Evaluation of Breast Cancer Tumor Classification with Unconstrained Functional Networks Classifier

Emad A. El-Sebakhy, Kanaan Abed Faisal, T. Helmy, F. Azzedin, and A. Al-Suhaim

Department of Information & Computer Science
College of Computer Science & Engineering
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia
dodi05, faisal, helmy, fazzedin, suhaim@ccse.kfupm.edu.sa

Abstract

This paper proposes functional networks as an unconstrained classifier scheme for multivariate data to diagnose the breast cancer tumor. The performance of this new technique is measured using two well known databases under the minimum description length criterion, the results are compared with the most common existing classifiers in both computer science and statistics literatures. This new classifier shown reliable and efficient results with better correct classification rate, and much less computational time.

Keywords: Pattern Classification; Functional Networks; Breast cancer detection; Minimum Description Length.

1. Introduction

Breast cancer has become a major cause of death among women in developed countries [2]. The most effective way to reduce breast cancer deaths is detect it earlier. However, earlier treatment requires the ability to detect breast cancer in early stages. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones. Thus, finding an accurate and effective diagnosis method is very important. Biopsy is the best way to accurately determine whether the tumor is benign or malignant. However, it is invasive and expensive, and positive findings at biopsy for cancer are low, between 10% and 31% [1, 10, 16].

Recently, several researchers have used statistical and artificial intelligence techniques to successfully "predict" breast cancer [14, 15]. Basically, the objective of these prediction techniques is to assign patients to either a "benign" group that does not have breast cancer or a "malignant" group that has strong evidence of having breast cancer. Thus, breast cancer diagnostic problems are basically in the scope of the widely discussed classification problems [13]. The Breast Cancer Diagnosis (BCD) problem has attracted many researchers in computational intelligence, data mining, and statistics fields [6]. Artificial neural networks (ANNs) [9] and support vector machines [5, 11] have been recently proposed as a very effective method for pattern recognition, machine learning and data mining. Generally, linear/nonlinear discriminant analysis (LDA) and multinomial logistic regression (MLR) are the most commonly used data mining techniques to construct classification models. ANN and radial basis function (RBF), probabilistic neural network (PNN), and SVM became efficient alternatives in modeling classification problems due to their capability to capture complex nonlinear relationships among variables.

The rest of the paper is organized as follows: Section 2 introduces the meaning of breast cancer and its existence world wide. Functional networks are introduced in Section 3 with general methodology, including their selection of initial architecture and learning methods. The maximum likelihood functional networks classifier methodology is explained in Section 4. Implementation and a comparative study are investigated using the two Wisconsin breast cancer databases are presented in Section 5. Section 6 contains both conclusion and future work.
2 Breast Cancer: An Overview
Breast cancer is the most common cancer disease among women, excluding nonmelanoma skin cancers. The information about the tumor from certain examinations and diagnostic tests are gathered using staging to determine how widespread the cancer is. The stage of a cancer is one of the most important factors in selecting treatment options, and it uses the Tumour, Nodes and Metastasis (TNM) system, which in turn determines treatment recommendations RWWD03. Breast cancer is a malignant tumor that has developed from cells of the breast. Breast cancer is cancer of breast tissue. Worldwide, it is the most common form of cancer in females, affecting approximately 10% of all women at some stage of their life in the Western world. Although significant efforts are made to achieve early detection and effective treatment, about 20% of all women with breast cancer will die from the disease. Although scientists know some of the risk factors (i.e. ageing, genetic risk factors, family history, menstrual periods, not having children, obesity) that increase a woman’s chance of developing breast cancer, they do not yet know what causes most breast cancers or exactly how some of these risk factors cause cells to become cancerous. Research is under way to learn more and scientists are making great progress in understanding how certain changes in Deoxyribonucleic acid (DNA) can cause normal breast cells to become cancerous.

The information about the tumor from certain examinations and diagnostic tests are gathered using staging to determine how widespread the cancer is. The stage of a cancer is one of the most important factors in selecting treatment options, and it uses the Tumour, Nodes and Metastasis (TNM) system, which in turn determines treatment recommendations RWWD03. The TNM system is specific for each type of cancer, it is a standardized way in which the cancer care team describes the extent to which the cancer spread, where the letter T followed by a number from 0 to 4 describes the tumor’s size and spread to the skin or chest wall under the breast, the letter N followed by a number from 0 to 3 indicates whether the cancer has spread to lymph nodes near the breast, and the letter M followed by a 0 or 1 indicates whether or not the cancer has spread to distant organs. Once a patient’s T, N, and M categories have been determined, this information is combined in a process called stage grouping to determine a woman’s disease stage. This is expressed in Roman numerals from Stage 0 (the least advanced stage) to Stage IV (the most advanced stage).

3. Functional Networks
Functional networks are a generalization of neural networks that combine both knowledge about the structure of the problem, to determine the architecture of the network, and data, to estimate the unknown functional neurons [3, 4] and [7, 8]. We call the node $X_j \in \mathbf{X}$, for all $j$ as a multiple node, if it is an output of more than one neural functions. Otherwise, it is called a simple node.

As it is shown in Figure 1, a functional network consists of: a) several layers of storing units, one layer for containing the input data ($x_i; i = 1, \ldots, 4$), another for containing the output data ($x_7$) and none, one or several layers to store intermediate information ($x_5$ and $x_6$); b) one or several layers of processing units that evaluate a set of input values and delivers a set of output values ($f_i$) and c) a set of directed links. Functional networks extend neural networks by allowing neural functions $f_i$ to be not only true multiaargument and multivariate functions, but to be different and learnable, instead of fixed functions. In functional networks, the activation functions are unknown functions from a given family, i.e. polynomial, to be estimated during the learning process. In addition, functional networks allow connecting neuron outputs, forcing them to be coincident. Some differences between a neural network and a functional network are shown in Figure 1. Functional networks methodology can be more easily understood by organizing it into seven steps as it is shown in [7, 8].

4. Constrained-Functional Networks Classifier
We applied functional networks with maximum likelihood technique, [8] to investigate the breast cancer tumor. Generally, we assume that the initial functional network model is written as

$$\pi_{ik} = h_k(x_i, \Theta_k) = p(g_k(x_i, \Theta_k)), \quad (1)$$

![Figure 1. (a) A Neural Network, (b) A functional network.](image-url)
for all $i = 1, \ldots, n$, where $x_i = (x_{i1}, \ldots, x_{ip})$ is the $i^{th}$ observation. The vector $\Theta_k = [\theta_{1k}, \ldots, \theta_{pk}]^T$ is the vector of the functional network parameters needed to be learned. The functions $h_k(x_i, \Theta_k)$ is to be estimated based on the available breast cancer data, bearing in mind the probability restrictions on $h_k(x_i, \Theta_k)$. The functions $g_k(x_i, \Theta_k)$ are unknown, but unrestricted functions to be learned from the data, and the function $p(.)$ must satisfy the probability conditions. Since functional networks do not make any assumptions about the function $g_k(x_i, \Theta_k)$, then it can be known or unknown in form and/or in the parameters. It could be linear or non-linear. However, since we approximate $g_k(x_i, \Theta_k)$ by a linearly independent family, which depends on parameters $a_{ski}$, [8]. Figure 2 shows the architecture of the corresponding functional network model (1).

![Figure 2. Unconstrained-functional networks classifier topology with $p$ Predictors.](image)

Follow both learning techniques and classification criterions in [8] and select the best network model according to the following criterions:

1. **Correct Classification Rate (CCR) and Average Squared Classification Error (ASCE):**

   \[
   CCR = \frac{\sum_{k=0}^{c-1} CC_k}{n}; \quad ASCE = \frac{\sum_{k=0}^{c-1} [n_k - CC_k]^2}{n},
   \]

   where $n_k$ is the number of observations in class $k$, and $CC_k$ is the number of correctly classified observations in the class $k$. The best functional network is the one with both highest CCR and smallest ASCE. We construct the confusion matrix, which is a $c \times c$ matrix, its diagonal contains the number of correctly classified observations, $CC_k$, and the off-diagonal elements are the number of misclassified observations, $mck$, for $k = 0, \ldots, c - 1$.

2. **Computational cost** (Time of execution): It is the time needed to execute the classifier till obtaining the best model in both calibration and validation. The less computation cost is the better classifier.

3. **The Minimum Description Length (MDL) criterion:** As explained in [8], and then the best model is the one with the smallest MDL value. The form of the description length for the classification problem using the functional network is defined as

   \[
   L(\Theta_k) = \frac{m \log(nk)}{2} + \frac{n k}{2} \log \left( \frac{1}{nk} \sum_{i=1}^{n} \epsilon_i^2(\Theta_k) \right),
   \]

   for all $k = 0, \ldots, c - 1$, where $m$ and $k$ are the number of elements in the family and the category levels, respectively. We note that the principle $L(\Theta_k)$ is the code length of the estimated parameters $\Theta_k$, $\forall k = 0, 1, 2, \ldots, c - 1$. We note that the description length has two terms:

   (a) The first term $\frac{m \log(nk)}{2}$ is a penalty for including too many parameters in the functional network model.

   (b) The second term $\frac{n k}{2} \log \left( \frac{1}{nk} \sum_{i=1}^{n} \epsilon_i^2(\Theta_k) \right)$ measures the quality of the functional network model fitted to the training set.

   Therefore, the best model is the model with the smallest value of its description length. MDL is the best model performance.

5. **Implementation and Comparative Study**

   We apply the constrained-functional networks explained in section 4 to Wisconsin Diagnostics Breast Cancer (WDBC) databases and compare its performances to the most common classification methods in both computer science and statistics literatures discussed in Section 1. The databases can be downloaded from the machine learning repository database at university of California, Irvine 1. All the computations are implemented using MATLAB V6 under Pentium IV personal computer with a clock speed of 2.4 GHz.

   As in commonly done, we normalize the input variables to make sure that they are independent of measurement units. Thus, the predictors are normalized to interval of $(0, 1)$ using the formula:

   \[
   x_i^{new} = \frac{x_i^{old} - x_{i1:n}}{x_{n:n} - x_{1:n}},
   \]

   where $x_{i:n}$ is the $i^{th}$ order statistics of $x_1, x_2, \ldots, x_n$. We use the **stratified Sampling** technique to make sure that we get the same proportion from each group as in the original data, as recommended by [12]. We randomly hold a total

---

of \( k = \text{round} \left( \frac{n}{5} \right) \) or \( k = \text{round} \left( \frac{n}{10} \right) \) observations, with \( k_l = \text{round} \left( k + \frac{n_l}{n} \right) \) observations from the class \( l \), where \( n_l \) is the number of observations of the given data set in group \( l \), for \( l = 0, 1, \ldots, c - 1 \). Therefore, to evaluate the performance of each classifier on a real-application, we use either 5-fold or 10-fold cross-validation. Thus, we fold the given data into 5 or 10 parts, and we use 0.8 or 0.7 of the data for the learning the classification model (building) and 0.2 or 0.3 for external validation (testing). Both description and relevant work on each data set under study are represented below:

1. **Wisconsin Diagnostics Breast Cancer (WDBC):** This database is created by William H. Wolberg at University of Wisconsin at Madison. There are 683 data points in this database of which 357 are benign cases and 212 are malignant cases. We note that in this database that for each observation, there are 30 featured variables. These features are computed from digital images of fine needle aspirates (FNA) of breast masses. These features describe the characteristics of the cell nuclei in the image. The author of this database considered 10 real-valued features for each cell nucleus: (i) radius (mean of distances from center to points on perimeter); (ii) texture (standard deviation of gray-scale values); (iii) perimeter; (iv) area; (v) smoothness (local variation in radius lengths); (vi) compactness \((\frac{\text{perimeter}^2}{\text{area}} - 1)\); (vii) concavity (severity of concave portions of the contour); (viii) concave points (number of concave portions of the contour); (ix) symmetry; and (x) fractal dimension (coastline approximation \(-1\)). They computed the mean, standard error, and worst mean (the mean of the three largest values) of each feature. This process resulted in 30 feature variables for each image. For instance, field 1 is the Mean Radius, field 11 is the Standard Error of Radius, and field 21 is the Worst Radius.

2. **Wisconsin Breast Cancer (WBC):** The source of this database is the University of Wisconsin Hospital, Madison. There are 683 data points in this database of which 444 are benign and 239 are malignant. WBC is a nine-dimensional data set with the following features: (i) Clump thickness; (ii) Uniformity of cell size; (iii) Uniformity of cell shape; (iv) Marginal adhesion; (v) Single epithelial cell size; (vi) Bare nuclei; (vii) Bland chromatin; (viii) Normal nucleoli; and (ix) Mitoses.

Several researchers studied WDBC database and proved that the best three attributes are mean texture, worst mean area, and worst mean smoothness [7]. Based on these three features they create separating hyperplane that uses multi-surface method-tree (MSM-T) to construct a decision tree. They reported that the best correct classification rate is 97.5%. This estimate was obtained using a repeated 10-fold cross-validation method, their CCR values are 93.5% and 95.9%, respectively.

We are utilizing both internal and external Validation Techniques as it is shown in [8]. We repeat the estimation and validation processes for \( N = 1000 \) times, then compute all the quality measures explained in Section 4 for all classifiers. Next, we summarize the results by computing the average, the standard deviation, and the coefficient of variation of each quality measure over these 1000 runs. In addition, we draw two graphs: One for the mean of CCR versus its standard deviation over the 1000 runs, and the other for the mean of the ASCE versus mean of MDL. These graphs help us to decide which classifier is better in its performance. In both plots, each classifier is represented by a symbol. In the graph of the average of CCR versus its standard deviation, a good classifier should appear in the upper left corner of the graph. In the graph of average of MDL versus average of ASCE, a good classifier should appear in the bottom left of the plot. In addition, corresponding to these graphs, we summarize the results in Tables. In these Tables, the highest CCR’s are given in boldface [7].

For the sake of simplicity and space, we did the implementations for two predictors (feature variables and three predictor). We did the implementations and analysis with ten predictor variables (high-dimensional data) to check the performance of the functional network classifiers against other proposed classifiers.

1. **WDBC database with two predictor variables:** As we can see from the scatter plot of the data in Figure 3, there is some overlap between the two groups. The correlation coefficient between the two variables is 0.3295.

![Figure 3. WDBC Data: The scatter plot of Mean Texture and Mean Radius](image)

The quality measures and the corresponding scatter plots are shown in Table 1 and Figure 4, respectively.
### Table 1. WDBC Data: The External Validation Results with 2 Predictors

<table>
<thead>
<tr>
<th>Classification Method</th>
<th>No. Parameter</th>
<th>MDL mean</th>
<th>MDL StDev</th>
<th>Time of Exec. mean</th>
<th>Time of Exec. StDev</th>
<th>CCR mean</th>
<th>CCR StDev</th>
<th>ASCE mean</th>
<th>ASCE StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>3.000</td>
<td>-674.541</td>
<td>35.332</td>
<td>0.275</td>
<td>0.069</td>
<td>0.889</td>
<td>0.017</td>
<td>1.044</td>
<td>0.317</td>
</tr>
<tr>
<td>Linear Disc. Anal.</td>
<td>3.000</td>
<td>-759.904</td>
<td>41.170</td>
<td>0.002</td>
<td>0.069</td>
<td>0.884</td>
<td>0.017</td>
<td>0.981</td>
<td>0.335</td>
</tr>
<tr>
<td>KNN</td>
<td>3.000</td>
<td>-728.713</td>
<td>39.586</td>
<td>0.122</td>
<td>0.018</td>
<td>0.875</td>
<td>0.017</td>
<td>1.247</td>
<td>0.344</td>
</tr>
<tr>
<td>SVM</td>
<td>3.000</td>
<td>-561.948</td>
<td>39.586</td>
<td>0.375</td>
<td>0.038</td>
<td>0.759</td>
<td>0.033</td>
<td>5.148</td>
<td>1.316</td>
</tr>
<tr>
<td>PNN</td>
<td>2.000</td>
<td>-764.531</td>
<td>35.888</td>
<td>5.705</td>
<td>0.294</td>
<td>0.885</td>
<td>0.015</td>
<td>1.661</td>
<td>0.337</td>
</tr>
<tr>
<td>RBFN</td>
<td>2.000</td>
<td>-640.306</td>
<td>177.133</td>
<td>34.893</td>
<td>11.790</td>
<td>0.773</td>
<td>0.171</td>
<td>10.147</td>
<td>15.113</td>
</tr>
<tr>
<td>FFN</td>
<td>4.000</td>
<td>1.095</td>
<td>-826.403</td>
<td>34.387</td>
<td>2.728</td>
<td>0.892</td>
<td>0.013</td>
<td>0.981</td>
<td>0.238</td>
</tr>
</tbody>
</table>

### Table 2. WDBC Data: The Internal Validation Results with 2 Predictors

<table>
<thead>
<tr>
<th>Classifier</th>
<th>COR-OBST mean</th>
<th>COR-OBST StDev</th>
<th>MISS-OBST mean</th>
<th>MISS-OBST StDev</th>
<th>CCR mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogReg</td>
<td>196.000</td>
<td>344.000</td>
<td>16.000</td>
<td>57.000</td>
<td>0.994</td>
</tr>
<tr>
<td>Linear Disc. Anal.</td>
<td>212.000</td>
<td>344.000</td>
<td>0.000</td>
<td>20.000</td>
<td>0.992</td>
</tr>
<tr>
<td>KNN</td>
<td>210.000</td>
<td>354.000</td>
<td>2.000</td>
<td>3.000</td>
<td>0.991</td>
</tr>
<tr>
<td>SVM</td>
<td>162.000</td>
<td>352.000</td>
<td>50.000</td>
<td>7.000</td>
<td>0.968</td>
</tr>
<tr>
<td>PNN</td>
<td>201.000</td>
<td>350.000</td>
<td>11.000</td>
<td>3.000</td>
<td>0.990</td>
</tr>
<tr>
<td>RBFN</td>
<td>209.000</td>
<td>354.000</td>
<td>3.000</td>
<td>3.000</td>
<td>0.990</td>
</tr>
<tr>
<td>FFN</td>
<td>202.000</td>
<td>351.000</td>
<td>10.000</td>
<td>6.000</td>
<td>0.986</td>
</tr>
<tr>
<td>FNBF-MLE</td>
<td>4.000</td>
<td>1.095</td>
<td>-826.403</td>
<td>34.387</td>
<td>2.728</td>
</tr>
</tbody>
</table>

For the internal validation purpose, we summarize the output in Table 2.

2. **WDBC database with both three and ten predictor variables**: Figure 5 is a three-dimensional scatter plot of the data. As we can see, there is some overlap between the two groups. The pairwise correlation coefficients are: \(r_{1,2} = 0.3435\), \(r_{1,3} = 0.0775\), and \(r_{2,3} = 0.2091\).

The reason for choosing the high-dimensional data is to check the performance of the functional network classifiers against other proposed classifiers, and check the measure of quality. The quality measures and the corresponding scatter plots are shown in Tables 1 and Figure 6, respectively.

From the above two tables 1, 3, and 2 with the corresponding figures of average of CCR versus its standard deviation, we observe, for example the following:

1. The two classifiers: support vector machines, and
Feedforward neural networks are the worst performance. Moreover, the performance of SVM becomes even worse with the high-dimensional predictors. The linear discriminant analysis classifier, and radial basis functions networks are the second worst average CCR.

2. Both SVM and neural networks classifiers work well for internal validation with small number of predictors. In the case of high-dimensional data, their performance is affected, and the CCR values becomes less. The Feedforward neural network has by far the highest execution time. Yet, it has the second worst performance among all ten classifiers.

3. With the low-dimensional data, both Feedforward neural networks (FFNs) and support vector machines (SVMs) classifiers are the highest value of CCR. The functional networks classifier (MLE) is the second highest values of the average CCR in the low-dimensional data, but they are the highest in the high-dimensional data. Their performance is stable in both dimensions. The logistic regression, radial basis functions network and K-nearest neighbor classifiers perform more or less the same.

4. The functional networks classifier (MLE) is giving the highest values of the average CCR in the high dimensional data with less time of computations.

5. The functional networks classifier (MLE), logistic regression are giving the highest values of the average CCR. In addition, for the case of high-dimensional data, functional networks classifier (MLE) surpasses logistic regression.

6. All other classifiers perform more or less the same, with and logistic regression, linear discriminant analysis, K-nearest neighbor classifier probabilistic neural network, radial basis function, and functional networks classifiers (MLE).

7. The functional networks classifier (MLE) gives both smallest MDL and smallest ASCE. In addition, its execution time is much lower than its corresponding with both feedforward neural networks and support vector machines.

We draw our conclusion in section 6 by utilizing the useful information shown in both tables 1 and 3.

6. Conclusion and Future Work

We conclude that in the case of internal validation, the support vector machine and feedforward neural networks are the worst performance. Both feedforward neural networks and the constrained functional networks (FN-MLE) are the highest values of the average CCR. On the other hand, for the external validation, the results show that the linear discriminant analysis and probabilistic neural network classifiers give the lowest CCR value among all classifiers. The new functional networks classifier classifier with $q = 2$ has the highest CCR values, both smallest MDL and smallest ASCE with execution time much lower than its corresponding with both feedforward neural networks and support vector machines.

We showed empirically that the proposed approach has better performance, high quality and generalization than common existing approaches, with much lower computational cost. However, more work is needed in evaluating the performance of the proposed method on other medical and other science or business databases.

Acknowledgments

The authors wish to acknowledge the support and facilities provided by King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia in the development of this paper.

References


