Fuzzy Analysis of Breast Cancer Disease using Fuzzy c-means and Pattern Recognition

Indira Muhic
International University of Sarajevo, Faculty of Engineering and Natural Sciences
Hrasnicka Cesta 15, 71210 Sarajevo, Bosnia and Herzegovina
imuhic@ius.edu.ba

Abstract - Breast cancer is the second largest cause of cancer deaths among women. At the same time, it is also among the most curable cancer types if it can be diagnosed early. The automatic diagnosis of breast cancer is an important, real-world medical problem. In this article is introduced a new approach for diagnosis of breast cancer. The proposed approach uses Fuzzy c-means (FCM) algorithm and pattern recognition method. Algorithm has been applied to breast cancer clinic instances obtained from the University of Wisconsin. Using FCM algorithm clinic instances are grouped into two clusters, one with benign instances and other with malign instances. Further, input data are divided in train data and test data and success of each is evaluated. In pattern recognition method each input test data is assigned to one of the clusters obtained from the process of FCM classification. The proposed system has showed that the recommended system has a high accuracy.

Index Terms— Fuzzy c-means, Pattern recognition, Fuzzy logic, Breast cancer disease

1. INTRODUCTION

The most widespread disease today representing the first deadly cause for old age women is breast cancer. After thyroid cancer, melanoma, and lymphoma, breast cancer comes fourth in cancer incidences in women between 20 to 29 years. Breast cancer is most common type of cancer in women, with more than one million cases 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.

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, family history, menstrual periods, not having children, obesity, alcohol, overweight, etc. [3]. 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, contour, texture, 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. Some of the characteristics of malignant tumors are: clustered calcification, isolated ducts, poorly defined mass, etc. [2].

In order to diagnose breast cancer, there are currently four main methods used to distinguish benign lumps from malignant
Approach used in [7] combines two methodologies, fuzzy Wisconsin Breast Cancer (WBC) database was 95.75%. Another highest level of accuracy ever using different methods for [6] improves the accuracy up to 97.34%. Until this success the Hybrid system for diagnoses of the breast cancer represented in [11] with just 4% misclassified data. Victor in paper [11] gives one solution to computerized tool used for diagnose of breast cancer. Fact that fuzzy logic can significantly help in diagnostic of breast cancer is proposed in [12,17,18,19,20].

In this paper is presented a novel approach to automatically detect the breast cancer. The proposed approach utilizes fuzzy c-means clustering for classification of the data from the WBCD database, and pattern recognition method. The rest of paper is organized as follows. Section 2 describes data set of breast cancer disease. Section 3 presents main feature of fuzzy c-means algorithm for classification as well as main features of pattern recognition method. Section 4 presents new model and shows results and finally section 5 conclude the paper.

2. DATA SET OF BREAST CANCER DISEASE

The Wisconsin breast cancer diagnosis (WBCD) database [14] is the result of the efforts made at the University Of Wisconsin Hospital for accurately diagnosing breast masses based solely on an FNA test. This dataset is created by Dr. William H. Wolberg from University of Wisconsin Hospitals.

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

This dataset contains a total of 699 clinical instances, with 458 benign and 241 malignant cases. Each clinical instance has 9 attributes with assigned integer values ranging from 1 to 10 and one class output with a binary value of either 2 or 4, indicating benign and malignant breast cancer diagnoses, respectively. The physical meaning of the nine attributes is shown in table 1. Among the 699 clinical instances, 16 instances are each missing one of the nine attributes. The database itself consists of 683 cases, with each entry representing the classification for a certain ensemble of measured values.

Common practice is to eliminate all individuals from analysis for whom information on a variable is missing. For a consistently high accuracy, the 16 instances each missing one attribute are removed from this dataset. The resulting dataset has 683 clinical instances, with 444 (65.01%) benign and 239 (34.99%)
malignant diagnoses. The evolutionary experiments performed fall into three categories, in accordance with the data repartitioning into two distinct sets: training set and test (or evaluation) set. The experimental categories are: (1) data set which contains all 683 cases of the WBCD database (2) training set that contains 400 cases, 200 benign and 200 malignant cases, while the test set is empty; (3) testing set contains 58% of the WBCD cases with 29% of benign and 29% malignant.

In the last two categories, the choice of training-set cases and test set cases are done randomly, where data for benign case are not overlapped.

### 3. Fuzzy C-means and Pattern Recognition

Complexity of medical problems has showed that using traditional methods in solving this issue is not appropriate. In medicine, the lack of information, and its imprecision, and, many times, contradictory nature is common facts. Fuzzy logic plays an important role in medicine for diagnosis of the disease. Some examples showing that fuzzy logic crosses many disease groups are the following [21]:

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, or Prostate 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

#### A. Fuzzy c-means algorithm

Clustering is a process of grouping a data in clusters, where data placed in one cluster are more similar to each other than those in other clusters. In this task choosing cluster centers is crucial to the clustering. Fuzzy c-means algorithm uses the reciprocal of distances to decide the cluster centers [15].

Fuzzy logic was introduced by Zadeh during 1960s for handling uncertain and imprecise knowledge in real world applications [13]. FCM method was developed by Dunn and improved by Bezdek. With fuzzy c-means centroid of a cluster is calculate as mean of all points, weighted by their degree of belonging to the cluster. The application of this method causes a class membership to become a relative one and one data can belong to several classes at the same time with different degrees, which represents an important feature for increasing sensitivity in medical diagnostic problems.

Advantages of this algorithm are that this method gives better results than k-means algorithm, and each data is assigned membership to every cluster and as results one data may belong to more than one cluster center. Furthermore the greatest advantage of using fuzzy logic lies in the fact that scientists can model non-linear, imprecise, complex systems by implementing human experience, knowledge and practice as a set of inference (or fuzzy) rules that use linguistic (or fuzzy) variables [11]. Disadvantages of this method are: apriority defined number of clusters, and Euclidean distance measures can unequally weight underlying factors.

In the 70’s, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise. FCM method is based on minimization of the function:

$$ J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} \frac{1}{m-1} \frac{1}{m-1} \| x_i - c_j \|_m^{m-1} $$ (1)

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

$$ u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( \frac{\| x_i - c_k \|_m}{\| x_i - c_j \|_m} \right)^{m-1}} $$ (2)

Where centers of the clusters can be calculated as follows:

$$ c_j = \frac{\sum_{i=1}^{N} u_{ij} x_i}{\sum_{i=1}^{N} u_{ij}} $$ (3)

Algorithm performs calculation as follows:

1. Initialization of \( U = [u_{ij}] , U^{(0)} \)
2. Calculation of centers of the vectors \( C^{(k)} = [c_j] \ and \ U^{(k)} \)
\[
    c_j = \frac{\sum_{i=1}^{N} u_{ij} x_i}{\sum_{i=1}^{N} m u_{ij}}
\]

3. Update of \( U^{(k)} \) to \( U^{(k+1)} \)

\[
    u_{ij} = \frac{1}{C \sum_{k=1}^{m} \left( \left\| x_i - c_j \right\|^2 \right)^{m-1}}
\]

4. Comparison, if absolute value \( \| U^{(k+1)} - U^{(k)} \| < \varepsilon \), where \( \varepsilon \) represents predefined criteria, then STOP, otherwise return to step 2.

Fuzzy sets, fuzzy membership functions, and fuzzy rules form the elemental components of the fuzzy logic decision making systems.

**B. Pattern recognition model**

Pattern recognition can be defined as a process of identifying structure in data by comparison to known structure, where the known structure is developed through methods of classification [13]. In pattern recognition system each input data is assigned to one of the clusters obtained from the process of classification. Input data are often divided in training data and test data, where train data are included in process of classification and test data are new input data in system and assigned to one of the clusters. Problem that exists in pattern recognition is to collect input data and classify in known patterns. Known patterns are represented as typical classes [13]. Similarity is expressed by membership function. M typical patterns can be expressed as fuzzy sets \( A_i \) on \( X \), where \((i=1,2,3,\ldots,m)\). New data is presented as \( B \). So, problem is now to find which \( A_i \) and \( B \) most closely matches. To solve this issue it is needed to introduce fuzzy vectors. New fuzzy vectors can be noted as \( a=(a_1, a_2, a_3, \ldots, a_n) \) where \( 0 \leq a_i \leq 1 \). Inner product (or maximum of minimum) of these vectors can be expressed as union of intersection of fuzzy vectors:

\[
    a \bullet b^T = \bigcup_{i=1}^{n} \left( a_i \cap b_i \right)
\]

and fuzzy outer product can be expressed as:

\[
    a \oplus b^T = \bigcap_{i=1}^{n} \left( a_i \cup b_i \right)
\]

Where \( b^T \) is a transpose of a fuzzy vector. If two vectors are identical, the inner product metric will yield a maximum value, and if the two vectors are completely dissimilar the inner product will yield a minimum value [13]. These two norms, the inner product and the outer product, can be used simultaneously in pattern recognition studies because they measure closeness or similarity.

Now fuzzy vectors can be extended to the fuzzy sets \((A,B)\). Following expressions describe two metrics to assess the degree of similarity of the two sets:

\[
    (A, B)_1 = (A \bullet B) \cap (A \oplus B)
\]

\[
    (A, B)_2 = \frac{1}{2} \left[ (A \bullet B) + (A \oplus B) \right]
\]

When values \((A,B)\) approaches 1 two fuzzy sets \( A \) and \( B \) are more similar, apart when they approach 0, when they are considered that they are “more far apart”[13].

Represented model shows pattern recognition when there is only one known pattern. Usual problem in pattern recognition is comparison of data to a number of known patterns. One most common metric in solving this problem is that first determine the approaching degree value for each pairwise comparisons, and after to choose a pair with the largest approaching degree. The known pattern is involved in maximum approaching degree value is then the pattern the data sample most closely resembles in a maximal sense. This method is called maximum approaching degree [13].

If there are \( m \) known patterns considered problem can be expanded as:

\[
    (B, A_j) = \max \left\{ (B, A_1) \left| (B, A_2) \ldots (B, A_m) \right. \right\}
\]

4. Results

In this study, 683 clinical instances in the Breast Cancer Wisconsin (Original) Database were used for this model. Even original data from [14] has 699 clinical instances 16 were removed because of missing of one or more attributes. In rest 683 clinical instances there are 444 benign which represents 65% and 239 malignant breast cancer cases which represents 35%. The class output of an original binary value was 2 or 4 indicating benign and malignant breast cancer, respectively. Estimation of error of this model is done using two approaches.

In Fig.1 is shown scheme of the model and steps performed during the evaluation of the results. In Table2 are shown results using this method. Fuzzy c-means is done with initial data set as well as training data. Pattern recognition is done with two
clusters and test data. These two new clusters represent the results of process of classification of training data with Fuzzy c-means classification algorithm.

![Diagram](image)

Figure 1. Model of Fuzzy c-means and Pattern recognition

Note that 683 clinical instances in the dataset are divided in 11 columns like is shown in Table 1. From this data firstly is removed first column which represent ID. Identification number of clinical instance is not important for classification.

Also last column indicating benign or malignant data is exempted. New formed data is classified into two clusters with a success of 100% true positive, 87% true negative, 0% false positive, 13% false negative after eighteen iteration of algorithm.

Stopping criteria for the algorithm was when \( \| U^{(k+1)} - U^{(k)} \| < 10^{-6} \). Cluster 1 contained 225 data for benign instances and cluster 2 contained data for malignant instances with 458 data.

In the second approach from 683 clinical instances, 400 training data is randomly chosen. The first data training data has 400 clinical instances, 200 benign and 200 malignant breast cancer cases randomly chosen from the input clinical instances.

These data are run through Fuzzy c-means algorithm with stopping criteria for algorithm \( \| U^{(k+1)} - U^{(k)} \| < 10^{-6} \) again. After ten iterations the same classification results returned as before. From this point it is natural to conclude that as number of data is smaller, convergence of the algorithm is faster. As a result from this step two clusters are formed, one with 174 data in the cluster, so called cluster 1, and cluster 2 with 226 data.

To use the clustering information in the diagnosis of the disease, from 683 data, 400 data is used for test data. 400 test data consists of 200 data for benign instances and 200 data for malignant instances randomly chosen, but different from training data. Some test data for malignant instances are not unique, because in whole data set 239 malignant instances are present and some of these are repeated. So, 58% WBCD cases with 29% of benign and 29% malignant are used as a test data.

In pattern recognition method clusters from FCM classification of training data are used as known patterns, since the success is 100 % and they can be considered as highly reliable, in testing the new formed test data. Pattern recognition method consists of calculating minimum of the distances between one test data and all members of the two clusters obtained in the clustering stage. As a result which distance is smaller test data is recognized to belong to that cluster. Distance calculated using the two metrics in (8) and (9) in the above.

Success of disease diagnosis is 100% true positive, 80.5% true negative, 0% false positive, 19.5% false negative.

<table>
<thead>
<tr>
<th>Data</th>
<th>FCM Classification</th>
<th>Disease Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive</td>
<td>100 %</td>
<td>100 %</td>
</tr>
<tr>
<td>True negative</td>
<td>87 %</td>
<td>80.5%</td>
</tr>
<tr>
<td>False positive</td>
<td>0 %</td>
<td>0 %</td>
</tr>
<tr>
<td>False negative</td>
<td>13 %</td>
<td>19.5%</td>
</tr>
</tbody>
</table>

5. CONCLUSION

Breast cancer is one of the major causes of death among women. So early diagnosis through regular screening and timely treatment has been shown to prevent cancer. Medical oncologists diagnose breast cancer based on past professional experience and knowledge which can lead to wrong diagnosis. That was the reason why many researches have started making algorithms and model for precise diagnosis.

In this article, is introduced new alternative approach for breast cancer disease diagnosis and classifying benign and malignant breast cancer using fuzzy c-means algorithm and pattern recognition model. This proposed approach was based on the three steps model including classification of input data, training data and test data. Breast cancer diagnosis is done on data set from UCI machine learning repository. This set contains 683 data where 65% data is for malignant cases and 35% of the data are for benign cases.

The initial data, 683 clinical instances is run through FCM algorithm with the success of 100% true positive, 87% true negative, 0% false positive, 13% false negative.

Success of disease diagnosis is 100% true positive, 80.5% true negative, 0% false positive, 19.5% false negative on the Breast Cancer (WDBC) dataset. This is a better classification rate compared to the results in the referenced papers in diagnosis of the breast cancer disease, which allows FNA with a highly accurate diagnosis percentage rate without the need for a surgical biopsy. Although a surgical biopsy results in almost 100% accuracy in diagnosing benign and malignant breast cancer, it is invasive, expensive, and inconvenient for the breast cancer patient. FNA with visual interpretation varies from 65% to 98% in accuracy.

It may be noted that a single fuzzy c-means algorithm with pattern recognition is not always sufficient for real-world pattern classification problems. While each approach is very simple and has some drawbacks as discussed above, this model can facilitate the doctor to detect breast cancer in the early stage of diagnosis process.

For future research, it is possible to further enhance the model accuracy of benign and malignant breast cancer diagnosis by increasing the number of instances in the model, changing stopping criteria in fuzzy c-means algorithm, etc. Also for future
research, this method can be extended to include parameters like the number of invaded axillary nodes, calcifications, disease extent, and usage of the other artificial intelligence devices such as neural networks. These will improve the prognostic risk estimation.

6. REFERENCES


