_i denotes the value of individual **i**. The most popular fuzzy clustering method to date is the fuzzy c-means which is a generalization of the hard c-means clustering. The hard c-means method minimizes the function J(U, V), which represents the within-class sum-of-square errors between classes under conditions (1), (2), and (3):

$$J(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}\, d^2(x_i, v_j) \qquad (5)$$

where V=(v_j) is an (c x p) matrix of centre of classes, $\mathbf{v_j}$ denotes the value of the centre of the **j**th class. $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is the vector which represents the individual **i,** $v_j = (v_{j1}, v_{j2}, \ldots., v_{jp})^T$ is the vector which represents the centre of class **j,** and $d^2(x_i, v_j)$ is the square of the distance between $\mathbf{x_i}$, and $\mathbf{v_j}$ according to a given definition of distance, further denoted by d_{ij}^2 to simplify. J (U, V) is the sum of the square errors of the distance of each individual from the centre of the given classes (Kolen J. Hutcheson T, 2002). A fuzzy generalization of J (U, V) is obtained by a modification of the membership with an exponent. This weighting exponent controls the extent of membership sharing, or the "degree of fuzziness", among the resulting clusters. Hence, a new J_F (U, V) function is defined as:

$$J_F(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^{m}\, d^2_{ij} \qquad (6)$$

This function is minimized under conditions (1), (2), and (3). The value of **m** is chosen from (0, ∞). If m =1, the solution of (6) is a hard partition, i.e., the result is not fuzzy. As m → ∞, the solution approaches its maximum degree of fuzziness, with $\mu_{ij} = 1/c$ for every pair of **i** and **j**. There is no theoretical basis for an optical selection of **m**. It is often chosen on empirical grounds to be equal to **2**. If m>1, equation (6) could be minimized by Picard iteration (Bezdek, 1984):

$$\mu_{ij} = 1 / \sum_{k=1}^{c} \{ d_{ij} / d_{ik})^{2/(m-1)} \} \quad i=1,2, \ldots, n \qquad j=1,2, \ldots.. ,c$$

$$v_j = \sum_{i=1}^{n} (\mu_{ij})^m x_i \;/\; \sum_{i=1}^{n} (\mu_{ij})^m \qquad j=1, 2, \ldots.., c$$

IV. METHOD

To determine the performance of this model in practical usage, the database has been divided randomly into two separate sets, one for training and another for validation: (a) the training samples comprising **500** patients records (**303** benign, **197** malignant) and (b) the validation samples comprising **183** patients records (**141** benign, **42** malignant). Using the patients' records in training sample the model has been trained for estimating the membership function needed to establish the classification rules for fuzzy model. Then, the patients' record in validation samples (n = 183) have been utilized to evaluate the generalizing ability of the above said model. The best of this model has been compared in terms of accuracy, sensitivity, specificity, false positive and false negative.

V. EXPERIMENT AND RESULTS

The Researcher has already applied statistical and neural network classifications (Abdul Hameed. W, and Karthikeyan, K, 2016). He is now applying fuzzy classification using fuzzy c-means algorithm to **683** samples of WDBC data set. Table-1 presents results that have been obtained using this model. Out of **444** benign and **239** malignant instances, the fuzzy model has successfully identified **436** (**93.16%**) instances as negative and **217 (96.44%)** instances as positive. The same table also presents the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) results. According to these observations, Table-2 shows the sensitivity, specificity and efficiency of the fuzzy model using fuzzy c-means algorithm. The table presents the predicted values of a positive/negative test results.

Data	Group	Actual Instances	Fuzzy Logic-Predicated instances	
			Positive	Negative
Training Data (500)	Benign	303	08 (FP)	295 (TN)
	Malignant	197	175 (TP)	22 (FN)
Testing Data (183)	Benign	141	0 (FP)	141 (TN)
	Malignant	42	42 (TP)	0 (FN)
Total Data (683)	Benign	444	8 (FP)	436 (TN)
	Malignant	239	217 (TP)	22 (FN)

Table.1: *Performance of the model*

Data	Measurements	Fuzzy Logic
Training Data (500)	Sensitivity Specificity Efficiency Predictive Value (Positive) Predictive Value (Negative)	88.83 97.36 94.00 95.60 93.05
Testing Data (183)	Sensitivity Specificity Efficiency Predictive Value (Positive) Predictive Value (Negative)	100 100 100 100 100
Total Data (683)	Sensitivity Specificity Efficiency Predictive Value (Positive) Predictive Value (Negative)	90.80 98.20 95.60 96.44 93.16

Table-2: *Rresult Analysis of the model*

VI. CONCLUSION

This Paper has compared two models namely neural Network model and fuzzy model for the diagnosis of breast cancer. The ability of these models to differentiate malignant from benign tumor the Researcher has compared a group of **683** patients. The main aim of this Paper is to investigate which model obtains more reasonable specificity while keeping high sensitivity. The benefit is that the number of breast cancer patients for biopsy can be restricted. The seriousness of the ailment can easily be assessed.

The output of the fuzzy model has yielded a sensitivity of **90.8%**, maximum specificity of **98.2%** and efficiency of **95.6%** for the total data set, demonstrating that fuzzy model also differentiate a malignant from a benign tumor. The results of this study suggest that the diagnostic performance of this model is relatively better than the other models.

This Paper developed a diagnostic system that performs at or above an accuracy level in any procedure short of surgery. The results have also suggested that fuzzy model is a potentially useful multivariate method for optimizing the diagnostic validity of laboratory data. The Physicians can combine this unique opportunity extended by fuzzy model with their expertise to detect the early stages of the disease.

REFERENCES

[1] Abdul Hameed, W and Karthikeyan, K (2016), “Canonical Discriminant Analysis of Statistical Model and Learning Vector Quantization technique of Neural network in Diagnosing Breast Cancer” , IJPT/Sep 2016/Vol. 8/Issue No. 3/ 16198 – 16206.

[2] Bezdek J(1974),“Cluster validity with fuzzy sets”, J. Cybern (3):58-71

[3] Bezdek JC (1980), “A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms”, IEEE Trans. Pattern Anal. Machine Intell., vol. PAM 1-2, No. 1, pp.1-8.

[4] Bezdek JC., Ehrlich R and Full W (1984), “FCM: the fuzzy c-means clustering algorithm”, Computer & Geosciences, vol. 10, 191-203

[5] Chennakesava R. Alavala (2008), “Fuzzy Logic and Neural Networks: Basic concepts & Applications”, New Age International Publishers, New Delhi.

[6] Duda RO and Hart P (1973), “Pattern classification and Scene Analysis”, John Wiley & Sons

[7] Fletcher SW, Black W, Harrier R, Rimer BK and Shapiro S (1993), “Report of the International workshop on screening for breast cancer”, Journal of the National Cancer Institute, 85:1644-1656

[8] Giard RWM and Hermann J (1992), “The value of aspiration cytologic examination of the breast: A statistical review of the medical literature”, Cancer, 69:2104-2110

[9] ICMR (2001), National Cancer Registry Programme, Consolidated report of the population based cancer registries, 1997. Indian Council of Medical Research, New Delhi.

[10] ICMR (2002), National Cancer Registry Programme, 1981-2001, An Overview. Indian Council of Medical Research, New Delhi.

[11] Kent JJ and Mardia KV (1988), “Spatial Classification Using Fuzzy Memberships Models”, IEEE Trans. on PAMI, vol. 10 No. 5, September, pp. 659-671

[12] Kolen J. Hutcheson T (2002), “Reducing the time complexity of the fuzzy c-means algorithm”, IEEE Trans. Fuzzy Syst. 10(2): 263-267.

[13] Marshall E (1993), “Search for a Killer: Focus shifts from fret to hormones in special report on breast cancer”, Science, 259:618-621

[14] McBratney AB and de Gruiter JJ (1992), “A Continuoum Approach to Soil Classification by Modified Fuzzy k-Means with Extra-grades”, Journal of Soil Sciences, vol. 43, pp.159-175

[15] Zadeh LA (1965), “Fuzzy Sets, Information and Control”, vol. 8, pp. 338-353.