ISSN (Print): 0974-6846
ISSN (Online): 0974-5645
Indian Journal of Science and Technology, *Vol 12(8), DOI: 10.17485/ijst/2019/v12i8/141814, February 2019*

Performance Analysis of Different Classifiers in Prediction of Breast Cancer

S. Roobini and J. Fenila Naomi

Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore - 641008, Tamil Nadu, India; roobinis@skcet.ac.in, fenilanaomij@skcet.ac.in

Abstract

Objectives: The major motivation is to build the prediction model for diagnosis. The fundamental exploration of prediction is to anticipate breast cancer at a prior stage that guarantees a long survival of patients. **Methods/Statistical Analysis**: In medical field, the classification of tissues surrounding the malicious cancer cells into benign and malignant categories is extremely challenging task to predict. For diagnosis of a disease, Naive Bayesian [NB], Support Vector Machine [SVM] and Artificial Neural Network [ANN] Classification systems are investigated and Fuzzy C-Means Clustering are analyzed to make clusters. Fuzzy C-Means Clustering [FCM] algorithm clusters the data with simulated annealing which is classified using the above mentioned classifiers in furtherance of developing best prediction model with predefined rules. The performance is validated with K-fold cross validation. **Findings:** The Wisconsin Breast Cancer Dataset [WBCD] from UCI dataset storehouse is utilized to test the execution of classifiers. This dataset holds 10 properties with 699 records. This dataset has been clustered as benign and malignant. In the clusters, to achieve global optima simulated annealing technique is used and the classifiers are applied for clusters. In this examination, Fuzzy C-Means Clustering [FCM] with simulated annealing and Naive Bayesian classifier serves to be the best one with 89.2% accuracy and its F-measure is computed as 0.9417. The various performance metrics are computed for proposed novel model and its results are compared with existing values which indicates, the Naive Bayesian classifier works well for non-dependency data as there is no affinity between attributes and is considered as most noteworthy among them. **Application/Improvements:** Prediction model can be used for predicting any disease in medical field domain, which can be further improved by using Farthest First Clustering [FFC] algorithm.

Keywords: Artificial Neural Network [ANN], Breast Cancer, Classification, Fuzzy C-Means Clustering [FCM], Naive Bayesian [NB], Semi-Supervised

1. Introduction

Machine learning techniques are extensively used in medical application which includes identifying and classifying of tumours. Machine learning is upgrading characteristic, forecasting conclusions, and begins to blemish the superficial of personalized care. It is mainly used as an aid for cancer analysis and prediction. Cancer inquisitions in recent times have endeavoured to concern the machine learning methods towards cancer prediction and cast. Machine learning hushed deeply a more powerful arena because it allows choice to be made which could not be possibly made using accepted methodologies[1]. Prescient examination is one of the imperative segment regions in information mining which settlement with concentrating data from information and used to anticipate the patterns and personal conduct standards.

Prescient investigation is a prominent measurable technique which has a capability to build predictive models[2]. In data mining, breast cancer is an important research topic in medical science. In women, the most probable intrusive disease is Breast cancer, with more than one million cases and deaths occurring extensive annually.

Detecting of breast cancer at an earlier phase is an efficient way to reduce death occurring due to cancer. The objective is predicting breast cancer in an initial phase which ensures a long survival of patients. A confounded test for the primary finding makes it hard to get the last outcomes[3].

In prescient investigation, choosing the consequence of an ailment is a standout amongst the most captivating and testing undertakings. The absence of examination of right and imperative data in medicinal science is to deal with colossal measure of datasets by machine learning systems. These calculations could be utilized expressly to locate the last outcome by misusing different arrangement strategies in information mining. For predicting Breast Cancer accurately, there are various possible solutions with earlier interpretation such as supervised and unsupervised learning. Supervised Learning includes Decision tree a popular classification approaches in knowledge discovery and data mining, which classifies the labeled trained data into a tree or rules, Artificial Neural Network (ANN) is a scientific model or computational model dependent on organic neural systems, K-closest neighbor which characterizes the building model, SVM built optimal separating boundary between datasets to solve optimization problem and the construction of classification system by association rule discovery techniques. Unsupervised learning includes clustering which discovers useful patterns within the data. Semi-Supervised learning is also called as inductive learning which is deduces the exact label for unlabeled dataset[4].

The study is composed as pursues: The investigation of related specialists on the conclusion of bosom disease is displayed in area II. Area III gives brief clarification for the strategies utilized in existing and proposed models. The framework configuration is introduced in area IV. The exploratory outcomes and informational index is portrayed in Section V. At long last, Section VI gives finish of the study.

2. Literature Survey

In[5] compared various models of classification such as Bayes Net, Naive Bayesian, Sequential Minimal Optimization [SMO] for cancer prognosis. In classification technique, dimensionality reduction is used inorder to remove the features which do not contribute more or does not influence the result. Gain ratio technique is accustomed to remove the undesirable feature and ranker algorithm is influenced to rank the feature depending on the ratio values. Reduction techniques take off the features which has lowest gain ratio values. Among ten classification algorithms, Bayes net classifier provides best accuracy but time taken to accomplish the model is large.

In[6] compared various algorithms of decision tree such as ID3, CART and C4.5. ID3 uses information gain approach to resolve advisable property for each node of a decision tree which was generated. The disadvantage in ID3 algorithm is it cannot handle Continuous values, accepts only definite attribute. C4.5 is an extensibility of ID3, it depends on hunt's algorithm which can hold both definite and constant attributes to build a decision tree. Gain ratio as an feature selection part to build decision tree which removes proneness of information gain. The disadvantage is time taken to accomplish the model is too large.

In[7] diagnosed cancer by combining the approach of farthest first clustering, Outlier detection algorithm (ODA) and J48 decision tree. After clustering the data, ODA is accustomed to identify deviations within the clusters formed.

The clusters are given as input to J48 which has two parts such as tree building and pruning. The advantage is better performance which speeds clustering and outliers are removed. The limitation of this technique is expensive for estimation and time consumption to build decision tree.

In[3] proposes a hybrid approach of DT-SVM as a predictive framework for breast cancer disease. The first state is treatment of information and option extraction followed by DT-SVM hybrid model predictions. The intake features for SVM were optimized using DT algorithm. The advantage of hybrid model is to yield accurate results and robust to noise which yields a good accuracy. The disadvantage is accuracy depends on selection of kernel and computationally expensive.

In[8] described the distinguishing of different clustering techniques like FCM, K-means and EM (Expectation Maximization) cluster. FCM and K-means plays a fundamental role for intrusion detection system because clustering does not desire any labeling information. K-Means is a repetition clustering algorithm is moving an item surrounded by the set of clusters until the covet set is reached. Among them, K-means contribute superior results but FCM also provides results closer to K-means. K-means is said to be an exclusive clustering and FCM is an overlapping clustering. FCM is better for detection as it has high detection rate and low false positive rate though it is time consuming.

In[9] proposes a simulated annealing based Fuzzy Classification System (SAFCS). Initially, if-then fuzzy rules are developed and perturb operations are applied to

new fuzzy rules. SAFCS is distinguished with C4.5 which depends on entropy criteria and pruning techniques, these method of classification are applied to different datasets, among them SAFCS achieves better results in premises of accuracy for both training set and testing set. The disadvantage is execution time for prediction model and cooling rate is difficult to assess.

3. Methodologies

The following exploration methods are employed in this study.

3.1 Data Collection and Pre-Processing

The dataset is collected from UCI Machine learning data repository of Wisconsin (Original) Breast cancer dataset (WBC). WBC has 699 instances, 2 class labels (2 for Benign and 4 for Malignant) and 11 attributes. The attributes are cardinal valued. The dataset contains missing values '?'. The dataset is pre-processed by single imputation method, i.e., the replacement of mean value of a variable. The advantage is sample mean remains unchanged. The Breast Cancer dataset is provided in Table 1.

Table 1. Dataset Description

S. no	Attribute	Domain
1.	Sample Code Number	Id number
2.	Clump Thickness	1-10
3.	Uniformity of Cell Size	1-10
4.	Uniformity of Cell Shape	1-10
5.	Marginal Adhesion	1-10
6.	Single Epithelial Cell Size	1-10
7.	Bare Nuclei	1-10
8.	Bland Chromatin	1-10
9.	Normal Nucleoli	1-10
10.	Mitoses	1-10
11.	Class	2-benign, 4-malignant

3.2 Fuzzy C-Means Clustering

The Fuzzy C-Means Clustering (FCM) algorithm is a soft clustering where one data point can reside to more than one cluster. FCM is an unsupervised clustering algorithm which is enforced in agricultural engineering, astronomy, image analysis, medical diagnosis[8]. In FCM, degree of membership is designated to each data point, based on which the data points are designated to clusters.

3.3 Simulated Annealing

Simulated Annealing is a repetitive method which was inspired for annealing for metals. Simulated Annealing mainly used as an escalation search paradigm to evade from local minima and to attain global optima. SA has been extensively accustomed on a wide range of combinatorial optimization and achieves good results[10]. This optimization can be done by accepting moves which degrades the feature on a parameter called temperature. The temperature is step by step diminished by utilizing cooling plan. The conduct of a calculation closes, when the temperature scopes to zero.

The parameter required for recreated strengthening are beginning temperature, last temperature and temperature decrement. One approach to reduce the temperature is basic direct technique. The temperature decrement,

$$f(t) = t\alpha \qquad (1)$$

Where, t = time in minutes [1, ∞]
α = Cooling rate [0.5 - 0.99]

3.4 Decision Tree [C4.5]

In Classification Problem, C4.5 is a supervised algorithm that generates decision tree (DT). It is improved from ID3 algorithm by dealing with both consecutive and discrete attributes, missing values and pruning trees[6]. C4.5 builds decision trees from a set of training data by calculating the information gain for each attribute.

$$Entropy(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \qquad (2)$$

The property with most elevated data gain is taken as a root for choice tree. Data gain for each quality is calculated and sorted in descending order is shown in Table 2.

Table 2. Information Gain for Each Feature

Features	Information Gain
Clump Thickness	0.969
Bare Nuclei	0.965
Marginal Adhesion	0.962
Uniformity of Cell Shape	0.956

Normal Nucleoli	0.942
Bland Chromatin	0.918
Single Epithelial Cell Size	0.8515
Mitoses	0.8017
Uniformity of Cell Size	0.7040

3.5 Support Vector Machine

Support vector machine [SVM] is used to solve binary class of problems which maps linear into non- linear space. In furtherance to enforce mapping, kernel implementation is required[11]. The kernel functions are accustomed to train the classifier which selects the support vector for the kernel.

3.6 Artificial Neural Networks

The Artificial Neural Network [ANN] is implemented using three layer neural network of Back propagation approach. This approach has Input layer, Output and hidden layers[12]. Each layer contains an element called neurons. The neurons are associated via links. The output layer consists of 2 neurons which classifies either as benign or malignant. The back propagation approach was used to train the network, in which all the activations are calculated in forward pass. The target node is directly measured for output node by comparing the output of training set.

3.7 Naive Bayesian

In machine learning, naive bayes algorithm is considered to be a simple probability based classifier which depends on Bayes theorem with strong independence assumptions between the features[13]. Naive Bayes classifiers are immensely adaptable in which number of parameters is linear to the number of variables (features/predictors) in a learning problem. Naive Bayesian classifier depends on Bayes' hypothesis and the hypothesis of all out likelihood. The likelihood with vector x = < x1... xn> has a place with speculation h is

$$P(Y \mid X_1,..X_n) = P(X_1,..X_n \mid Y) / P(X_1, X_2...Xn) \qquad (3)$$

4. System Design

4.1 Existing System

In existing method, the combined approach such as Fuzzy C-means clustering [FCM] with simulated annealing and Decision tree (C4.5) classifier is used for diagnosis of breast cancer[14]. The pre-processed dataset is clustered using FCM algorithm. In this algorithm, 'm' is a fuzziness index whose value lies between [1,∞]. Fuzziness index, measures the tolerance of required clustering. If the value of 'm' is larger, it has larger overlapping between clusters. In general, m=1 for crisp and 2 for fuzzy clustering. In this investigation, m=1.4 is chosen as fuzzy index. The fuzzy membership degree μ_{ij}, lies between [0, 1]. Then, the clustered data is annealed for which the cooling schedule is chosen as f(t)=4. The starting and final temperature is chosen as minimum and maximum of a feature in a random manner. After clustering, C4.5 classifiers are accustomed to divide the clustered dataset and labels are predicted either as Benign or Malignant. The model is then cross validated by applying K-fold cross validation, here K=10. The existing system flows as it is being provided in Table 3.

Table 3. Existing System Diagnosis of Breast Cancer

Input: Pre-processed Wisconsin Breast Cancer data set.
Output: Benign or Malignant cancer with better accuracy
Procedure:

a. Get dataset WBCD from the UCI Machine Learning Repository.
b. Pre-processed dataset is enforced for Fuzzy C-means Clustering.
c. The clustered data is applied for simulated annealing. Again, the output is applied for FCM.
d. Repeat steps 2 -3 until minimum objective function is achieved.
e. The C4.5 classification algorithm is applied on clustered data.
f. Diagnosis of tumor patient either benign or malignant with better accuracy using 10-fold cross validation.

4.2 Disadvantages of Existing Model

The drawbacks of existing model are tree structure will be prone to sampling. Generally, trees will be robust to outliers, due to over fitting, decision tree tend not to produce greater results. Decision Tree is said to be greedy algorithm which actually produces local optima.

4.3 Proposed System

In proposed model, the pre-processed data is clustered by Fuzzy C-Means Clustering [FCM] algorithm with Simulated Annealing. Then, clustered record is classified by several classifiers such as Naïve Bayesian (NB), Support

Vector Machine (SVM) and Artificial Neural Network (ANN) models are used for diagnosis of breast cancer. In existing, Decision tree classifier was used to classify the samples due to over fitting it provides only local optima. In proposed model, after clustering some of the classifier models are used to classify the clustered dataset and labels are predicted either as Benign or Malignant. The model is then cross validated using K-fold cross validation (Here k=10). The proposed system flow as it is being provided in Table 4 and 5.

5. Experiments And Results

5.1 Evaluation Metrics

In this section, a relative report on the execution of existing and proposed grouping model is talked about dependent on Accuracy, Error rate, F - measure, exactness and review. Precision quantum's the means by which profound the settled tuples are being ordered effectively[7]. TP embodies to positive tuples and TN epitomizes to negative tuples characterized by the essential classifiers. So also FP ascribes to positive tuples and FN attributes to negative tuples which is inaccurately grouped by the classifiers.

5.2 Precision

Precision is a ratio of true positive tuples and all positive tuples in a dataset. Precision is given by,

$$\text{Precision} = TP / TP + FP \quad (4)$$

5.3 Recall

Recall is a ratio of true positive tuples against positive and negative tuples. Recall is given by,

$$\text{Recall} = TP / TP + FN \quad (5)$$

5.4 F-Measure

F - Measure is also called as F - Score. F - Measure is a mean of precision and recall. F- Measure value varies from 0 to 1. If the value of F-Measure is higher, then it is said to be a better classifier. It is given by

$$F - Measure = 2 * ((precision * recall) / (precision + recall)) \quad (6)$$

5.5 Accuracy

The classifiers accuracy is an important metric for evaluation. It is a ratio of positive tuples and negative tuples against all the tuples. It is given by,

$$Accuracy = TP + TN / TP + TN + FP + FN \quad (7)$$

Table 4. Comparison Chart for Confusion Matrix of All Classifiers

Classifiers	True positive (TP)	True Negative (TN)	False positive (FP)	False Negative (FN)
DECISION TREE	62	18	18	42
SUPPORT VECTOR MACHINE	96	6	8	30
ARTIFICIAL NEURAL NETWORK	118	0	22	0
NAÏVE BAYESIAN	125	0	0	15

Table 5. Comparison of Values Obtained For Evaluation Metrics

Methodology	Classification Accuracy (%)	Precision	Recall	F-measure	Error rate
EXISTING SYSTEM [FCM with simulated annealing and Decision tree]	57.1%	0.775	0.596	0.673	0.429
PROPOSED SYSTEM [FCM with simulated annealing and Support Vector Machine]	72.8%	0.7619	0.9230	0.8434	0.2714
Proposed System [FCM with simulated annealing and Artificial Neural Network]	84.2%	0.842	1.0000	0.914	0.158
Proposed System [Fuzzy C-Means Clustering and Naïve Bayesian]	89.2%	1.0000	0.8772	0.9417	0.108

5.6 Error Rate

The error rate is an essential measure for evaluation. Lower error rate is said to be a better classifier. Error rate determines the error between the prediction and actual. It is given by,

$$Error - rate = FP + FN / TP + TN + FP + FN \quad (8)$$

5.7 Results

In this research, 10-fold cross validation is used to validate the results. The dataset is divided into ten equal subsets randomly. One of the partition act as a testing set, whereas the rest of the partitions act as training set to train the model. A relative report on the execution of existing and proposed grouping model is talked about dependent on Accuracy, Error rate, F - measure, exactness and review. Precision quantum's the means by which profound the settled tuples are being ordered effectively[7]. TP embodies to positive tuples and TN epitomizes to negative tuples characterized by the essential classifiers. So also FP ascribes to positive tuples and FN attributes to negative tuples which is inaccurately grouped by the classifiers.

6. Conclusion

The study presents the comparative analysis of several classifiers with clustering which is used for prediction of breast cancer. The performance of Fuzzy C-Means Clustering [FCM] with Naive Bayesian classifier provides a better prediction when compared to other classifiers. Therefore, Fuzzy C-Means Clustering [FCM] with Naïve Bayesian model achieves highest accura cy with lower error rate. F-Measure value is high which also indicates Fuzzy C-Means Clustering [FCM] with Naive Bayes is a better Classifier and it is suggested as a better prediction model for diagnosis of bosom malignant growth.

7. References

1. Kharya S, Dubey D, Soni S. Predictive Machine Learning Techniques for Breast Cancer Detection. International Journal of Computer Science and Information Technologies. 2013; 4(6):1023-8.
2. Enterprise data analytics strategy: A guide for CIOs. Available from: https://searchcio.techtarget.com/essentialguide/Enterprise-data-analytics-strategy-A-guide-for-CIOs. Date accessed: 2018.
3. Sivakami K. Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model. International Journal of Scientific Engineering and Applied Science (IJSEAS). 2015; 1(5):418-29.
4. Data Mining Concepts. Available from: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004. Date accessed: 2018.
5. Garg R, Mongia V. Dimensionality Reduction and Comparison of Classification Models for Breast Cancer Prognosis. International Journal of Computer Sciences and Engineering. 2018; 6(1):308-12.
6. Singh S, Gupta P. Comparative Study Id3, Cart and C4.5 Decision Tree Algorithm: A Survey. International Journal of Advanced Information Science and Technology (IJAIST). 2014; 27(27):97-103.
7. Howsalya Devi RD, Indra Devi M. Outlier Detection Algorithm Combined With Decision Tree Classifier for Early Diagnosis of Breast Cancer. International Journal of Advanced Engineering and Technology (IJAE). 2016; 7(32):1-6.
8. Singh T, Mahajan M. Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm. International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE). 2015; 4(5):89-93.
9. Mohamadi H, Habibi J, Saniee Abadeh M, Saadi H. Data mining with a simulated annealing based fuzzy classification system. Pattern Recognition. 2008; 41(5):1824-33. https://doi.org/10.1016/j.patcog.2007.11.002
10. Simulated Annealing. Available from: https://www.geeksforgeeks.org/simulated-annealing/. Date accessed: 2017.
11. Raikwal JS, Saxena K. Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set. International Journal of Computer Applications. 2012; 50(14):35-9.
12. Mohamed Junaid KA. Classification Using Two Layer Neural Network Back Propagation Algorithm. Scientific Research. 2016; p. 1207-12.
13. Hazra A, Mandal SK, Gupta A. Study and Analysis of Breast Cancer Cell Detection using Naive Bayes, SVM and Ensemble Algorithms. International Journal of Computer Applications. 2016; 145(2):39-45.
14. Acharya S, Saha S, Thadisina Y. Multiobjective Simulated Annealing-Based Clustering of Tissue Samples for Cancer Diagnosis. IEEE Journal of Biomedical and Health Informatics. 2016; 20(2):691-8. https://doi.org/10.1109/JBHI.2015.2404971 PMid:25706936