**IJCTA,** 9(40), 2016, pp. 111-117


# Selection of Features using F-Score Method Classification of Breast Cancer Dataset using WEO Classifier

**P. Suganya* and C.P. Sumathi****

***Abstract :*** Data Mining is basically forecasting various decisions based on the information provided to predict future trends and behavior. Data Mining is highly accurate to knowledge driven decision in the field of scientific, physiology, sociology and business decision. The features of the datasets are important which decide the performance of the classification algorithms. The feature selection algorithm is an optimization technique which is used to remove the irrelevant features from the data set to improve the efficacy of various models. The categories of feature selection algorithm fall into filter, hybrid and wrapper to produce useful subsets of prediction. In feature selection the features are discriminated using various measures and they can be deployed in medical data. Breast cancer is the second leading cause of cancer deaths especially for women. The basic aim of this article is to predict the Breast cancer data set using different classifier models. The breast cancer occurs when a cell in the breast undergoes a change and solid masses can be non-cancerous tumor or they may be Breast cancer. There are nine attributes in the data set which represent cytological characteristics of breast fine aspirates with two classes with one being malignant and other being benign. The Wisconsin Breast Cancer Data (WBCD) set is an imbalanced data set with 694 samples out of which 250 samples are benign and 44 samples are malignant. The data set is balanced with class balancer to undergo feature extraction using FScore method and to classify them. The classification results have indicated that the network gave the good diagnostics performance of 99.27%.

***Keywords:*** Breast cancer, Support Vector Machine, Logistic Regression, Multi-Layer Perceptron, F-Score.

## 1. INTRODUCTION

Cancer is a general term used to refer a condition where the body cell begins to grow and reproduce uncontrollable ways. Cancer is a serious health problem. There are hundreds of different types of cancer. Commonly breast, prostate, lung, bowel cancer, skin are the various types. The biggest risk of developing cancer is age, smoking, drinking alcohol, obesity poor diet, lack of exercise etc.

Breast cancer is caused when abnormal tissue in the breast begins to multiply. The good news is that if is detected earlier can be treated successfully. Breast cancer is most type of cancer in women. The incident of breast cancer rises after age of 40. Ninety percent of breast cancer is adenocarcinomas, which arise from glandular tissue. The earliest form of the disease ductal carcinoma in situation comprises about 15-20% of all breast cancers and develops in milk ducts. Cancer that begins in the lobes or lobules is called lobular carcinoma and is found in the breast. The risk factors of breast cancer are unknown although studies suggest that estrogen, the female hormone produced by the ovaries is involved. About 5-10% of all breast cancers are thought to be related to genetic predisposition. The signs and symptoms of breast cancer can also be caused by other health conditions. A lump in the breast and the irregular shape are the most common first signs of breast cancer. Itching of the breast or nipple may be a sign of inflammatory and the

* Ph.D Scholar, Bharthiar University, Coimbatore-641046, TamilNadu, India.

** Associate Prof. & Head, Dept. of Computer Sci, S.D.N.B Vaishnav College for Women, Chromepet, Chennai -44. *suganyadgvc@gmail.com*

skin of the breast may become dimpled or puckered which shows the sign of breast cancer. Breast cancer can be examined by physical exam, ultrasound testing and biopsy. Treatment of breast cancer depends on the type of cancer and its stages and may involve radiation or chemotherapy.

## 2. RELATED WORK

The SVM and modified fuzzy c-means clustering has been used in classifying the brain tumor. The algorithm uses the concept of highly compressed data for the diagnosis and detection of brain tumor. The similar features are grouped by SVM which makes high dimensionality [1].The two transformation frequencies are used for the classification problem. The wavelet transformation is much better than the Fast Fourier transformation for the segmentation problem which yields better result [2]. In diagnosing the Parkinson disorder various classification approaches are used to yield a better result. The logistic regression showed a better accuracy in case of discrimination among the IPD from APS and MSA from PSP at various levels. Comparison for accuracy with different network based algorithm was taken for the detection of Parkinsonism [3]. The four various cardiovascular disease predictions was observed with different classifiers such AdaBoost, Naïve Bayes, SVM and LR. The preprocessing and feature selection was done with weka tool for performance analysis. The AdaBoost was comparatively higher than the other classifier giving 98% accuracy whereas the LR being the least with 68% which showed the under fitting of data [4]. The Deep belief network path [6] tested on the Wisconsin Breast Cancer Dataset which produced an accuracy of 99.68% proved to be an effective classification model. Leave One out cross validation was used [7] used for the MRI images of Breast cancer with statistical significance values .Various varied features was evaluated with $p<0.05$ and the predictive model was able to predict the overall accuracy. The article [8] aims to support the oncologists for the classification problem with breast cancer. Feature extraction has been carried out and the accuracy has been resulted to 89.3% to 64.7% as benign/ malignant. In [10] the study was to evaluate the number of events per variable (EPV) for simulating the data of cardiac patients with various iterations. The result showed that the regression coefficients were analyzed with precision, bias and significance testing. The bias was both positive and negative directions showing 90% confidence limits for regression coefficient. In [13] the output vector is defined by fuzzy class membership values and the classification is based on multilayer perceptron model. The membership values and the weights are calculated using back propagation method which reduces the learning error rate. The data taken was speech recognizer where the models are compared for performance analysis. In [16] s a feed forward neural network is designed and the Back propagation algorithm is used to train the network which produced 99.28% for six neurons. The [17] work claims that combination five features *i.e.* clump thickness, uniformity of cell size, Bare Nuclei, Bland Chromatin, Normal Nuclei derived using CSF shows a better performance with 94.29% of classification accuracy.

## 3. METHODS

The classifiers used in this article are SVM, Logistic regression, Multi-Layer perceptron and Weighted Empirical Optimization. The different algorithms perform well with numeric values for the dataset.

### Support Vector Machine

It is a supervised learning algorithm used for classification problem. It is used to find a hyper plane that separates the classes minimize the training error and maximize the margin in order to increase the generation capability. SVM learns [9] the unknown and nonlinear dependency which is a mapping function for the high dimensional input vector $x$ and output $y$. The training dataset from Eqn. (1) is subjected to distribution free learning concept where I = 1 to l , l denoting the number of training data pairs which is equal to the training data set D and Y denote the desired output .

$$D = \{(x_i, y_i) \in X \times Y\} \quad (1)$$

1. SVM learns from parameter functions from experimental data.
2. It uses normal probability distribution law which follows Gaussian distribution for an inference from the data.
3. The normal distribution makes the sum of errors squares cost function to be much reduced with maximum likelihood.

$$g(x) = w^{T}x + b \tag{2}$$

From Eqn. (2) the $x$ describes data points, $w$ is a coefficient vector and $b$ shows offset from the origin.

**Logistic regression**

The logistic regression [11] model describes the integral component of any data analysis concerned with describing the relationship between a dependent variable and one or more explanatory variables. Goal of any analysis using the method, build a model to find the best fitting and most parsimonious where the outcome variable in logistic regression is binary or dichotomous. Logistic regression uses equation for all the attributes and it finds the expected class. The best attribute is chosen from the equation and it makes the classifier trained. Under fitting problems in classification has low accuracy with logistic regression.

The relation between the independent and dependent variable is given by the mean value of the outcome variable. The quantity E(Y/$x$) denotes the conditional mean where $y$ is the predicted variable and $x$ denotes the independent variable. When the error terms are normally distributed the maximum likelihood which yields values for the unknown predictors in turn maximizes the probability of obtaining the observed set of data. The likelihood equations are as follows:

$$\sum[y_i - \pi(x_i)] = 0 \tag{3}$$

$$\sum x_i[y_i - \pi(x_i)] = 0 \tag{4}$$

The Eqn. (3) and Eqn.(4) are varied different ranges of instances of records in the data. The equations are iterative procedures to maximize the likelihood of observed predictions.

**Multi-Layer Perceptron**

Multi-Layer perceptron has computational intelligence which used back propagation technique. Various features are chosen at each level of training which yields a good accuracy. The [12] single layer perceptron can only represent linear decision surfaces. The target functions in nonlinear decision surfaces can be used as an input for the multilayer perceptron at various levels. The sigmoid activation function is used for nonlinear output function. They produce difference gradient descent of weights to differentiate the error functions. The derivative for sigmoid function as follows:

$$\frac{d\sigma(z)}{dz} = \sigma(z).(1 - \sigma(z)) \tag{5}$$

The above Eqn. (5) is a sigmoid function to derive the back propagation problem. Target function is output unit should have activation 0.9 if it corresponds to correct classification. Otherwise output unit should have activation 0.1

The Eqn. (6) shows the gradient descent '$w$' to minimize the sum squared error over training examples and '$k$' is the index value of E.

$$E(w) = \frac{1}{2}\sum_{k=0}^{m}(t^k - 0^k)^2 \tag{6}$$

**Weighted Empirical Optimization**

The best optimized data [5] is achieved through various fitness value and the error matrix or confusion matrix is generated .This result is compared with the various performance measures of various classifiers.

**Step 1:** To accumulate $\sum_{1}^{n} f(x)$ no of various frequency data.

**Step 2:** Evaluate the data with weightage constrains [Decide on which constraint is needed and evaluate it from the rest of it]

**Step 3:** To arrive at the optimized data regarding the value of fitness for each weightage constraints. (*i.e.* : $f(x)$ )

**Step 4:** To remember the best optimized data through its fitness value and store it in the given WEO($x$).

**Step 5:** Repeat the Step 3 and 4 again, until the data regarding the $f(n)$ is complete.

**Step 6:** Exchange the data of weightage in the given WEO($x$) to determine the optimal decision making.

## 4. PREPROCESSING

### Experimental Data

The Wisconsin Breast Cancer Data set consists of two class attributes. The dataset is taken form UCI repository which has 699 records with nine attributes. The attributes are 1) Clump thickness, 2) Uniformity of cell size, 3)Uniformity of cell shape, 4) Marginal adhesion, 5) Single epithelial cell size 6) Bare nuclei 7) Bland chromatin 8) Normal nucleoli, 9) Mitosis. The class value has two hundred and forty one records (65.5%) which are malignant and four hundred and fifty eight records (34.5%) which are benign.

### Data Cleaning

The data collected has to be cleaned where missing values are identified. Several methods are adopted to solve missing values in which the missing values can be replaced or deleted. Some of the missing values are substituted using median value for each attribute. In breast cancer dataset 16 missing values for Breast cancer data set are removed in this article. Normalization helps in training time for the classifiers. The Min Max Normalization is used to transform all values of the attributes between 0 and 1. The Min Max Normalization [14] applies linear transformation on the raw data, keeping the data values in the same range. The Min Max formula is written in Eqn. (7).

$$V^1 = (v - \text{Min}(v(i))/(\text{Max}(v(i) - \text{Min}(v(i))) \quad (7)$$

Where $v$ is the observed value, Min ($v(i)$) is the minimum value for a particular attribute, Max ($v(i)$) the maximum value in a particular attribute and $v^1$ is the standard value of an attribute. Using [15] data mining techniques the missing values are filled up, thus by calculating the average of all available values for an attribute and replace the missing instance with the average. Using neural network models the inputs are normalized from maximum value of an attribute and dividing the rest of the values by this maximum value. The normalized data turn to be either zero or one.

### Architecture of the Proposed System

The architecture of the proposed system is shown in Fig (1). The mean value is computed and set as threshold value. The feature value which is greater than the threshold value is selected. The feature values below the mean F-Score are removed. The selected features constitute the resultant dataset and are involved in computing the classification accuracy by using various cross validation procedure and benchmarked algorithms.

## 5. ANALYSIS OF RESULTS

In this article various network models for classification of breast cancer data set was used. The proposed model learns through different iteration process with weight adjustments to the initial weights of the stages. After each iterative process the fitness weight function usually sums the overall. These weights decide to optimize the data through fitness weight constraint. The best weight is assigned and stored in

hash table. The iteration completes until the rest of the instances for all. Finally exchange of weights correlates with the hash table to determine the best solution from the attributes. The simulation results are measured among the various performance measures for the benchmarked classifier algorithms.

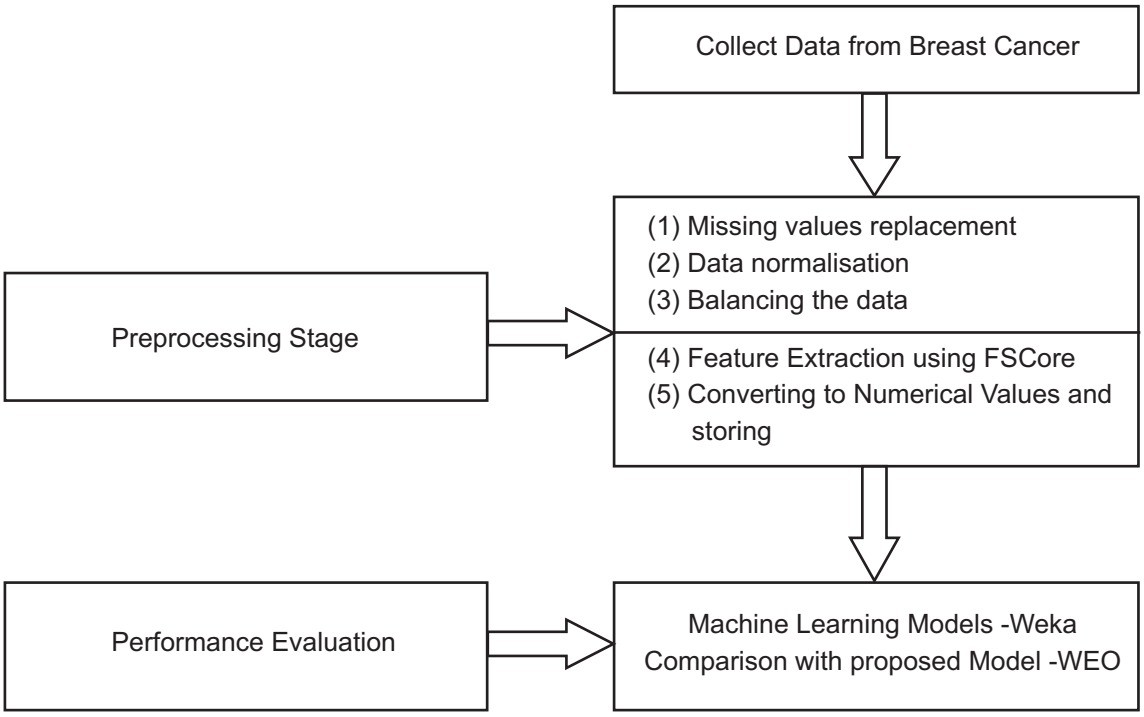


**Figure 1: Block Diagram of Proposed System**

**Table 1**

**Details of features obtained after feature selection method (Improved FScore)**

| *Dataset* | *Instances* | *Original Features* | *Classes* | *After Applying Feature Selection (Number of Features)* |
|---|---|---|---|---|
| Breast Cancer | 694 | 10 | 2 | 6 |

**Table 2**

**Classification Accuracy with various percentage Splits**

| *Method with FScore Feature Selection* | *Testing and Training split Percentage (with Accuracy)* | | |
|---|---|---|---|
| | 50-50% | 60-40% | 80-20% |
| SMO | 89.65 | 90.60 | 90.75 |
| Logistic Regression | 88.70 | 88.70 | 89.00 |
| MLP | 88.56 | 91.65 | 93.00 |
| WEO | 91.60 | 93.56 | 95.60 |

Table 2 shows the comparison of performance with varying training and testing partitions. In SMO classifier there is 89% for 50-50 % training test partition, 90% for 60-40% training test partition, 90% for 80-20% training test as accuracy evaluation. In Logistic Regression classifier there is 88% for 50-50 % training test partition, 88% for 60-40% training test partition, 89% for 80-20% training test as accuracy evaluation. In MLP classifier there is 88% for 50-50 % training test partition, 91% for 60-40% training test partition, 93% for 80-20% training test as accuracy evaluation. In WEO classifier there is 91% for 50-50 % training test partition, 93% for 60-40% training test partition, 95% for 80-20% training test as accuracy evaluation.

**Table 3**

**Classification Accuracy with Balancing and imbalancing in prediction Class**

| *Classifier Name* | *Imbalanced Dataset* | *Balanced dataset* |
|---|---|---|
| SMO | 89.58 | 89.51 |
| Logistic Regression | 84.29 | 87.76 |
| MulitlayerPerceptron | 88.58 | 88.55 |
| WEO | 93.30 | 94.65 |

**Table 4**

**Classification Accuracy after applying FScore Feature Selection**

| *Classifier Name* | *Accuracy* | *Sensitivity* | *Specificity* |
|---|---|---|---|
| SMO | 91.84 | 0.87 | 0.67 |
| Logistic Regression | 90.84 | 0.87 | 0.77 |
| MulitlayerPerceptron | 91.12 | 0.89 | 0.85 |
| WEO | 94.65 | 0.95 | 0.87 |

From the above Table 4, the various classifiers are compared using confusion matrix and the accuracy of proposed model showed an accuracy of 94.65% which is relatively higher than the other classifiers. The stability of the model with 0.95 measurements for the new proposed model shows a good sign with concern to other classifiers.

## 6. CONCLUSION

The Breast Cancer dataset is imbalanced and it is balanced using smote technique for further classification. In classifier models the dataset showed different accuracy and the proposed model has good performance than the other classifier models. With FScore as the feature selection technique the dataset is again processed for further performance measures. The important attributes are chosen using FScore technique which yielded six attributes as the best one for the classification. The model proposed exhibited an accuracy of 94.65% with 0.95 as its sensitiveness.

## 7. REFERENCES

1. Nichat, Aparna M., and S. A. Ladhake. “Brain Tumor Segmentation and Classification Using Modified FCM and SVM Classifier” *Brain* 5.4 (2016)
2. Srinivas, Mr Veerabathini, Mrs V. Rama, and CB Rama Rao. “Wavelet Based Emotion Recognition Using RBF Algorithm.” *brain* 4.5 (2016).
3. Tripathi, Madhavi, et al. “Automated differential diagnosis of early Parkinsonism using metabolic brain networks: a validation study.” *Journal of Nuclear Medicine* 57.1 (2016): 60-66.
4. Dominic, Vinitha, Deepa Gupta, and Sangita Khare “An effective performance analysis of machine learning techniques for cardiovascular disease.” *Applied Medical Informatics* 36.1 (2015): 23.
5. P.Suganya,Dr.C.P.Sumathi. “Weighted Empirical Optimization Algorithm to Classify Parkinson Disease” *Autralian Journal of Basic and Applied Sciences* June 2016.
6. Abdel-Zaher, Ahmed M., and Ayman M. Eldeib. “Breast cancer classification using deep belief networks” *Expert Systems with Applications* 46 (2016): 139-144.
7. Sutton, Elizabeth J., et al. “Breast cancer molecular subtype classifier that incorporates MRI features” *Journal of Magnetic Resonance Imaging* (2016).
8. Diz, Joana, Goreti Marreiros, and Alberto Freitas. “Applying Data Mining Techniques to Improve Breast Cancer Diagnosis” *Journal of Medical Systems* 40.9 (2016): 203.
9. Kecman, Vojislav. “Support vector machines–an introduction.” *Support vector machines: theory and applications* Springer Berlin Heidelberg, 2005.1-47.

10. Peduzzi, Peter, et al. “A simulation study of the number of events per variable in logistic regression analysis.” *Journal of clinical epidemiology*49.12 (1996): 1373-1379.

11. Hosmer Jr, David W., and Stanley Lemeshow *Applied logistic regression* John Wiley & Sons, 2004.

12. Cristianini, Nello, and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods* Cambridge university press, 2000.

13. Pal, Sankar K., and Sushmita Mitra. “Multilayer perceptron, fuzzy sets, and classification” *IEEE Transactions on Neural Networks* 3.5 (1992): 683-697.

14. Paulin, F., and A. Santhakumaran. “Classification of breast cancer by comparing back propagation training algorithms” *International Journal on Computer Science and Engineering* 3.1 (2011): 327-332.

15. Janghel, R. R., et al. “Intelligent decision support system for breast cancer” *International Conference in Swarm Intelligence* Springer Berlin Heidelberg, 2010.

16. Paulin, F., and A. Santhakumaran. “Back propagation neural network by comparing hidden neurons: case study on breast cancer diagnosis”*International Journal of Computer Applications* 2.4 (2010): 40-44.

17. Mitra, Malay. “A Neural Network Based Intelligent System for Breast Cancer Diagnosis.” (2013).