A Dynamic System for Diagnosing of Breast Cancer based on CF Algorithm

B. Gomathy [1], Shanmugam [2], S.M. Ramesh [3]

[1.] Research Scholar, Anna University, Coimbatore
[2.] Professor, Department of Electrical and Communcation Engineering, Bannari Amman Institute of Technology, Erode, India.
[3.] Assistant Professor Senior Grade, Department of Electrical and Communication Engineering, Bannari Amman Institute of Technology, Erode, India.

Abstract: Disease diagnosis and prognosis are two medical applications pose a great challenge to the researchers. The application of data mining and machine learning techniques has revolutionized the whole process of disease diagnosis. This paper intends to propose a hybridized approach, Correlative Framework (CF) algorithm for the implementation of diagnosing breast cancer, the most scare away disease. The major objective of this paper is to analyze and predict the vulnerability of disease in a patient. The association rule of data mining is deployed to correlate the interesting relation from the huge medical database. In this proposal, Linear Discriminant Analysis (LDA) is exploited for feature selection. Initially, base rule is generated, then for each rule feature support is computed which is followed by confidence. On the basis of feature supporting the base rule, supplementary rule and positive rule is generated. Regression is utilized for each rule with its corresponding feature value. Classification result is obtained based on minimum and maximum of residual support values. The significant performance of our newly devised algorithm is evaluated using confidence metrics. Experimental results expose that prediction level of our adduced work is more factual than other existing algorithms.

[B. Gomathy, Shanmugam, S.M. Ramesh. **A Dynamic System for Diagnosing of Breast Cancer based on CF Algorithm.** *Cancer Biology* 2017;7(2):95-103]. ISSN: 2150-1041 (print); ISSN: 2150-105X (online). http://www.cancerbio.net. 13. doi:10.7537/marscbj070217.13.

Keywords: Association rule, attribute selection, machine learning, mead linear regression, nedler, pruning, regression.

1. Introduction

Data mining technology provides a user oriented approach to the novel and hidden patterns in the data. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. However, these patterns can be utilized for clinical diagnosis. Since the available raw medical data are widely distributed, voluminous and heterogeneous in nature. These distributed data need to be collected in a well organized form. To form a hospital information system, these collected data can be then integrated. Health care organizations must have the ability to analyze data. Plenty of treatment records for millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently.

There exist many different kinds of breast cancer, prolong with spread or different stages, aggressiveness and genetic makeup. Detection in earlier stage will increase the survival rate for breast cancer. Nowadays, it is quite common that medical practitioner uses different classifier model to diagnose the diseases. To find the patterns in data, models can be designed using data mining. Though, implementation of mass screening would result in increased caseloads for radiologist to identify the breast cancer. Probably, this will increase chances of improper diagnosis. Research work is carried out for solid data analysis to detect such disease using contrary data mining algorithms. Association rule of data mining is used to analyze massive datasets. Early diagnosis requires an accurate and a reliable diagnostic procedure that allows the physicians to distinguish benign breast tumors from malignant ones.

An effective CF algorithm is presented in this paper to enhance the detection accuracy in breast cancer. At the initial stage, the dataset is formulated by choosing the features. For feature selection, Linear Discriminant Analysis (LDA) is used. Preprocessing should be carried out on maintaining dataset prior to feature selection. Since, a dataset may hold some anomalies like missing values, noisy data, etc. It is intended to apply preprocessing to remove such anomaly data. This preprocessing work will enhance the association rule mining to discover medically significant rules by assigning weights from a huge set of medical datasets. The offered CF algorithm divides the process into two scenarios.

In the first scenario, list of features have been chosen and a set of labels is assigned to selected features. Base rule is framed by collecting the frequency of each attribute, higher the frequency of all attribute becomes the base rule. For each rule, feature support is calculated. Following this rule, framework for positive and supplementary rules is collected. The feature supporting the base rule becomes positive rule, whereas the features not supporting the base rule becomes supplementary rule. Regression is exploited for each rule with its corresponding feature value.

Forward to the second scenario, Hessian matrix (square matrix) computed for finding second order partial derivative of chi square. The threshold is set to filter the items in a data set. Patterns are classified from the obtained threshold values. Based on the positive and negative results produced, diagnosing of disease is effectuated.

In our method, CF algorithm is presented which process any given data set using base rule. This process highly minifies the work of medical practitioner and radiologist from manually determining the massive data set, thus reduces the time required to process the items of a given dataset.

The rest of this paper is composed as follows: Section 2 describes some reviews related to our work. Section 3 presents our proposed work, including the description of CF algorithm. Section 4 experimentally evaluates the performance of our proposed work. Finally, Section 5 sums up the paper with some conclusion and future enhancements.

2. Related Works

Searching of related information about the enhancement technique for the diagnosis of breast cancer is an increasing criterion. Numerous works have been carried out to explore the analysis of most dreadful diseases. Here, some works of various scholars which are related to our proposal is presented.

There are many techniques to predict and classification of breast cancer pattern. A work proposed by [M. Karabatak, *et al*], presented an automatic diagnosis system for detecting breast cancer based on association rule (AR) and neural network (NN). In this system, AR is used for reducing the dimension of breast cancer database and NN is used for intelligent classification. The performance of this system is evaluated using a test stage criterion, 3-fold cross validation method applied to the Wisconsin breast cancer database. Research works shows that hybridized approach of association rule and neural network model can be used to obtain fast automatic diagnostic systems for other threatens diseases. A rule based reasoning algorithm was presented by [T. Singh, et al] for mammographic findings to provide support for the clinical decision to perform biopsy of the breast. The most important component of this system is the rule to decide whether a new case is similar to a case in the database. This case based reasoning algorithm is straightforward; it selects the matched test case from all cases in the database. The malignancy fraction is computed as the number of selected cases that have malignant outcomes divided by the total number of selected cases.

[K. J. Schilling, *et al*] found an exciting technology Computer-aided detection (CAD) with full field digital mammography (FFDM) in detection of breast cancer. However, this technology possesses several advantages over screen film mammography, observed higher contrast resolution, lower noise and better dynamic range. CAD with FFDM showed a high sensitivity in identifying cancers manifesting as calcifications and masses. Several data mining techniques have proposed and applied for the diagnosis of breast cancer. [M. A. Karaolis, *et al*] analyzed a new data mining system for classification of myocardial infarction (MI), percutaneous coronary intervention (PCI), and coronary artery bypass graft surgery (CABG) events based on decision trees. Experiments were carried out, and the study showed that the system achieved 66% of accurate classification for the MI, 75% of exact classification for the PCI, and 75% of perfect classification for the CABG events. [*S. Gupta, et al*] presented a paper on breast cancer diagnosis and explore the data mining techniques offer great promise to uncover patterns hidden in the data that can help the physicians in decision making. Knowledge Discovery in Databases (KDD) used in this paper focus the process of turning the low level data into high level knowledge. The prognostic problem is mainly analyzed under ANN (Artificial Neural Networks) and its accuracy came higher in comparison to other classification techniques. The paper organized by [E. A. El-Sebakhy, *et al*] proposes functional networks as an unconstrained classifier scheme for multivariate data to diagnose the breast cancer. The functional classifier depicted for classification shown reliable and efficient results with better correct classification rate and less computational time.

[M. M. A. Abdelaal, *et al*] framed a technique and provided a promising results for increasing diagnostic accuracy of classifying the mammographic mass features along with age that discriminates true and false cases. There are three classification techniques for extraction of the most important mammographic mass features. It comprises the classification of SVM over decision trees witnessed by minimal error measures and maximum average gain. Apart from the above three classification technique used, a neural network support vector machine method for classification was proposed by [Sudhir D, *et al*] for diagnosis of breast cancer. Here, SVM is used only for classification rather than functional approximation. Kernel Adatron algorithm implemented by SVM, maps inputs to a high dimensional feature space and separates data optimally into their respective classes by isolating those inputs. The efficiency of manual detection of breast cancer is 85% and the efficiency of SVM recognition obtained is nearly 97%. Another method proposed by [Jamarani S. M. h, *et al*] employed combination of artificial neural networks (ANN) and multi wavelet packet based sub band image decomposition for early breast cancer diagnosis. The system identifies micro calcification clusters in digitized mammographic images by decomposing the mammograms into different frequency sub bands and suppressing low frequency sub band to construct the mammogram from sub band containing only high frequencies. Hybrid method was evaluated by [S. Şahan, *et al*] using Wisconsin Breast Cancer Dataset (WBCD) and the results showed that this method has accurate detection that other methods proposed before the hybrid method. On the other hand, a hybrid model developed by [J. P. Choi, *et al*] combined the artificial neural network and the Bayesian network to obtain good estimation of prognosis in breast cancer. This hybrid approach achieved a prediction accuracy of 87.2% with a sensitivity of 93.3% and a specificity of 83.1%. The accuracy of neural network in predicting the breast cancer is also evaluated by [M. Lundin, *et al*] [C.L. Chi, *et al*] [H. B. Burke, *et al*]. Analyzing and determining risk factors in medical data set was a mind-numbing work, which consumes time. To overcome this problem, [W. G. Baxt, et al] used artificial neural network to classify the 356 record of patients with acute coronary occlusion. This detection system used back propagation of neural network to train the data set. This model accurately determined 80% of patients with infraction. [P. M. Ravdin, *et al*] followed this work, and used neural networks to predict 1008 clinical data set of breast cancer patients.

[M. U. Khan, *et al*] predict the breast cancer survivability using fuzzy decision trees was intended. This system is efficient alternative to crisp classifiers that are applied independently. To illustrate the performance, different combination like number of decision tree rules, inference techniques, types of fuzzy membership functions are evaluated. For this purpose, consider a collection of SEER breast cancer data set is taken which is more comprehensible source of information on cancer incidence in United States. This hybrid fuzzy decision tree classification is more robust and balanced than independently applied crisp classification. Review of methodologies for neural networks and logistics regression was presented by [S. Dreiseitl, *et al*], and results showed the performance of neural network was not significant as logical regression. Logical regression outperforms neural network, due to its property of interpretability of model parameter and easy to use. Following this work, cancer classification and prediction using logistic regression with Bayesian gene selection was proposed by [X. Zhou, *et al*]. The idea behind in this approach is in conjunction with a logistic regression model to relate the gene expression with the class labels. To discover important genes, Gibbs sampling and Markov Chain Monte Carlo (MCMC) methods are used. The same logistic regression is then used for cancer classification and prediction, once the important genes are identified. The accuracy of the classification based on this method is high. A rough set data reduction was used by [A. E. Hassanien, *et al*] to find class labels for classification. Rough set approach based rules were evaluated and compared with the ID3 classifier algorithms. Results showed that classification accuracy of Rough set is much better. Consecutively, [H.L. Chen, *et al*] proposed a rough set based supporting vector machine classifier for breast cancer diagnosis. RS reduction algorithm is employed as feature selection tool to remove the redundant features and further improve the diagnostic accuracy by SVM.

As by [M. Karabatak, *et al*], the classification accuracy is examined on Wisconsin Breast cancer dataset. Another feature selection method presented by [H.M. Lee, *et al*] is an efficient fuzzy classifier that selects the feature based on fuzzy entropy measure. The patterns are classified based on partition the pattern space into non overlapping decision region. Since, the decision regions do not overlap thus the complexity and computational load of the classifier are reduced which then leads the training time and classification time as extremely short. The performance of this classifier is measured by considering the Iris database and Wisconsin breast cancer database. This classifier works well for the pattern classification application. Nelder Mead's simplex method presented by [L. Riggs, *et al*] has a great deal of large scale optimization. It shows its ability in many real applications which other algorithm or other analytical methods may be ineffective because of their computing cost and complexity. The method achieves efficiency in terms of success rate and computing time.

3. Proposed Methodology

In this section, proposed CF algorithm for accurate diagnosis of breast cancer is analyzed. The two scenarios taken for processing of this algorithm is elaborated with detailed system design. Association rules (AR) employed in this work identify the relations among variables in huge medical data set. It acts as a more promising technique to enhance the diagnosing of disease. Fig 1 exposes the outer section of our proposed CF algorithm.

3.1 Attribute Selection

One of the crucial steps in classification is the feature selection. It is substantial that the model includes all relevant variables. In general, for any given set of data, more variables produce a better model fit to the data. Feature selection intends to avoid over fitting and enhance the model performance. Selecting highly informative features could enhance the accuracy of classification models. This leads to deep insight into the underlying processes that generated the data. In our algorithm, PCA (Principal Component Analysis) is used, which is a transformation based reduction, transforms the original features of the dataset with a typically diminished number of uncorrelated ones and is termed as a principal component. The integration of both PCA and LDA (Linear Discriminant Analysis) is employed in this paper in order to proficiently extract the features.

3.2 CF algorithm: Scenario 1

Initially, base rule is composed by assigning labels to the features.

3.2.1 Label Assignment

A preprocessed medical data set contains attribute set such that $A = \{a_1, a_2, a_3, \ldots\ldots\ldots, a_k\}$ representing 'k' number of attributes in the data set D. Each attribute contains set of labels assigned as $\delta_{i1}, \delta_{i2}, \delta_{i3}, \ldots \neq \infty$ where, i denotes i^{th} attribute in A. To build robust decision making tool subset, attribute selection is important. It is the most important set when analyzing a huge medical data set, which helps in predicting the outcome. Attribute selection also improves the performance of prediction of the risk level of each item $\{I_1, I_2, I_3, \ldots\ldots\ldots, I_n\}$ in data set D, since prediction may not scale up to the full attribute set A. Such attributes acted as the major role in diagnosing of diseases. Depends on our integrated attribute selection method PCA and LDA, $CF = \sum_{i=1}^{j} a_i$, a subset of attributes is obtained. A single label is chosen for each attribute in CF.

3.2.2 Base Rule Formation:

Next step is the formation of Base rules (BR) from the selected labels. Base rule is formed by choosing the n features and is evaluated as a row vector. Features supporting the base rule are taken, and for each rule feature support is calculated. The base rule representation is given as

$$BR(a_1 = \delta_{15} \,||\, a_2 = \delta_{22} \,||\, a_3 = \delta_{32} \,||\, \ldots\ldots\ldots \,||\, a_j = \delta_{j9}) .$$

The feature support is computed as,

$$S_{count} = \frac{\sum \delta p_i}{i} \qquad (1)$$

Base rule formation depends upon the decision of class labels, so labels that are chosen will act as the criteria for disease diagnosis. BR is the most preliminary rule framed by CF.

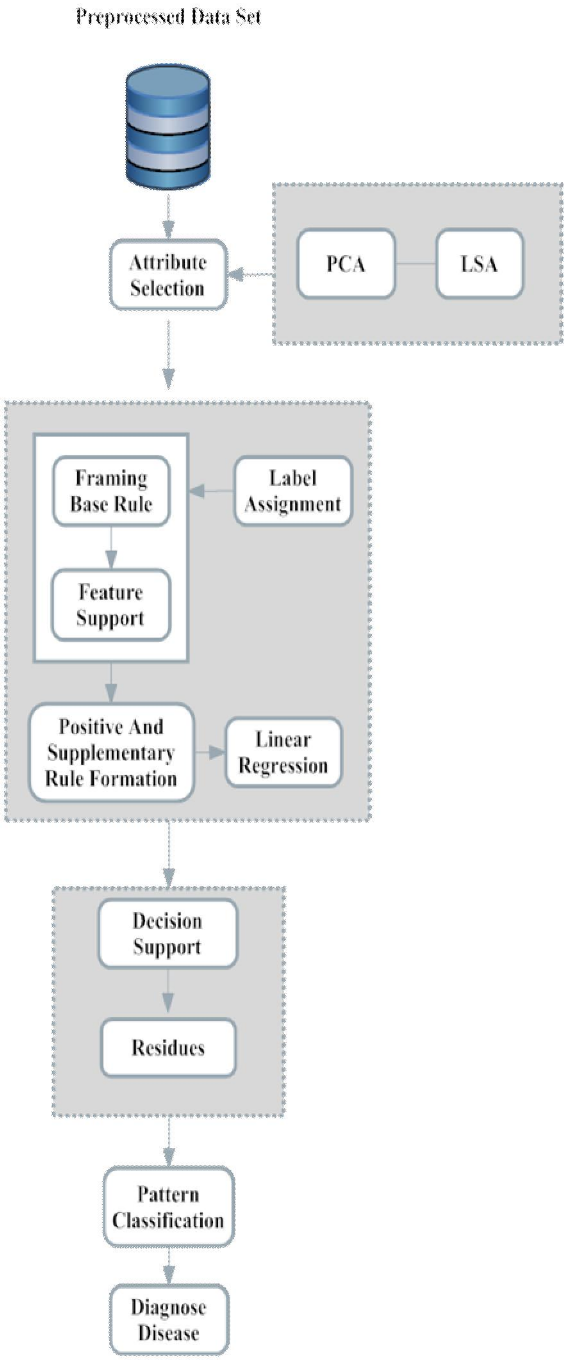

Fig1: CF Workflow Design

3.2.3 Positive and Supplementary Rule collection

Base rule is framed by collecting the frequency of each attribute, higher the frequency of all attribute becomes the base rule. Positive Rule (PR) and Supplementary Rule (SR) are the most sophisticated rules formed by our CF algorithm in an item I_n. These two rules PR and SR forms a subset of BR, represented in a mathematical form $PR \subset BR$ and $SR \subset BR$. In addition to that $PR \cup SR = BR$. As our algorithm, a feature supporting the base rule becomes positive rule, on the other hand features not supporting the base rule becomes supplementary rule. Mathematically, $a_k \exists\ BR => PR$ whereas $a_k \nexists\ BR => SR$.

3.2.4 Applying linear regression

Regression is formulated for each rule with its corresponding feature value. Essentially, Association rules are deployed in mining to discover set of rules that is shared among a large number of rules. Association rule is defined as the implication of the form L => R where $L, R \ \in I$ and $L \cap R = \emptyset$. The set of items shortly termed as item sets L and R are called antecedent (LHS) and consequent (RHS) of the rule respectively. Moreover, L => R represents x features in the rule. Regression is a statistical measure that determines the strength of the relationship between one dependent variable (denoted as y) and a series of other changing variables (denoted as independent variable). It is performed on set of x and y data through classically weighted regression, i.e. regressions in which the weighting factors are equated to the measurement errors in the dependent array (y) while another dependent array (x) is assumed to be error free. The weighted factors are derived from the measurement errors in both the dependent and independent arrays.

The weighted features denoted as w_i is equated which is of the form,

$$w_i \ => Y = \sum_{i=1}^{n} x_i \qquad (2)$$

Where,

Y is a dependent array and x_i is another dependent array with i features. Hence, the intrinsic attribute value for each rule is subjected. The regression is applied to get the result for the predictive modeling procedure.

3.3 CF algorithm: Scenario 2

Hessian matrix is computed by acquiring the second order derivative of chi square.

3.3.1 Heuristic optimization

Nelder and Mead Simplex method implement the Nelder and Mead heuristic simplex optimization procedure to minimize chi square. Chi square can be estimated as,

$$\chi 2 \ = \ \sum_{i=0}^{n-1} \frac{(\text{obs } y[i] - \text{calc } y[i])^2}{w[i]} \qquad (3)$$

Where,

$\text{obs } y[i]$ => Observed $y[i]$ values.

$\text{calc } y[i]$ => Calculated $y[i]$ values.

Nelder and Mead Simplex procedure projects to display a plot of the observed $y[i]$ values versus the calculated $y[i]$ values. It outputs the data of best estimates of both the estimated coefficients and the derived coefficients.

The best estimates obtained by nonlinear regression are calculated by assuming that the sum of squares surface approximates to one of the linear model close to the minimum. This calculation involves obtaining the co-variance matrix by inverting the Hessian matrix of second order derivatives of the weighted sum of squares. Such differentials are calculated analytically as,

$$\frac{\delta f\ (p,x)}{\delta p_i} \quad \text{And} \quad \frac{\delta f\ (p,x)}{\delta p_j} \qquad (4)$$

Where $f\ (p, x)$ denotes the fitted function.

p_i and p_j denotes the best estimate of a pair of parameter of the indices i and j. In this proposal, when the Positive definite of the hessian lies at critical point x, then the value of f (Hessian Matrix) accomplish a local minimum at x. Similarly, the value of f obtains a local maximum at x when the negative definite of hessian lies at critical point x. For instance, Hessian matrix simply performs the second order derivative test for one or two variable functions. Hessian encompasses only one second order derivative in the case of one variable function, depending on the positive and negative definite, local minimum and local maximum at x is exemplified. On the other hand, determinant is exploited by hessian while performing second order derivative test for two variable functions. Determinant is the product of the Eigen values, which accommodate either both positive/negative if the determinant is positive whereas negative determinant results the two Eigen values have dissimilar signs.

3.3.2 Decision support

Decision support is the most substantial task in prediction analysis, which is done by minimum and maximum pay off values. Prime Rule (PR_i) is devised by integrating the PR and SR to retrieve these minimum/maximum pay off values. The comparison is exhausted by PR_i features of class label with all features of original data set D. Support value is computed by counting the similar features of original data set D and prime rule. The resultant support value is then analyzed to accomplish efficacy in risk analysis. This support value computation is exemplified for all PR_i and is compared with S_{count}.

The following equation reveals the N_{count} calculation for all PR$_i$,

$$N_{count} = \sum_{i=1}^{j} N_{count} + 1 \qquad (5)$$

The above equation is computed only when $S_{count} \leq PR_i$. The minimum pay off is calculated as shown below:

$$Min_p = \frac{N_{count}}{T_n} \qquad (6)$$

Here, T_n implies the total number of transactions. Similarly, Maximum Payoff is computed using $Total_i$, which is obtained by total summation of support values with maximum S_{count}.

$$Max_p = \frac{Total_i}{T_n} \qquad (7)$$

In our proposed CF algorithm, two measures (support and confidence) are framed for mining association rules. To accomplish this task, support values for both Positive Rule (PR) and Supplementary Rule (SR) are computed independently, which represents the number of features in original dataset that support PR and SR.

The outcome of this support value for both positive and supplementary rule are acquired by comparing it with threshold values. After enforcing comparison, the rules are arranged in ascending order of data items and collect the values of x and y. The residues are collected to classify the data x and y by which the classification result produces the risk status of the disease.

4. Experimental Establishment

To illustrate the performance of our proposed CF algorithm, a data set of breast cancer is acquired, among that data set bmi, age first, brstproc, lastmamm, surgmeno are prioritized. Our newly proposed algorithm is applied on above mentioned data set and computes the feature support on the basis of LDA technique.

The LDA based preprocessing method is applied before employing the CF algorithm for the dataset. Preprocessing is the most important step in the decision making process. Since, a dataset may hold missing values, noisy data, etc. In order to remove such data, preprocessing is applied. Feature selection is the most needed step in the prediction algorithm because the accuracy of decision making depends upon the features selected. A strong prediction system is generated only through best feature selection.

There are two measures that involve statistical analysis of the data which include support and confidence. It improves the efficiency of our newly devised algorithm by reducing the number of rules generated by association rules. This measure aims to minimize the time and memory needed for decision making by reducing the number of rules. Based on the confidence, a small amount of pruning is performed in generating rule set.

After pruning, the residual support is calculated for pruned rule set. The value in residual support is taken and is compared with the threshold value. Classification result is obtained based on the minimum and maximum value of residual support values. Some of the sample association rules predicting the existence of breast cancer using a CF algorithm are depicted in table 1.

The classification performance is evaluated in terms of precision, recall and accuracy because of its clinical significance. The result of classification measure in terms of accuracy is depicted in fig 2. It is computed using equation 8.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \qquad (8)$$

Where TP, TN, FP, FN denotes true positive, true negative, false positive and false negative respectively. Each sample in the test set can be categorized in one of the four outcomes.

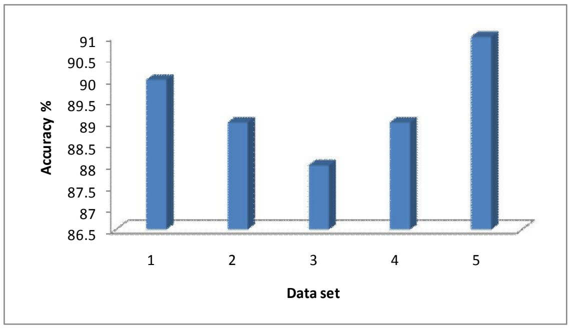

Fig 2: Classification measure in terms of accuracy

True positive are class members according to both the classifier and sample label. Likewise, true negative are non-class members according to both classifier and sample label. False negatives are samples that the classifier places outside the class, but sample labels are members. False positives are samples that the classifier places within the given class, but sample labels are non-members.

Accuracy is a percentage quantity for the number of times that the classifier is correct in its classification and it conveys the right intuition when the positive and negative populations are roughly equal in size.

$$Precision = \frac{TP}{TP+FP} \qquad (9)$$

Precision is another measure that evaluates the classification performance. It is the percentage of times that the classifier is correct in its classification of positive samples. Using the equation 9, the precision values are computed.

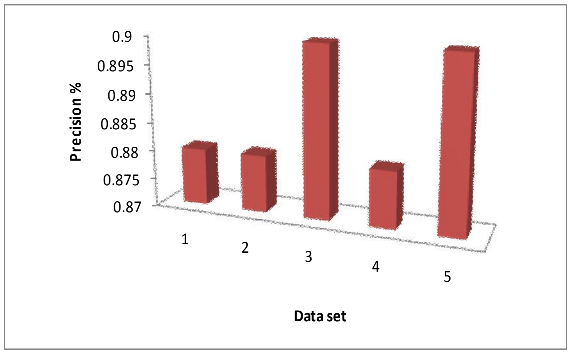

Fig 3: Classification measure in terms of precision

The result of classification measures in terms of precision is exposed in fig 3.

Recall is a measure of the ability of a predictive model to select instances of a certain class from the data set. It is commonly termed as sensitivity, which corresponds to the true positive rate. For instance, recall is the percentage of known positive samples that the classifier would classify as being positive.

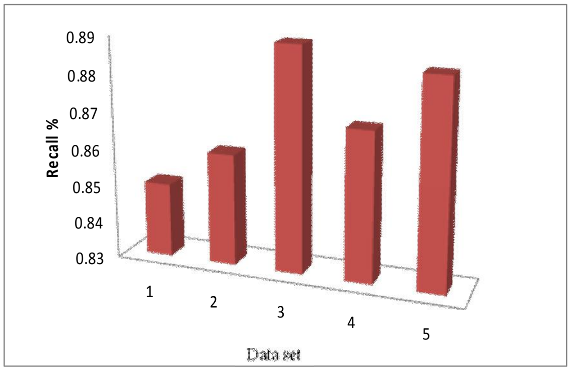

Fig 4: Classification measure in terms of Recall

$$Recall = \frac{TP}{TP+FN} \qquad (10)$$

The recall values are computed using the equation 10. The result of classification measures in terms of precision is presented in fig 4.

The average residual values are computed to analyze the classification result in disease diagnosis. The residual can be computed as follows:

$$Res_{avg} = Y - \hat{Y} \qquad (11)$$

Table 1: Sample Association rules predicting the existence of breast cancer using CF algorithm.

Items	Feature Support	Confidence	Rules	Rules after Pruning	Residual Support	Classified Result
1	94.16666667	0.431957187	brstproc:0,lastmamm:0,surgmeno:0==> bmi:9,agefirst:9,nrelbc:9,	bmi:1,nrelbc:0,brstproc:0,surgmeno:0====>agefirst:1,lastmamm:9,	0.0107	**No**
2	94.16666667	94.16666667	brstproc:0,lastmamm:0,surgmeno:0==> bmi:9,agefirst:9,nrelbc:9,	bmi:1,nrelbc:0,brstproc:0,lastmamm:0,surgmeno:0====>agefirst:2,	0.0107	**No**
3	85.33333333	42.66666667	brstproc:0,lastmamm:0,==> bmi:9,agefirst:9,nrelbc:9,surgmeno:1	bmi:1,nrelbc:0,brstproc:0,lastmamm:0,surgmeno:0====>agefirst:2,	0.0109	**No**
4	85.33333333	28.44444444	brstproc:0,lastmamm:0,==> bmi:9,agefirst:9,nrelbc:9,surgmeno:1	bmi:1,nrelbc:0,brstproc:0,lastmamm:0,surgmeno:0====>agefirst:2,	0.0109	**No**
5	93.83333333	23.45833333	brstproc:0,lastmamm:0,==> bmi:9,agefirst:9,nrelbc:9,surgmeno:9	bmi:1,nrelbc:0,brstproc:0,lastmamm:0,surgmeno:0====>agefirst:2,	0.0109	**No**
6	93.83333333	18.76666667	brstproc:0,lastmamm:0,==> bmi:9,agefirst:9,nrelbc:9,surgmeno:9	bmi:1,nrelbc:0,brstproc:0,lastmamm:0,====>agefirst:2,surgmeno:9	-0.9888	**Yes**
7	93.83333333	15.63888889	brstproc:0,lastmamm:0,==> bmi:9,agefirst:9,nrelbc:9,surgmeno:9	bmi:1,nrelbc:0,brstproc:0,lastmamm:0,====>agefirst:2,surgmeno:9	-0.9888	**Yes**
8	93.83333333	13.4047619	brstproc:0,lastmamm:0,==> bmi:9,agefirst:9,nrelbc:9,surgmeno:9	bmi:1,nrelbc:0,lastmamm:0,surgmeno:0====> agefirst:2,brstproc:1,	0.0112	**No**

In other words, residuals are computed by finding the difference between Symptoms and Symptoms predicted values in the sample data set.

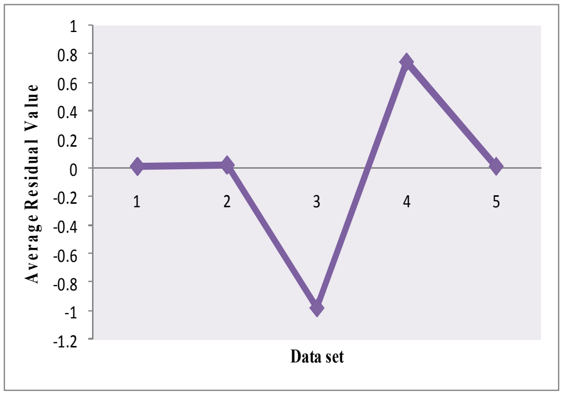

Fig 5: Computation of average residual values

Fig 5 unveils the computation of average residual values in terms of sample data set.

Finally, the regression is plotted for the experimental and calculated y values as mentioned in section 3.3.1. The approximation of regression plotted against calculated vs experimental y values using Nelder and Mead Simplex regression is exhibited in fig 6. Regression is plotted and the iteration is repeated to a maximum of 400 intervals.

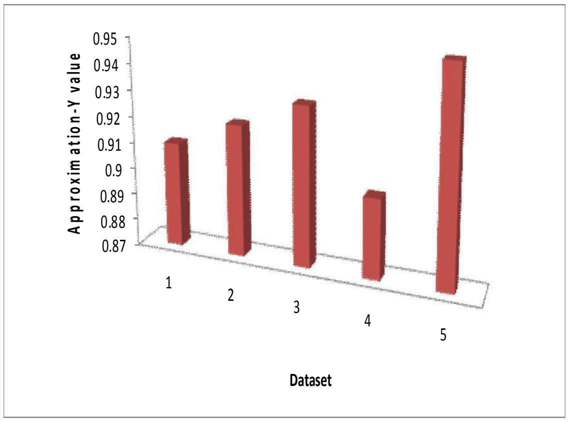

Fig 6: Dataset vs. Approximation of regression

This experimental study stipulates that our proposed CF algorithm outperform the existing model for decision making in breast cancer analysis using association rule.

5. Conclusion and Future Work

Medical diagnosis is regarded as an important yet intricate task that needs to be executed accurately and efficiently. Clinical decisions are often made based on radiologist intuition and experience rather than on the knowledge rich data hidden in the data base. This practice leads to unwanted errors and biases which affects the quality of service provided to patients. The knowledge rich environment in data mining has the potential to improve the quality of clinical decisions. In this study, dynamic system aiding radiologist for breast cancer diagnosis based on the CF algorithm is presented. The threshold is set to filter the items in the data set. Patterns are classified on having this threshold. The classification results are obtained by pruning the generated rule set and the computation of the residual values. The proposed algorithm plays an effective role in assisting radiologists with earlier detection of breast cancer.

The proposed work can be further enhanced and expanded for the automation of breast cancer prediction. Real data from Health care organizations and agencies needs to be collected and all the existing techniques will be compared for the optimum accuray.

References

1. M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis, "Assessment of the risk factors of coronary heart events based on data mining with decision trees," Information Technology in Biomedicine, IEEE Transactions on, vol. 14, pp. 559-566, 2010.
2. T. Singh, S. S. Bhadauoria, S. Wadhwani, and A. Wadwani, "Expert System Design and Analysis for Breast Cancer Diagnosis."
3. W. G. Baxt, "Use of an artificial neural network for data analysis in clinical decision-making: the diagnosis of acute coronary occlusion," Neural Computation, vol. 2, pp. 480-489, 1990.
4. E. A. El-Sebakhy, K. A. Faisal, T. Helmy, F. Azzedin, and A. Al-Suhaim, "Evaluation of breast cancer tumor classification with unconstrained functional networks classifier," in Proceedings of the IEEE International Conference on Computer Systems and Applications, 2006, pp. 281-287.
5. J. P. Choi, T. H. Han, and R. W. Park, "A hybrid Bayesian network model for predicting breast cancer prognosis," Journal of Korean Society of Medical Informatics, vol. 15, pp. 49-57, 2009.
6. M. U. Khan, J. P. Choi, H. Shin, and M. Kim, "Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare," in Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, 2008, pp. 5148-5151.
7. M. Lundin, J. Lundin, H. Burke, S. Toikkanen, L. Pylkkänen, and H. Joensuu, "Artificial neural

networks applied to survival prediction in breast cancer," Oncology, vol. 57, pp. 281-286, 1999.

8. M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," Expert Systems with Applications, vol. 36, pp. 3465-3469, 2009.
9. M. M. A. Abdelaal, H. Sena, M. Farouq, and A. Salem, "Using data mining for assessing diagnosis of breast cancer," in Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on, 2010, pp. 11-17.
10. S. Gupta, D. Kumar, and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, pp. 188-195, 2011.
11. S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," Journal of biomedical informatics, vol. 35, pp. 352-359, 2002.
12. K. J. Schilling, J. W. Hoffmeister, E. Friedmann, R. McGinnis, and R. G. Holcomb, "Detection of breast cancer with full-field digital mammography and computer-aided detection," American Journal of Roentgenology, vol. 192, pp. 337-340, 2009.
13. X. Zhou, K.-Y. Liu, and S. Wong, "Cancer classification and prediction using logistic regression with Bayesian gene selection," Journal of biomedical informatics, vol. 37, p. 249, 2004.
14. S. Şahan, K. Polat, H. Kodaz, and S. Güneş, "A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis," Computers in Biology and Medicine, vol. 37, pp. 415-423, 2007.
15. Sudhir D., Ghatol Ashok A., Pande Amol P., "Neural Network aided Breast Cancer Detection and Diagnosis", Proc. of the 7th WSEAS International Conference on Neural Networks, 2006.
16. C.-L. Chi, W. N. Street, and W. H. Wolberg, "Application of artificial neural network-based survival analysis on two breast cancer datasets," in AMIA Annual Symposium Proceedings, 2007, p. 130.
17. H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," Expert Systems with Applications, vol. 38, pp. 9014-9022, 2011.
18. L. Riggs, T. Roppel, M. Baginski, and W. Ku, "Improved Nelder Mead's Simplex Method and Applications," 2012.
19. Jamarani S. M. h., Behnam H. and Rezairad G. A., "Multi wavelet Based Neural Network for Breast Cancer Diagnosis", GVIP 05 Conference, 2005, pp. 19-21.
20. H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell, J. R. Marks, D. P. Winchester, and D. G. Bostwick, "Artificial neural networks improve the accuracy of cancer survival prediction," Cancer, vol. 79, pp. 857-862, 2000.
21. P. M. Ravdin, G. M. Clark, S. G. Hilsenbeck, M. A. Owens, P. Vendely, M. Pandian, and W. L. McGuire, "A demonstration that breast cancer recurrence can be predicted by neural network analysis," Breast Cancer Research and Treatment, vol. 21, pp. 47-53, 1992.
22. H.-M. Lee, C.-M. Chen, J.-M. Chen and Y.-L. Jou, "An efficient fuzzy classifier with feature selection based on fuzzy entropy," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 31, pp. 426-432, 2001.
23. A. E. Hassanien, "Rough set approach for attribute reduction and rule generation: a case of patients with suspected breast cancer," Journal of the American society for Information science and Technology, vol. 55, pp. 954-962, 2004.

6/25/2017