ISSN: 0975-766X
CODEN: IJPTFI
Available Online through **Research Article**
www.ijptonline.com

PERFORMANCE COMPARISION OF NAIVEBAYES AND IBK CLASSIFIERS

P.Hamsagayathri
2/183, Thanganagaram, Shenbagapudur, Sathyamangalam-638402, Erode (Dt), Tamil Nadu.
Email: palanisamy.hamsagayathri@gmail.com

Received on: 12-02-2017 *Accepted on: 25-03-2017*

Abstract

Breast cancer is one of the most common cancers among world’s women. Breast cancer is the cancer that develops from breast tissues and invades the surrounding tissues or metastasizes to distant parts of the body. Breast cancers can be either benign or malignant stage. Benign breast tumours are abnormal growth of the breast tissues and develop lumps in the breast, but it does not spread outside the breast. Malignant tumours are high risk and spread to other breast or organs through lymphatic system.

According to National Cancer Registry, approximately 182,000 new cases of breast cancer are diagnosed and 46,000 women die of breast cancer every year in India. Therefore, early detection and classification of breast cancer along with effective treatment is required to save women’s life. Classification is one of the most important techniques used to classify the data. This paper analyzes the different classifier algorithms for seer breast cancer dataset using WEKA software. The performance of the Naive Bayes and IBK classifiers are evaluated against the parameters like accuracy, Kappa statistic, Entropy, RMSE, TP Rate, FP Rate, Precision, Recall, F-Measure, ROC, Specificity, Sensitivity. k-nearest neighbour algorithms holds the high accuracy of 93% for SEER Breast cancer dataset.

Keywords: Classification, Naive Bayes, IBK, KNN, Accuracy, RMSE, Confusion Matrix.

1. Introduction

Breast cancer is one of the deadly cancers among the women worldwide. Breast cancers in the younger women (less than 40 years of age) grow aggressively than the older women. The survival rate of the younger women is actually 7% less than the older women. As, the occurrence of breast cancer is increasing every year by year, effective methods is required for diagnosis and prognosis of the breast cancer. The legitimate motivation of this research is to construct the classifier model for the raw medical data of breast and to classify with most accurate by physicians to save the life of women.

X-ray Mammography, is currently used most popular method of breast screening, but it has serious limitations. Mammographic screening is invasive and it is not recommended for women younger than 40 years of age. Also, there are various imaging techniques are available to evaluate the cancer in humans. In all imaging techniques diagnosis of the disease is made by the experienced physicians. However, the physicians are experienced, there always be the presence of human errors, which reduce the accuracy. The diagnosis of breast cancer with machine learning algorithms is more correct with approximated accuracy of 91.1% [5] and thus its usage has been increased in medical diagnosis.

The machine learning classifier algorithms helps not only the experienced specialist / surgeon but also the inexperienced physicians to diagnosis accurately by minimizing errors. The Machine learning can be applied to medical to improve the efficiency of the systems.

Research Objective

The genuine objective of this research is to undergo a performance analysis of IBK and NaiveBayes classifier algorithms and to select the best classifier for Breast cancer classification of SEER breast cancer dataset.

Research Scope

The scope of the research is to apply NaiveBayes and IBK classifiers on SEER Breast cancer dataset. The performance analysis on these classifiers includes classification accuracy, True Positive rate, True Negative, False Positive Rate, False Negative, ROC, Precision, Recall, Sensitivity, Specificity, and RMSE as performance metrics.

This paper is classified as follows. Section 2 gives a brief description on classification algorithms that are used to classify the data and section 3 provides the detailed description on datasets and discussed about the simulation results that are obtained for various algorithms.

2. Methodology

In most of the medical diagnosis system, classification is one of the most significant modules used for decision making in machine based learning algorithms. The main objective of the classification is to identifying the target class for each instance in the dataset accurately. There are two types of learning in the classification, supervised learning and unsupervised learning.

Supervised learning has specified class label and instance are analyzed against the target class where as in unsupervised learning there is no class label specified but to find the common patterns and grouping similarities. The process flow of various steps involved in classification is represented in below Figure 1.

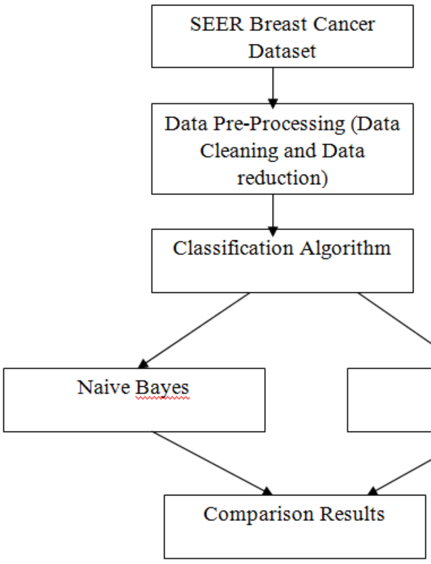

SEER Breast Cancer Dataset
Data Pre-Processing (Data Cleaning and Data reduction)
Classification Algorithm
Naive Bayes
IBK
Comparison Results

Figure 1: Methodology for data classification.

For the given dataset, Data pre-processing steps are performed to improve the quality of the data. There are several methods for Data pre-processing. In this research, we consider data cleaning and data reduction techniques for SEER breast cancer dataset.

Data cleaning: Data Cleaning is the first and fore-most steps to pre-process the data to handle missing values of attributes. Missing values or unknown values are replaced by the average value for that attribute.

Data reduction: Feature selection techniques reduce the dimensionality of the data by removing the redundant attributes from the dataset.

It excludes less significance attributes for the classification. Breast cancer dataset consists of 762691 instances with 134 attribute, using information gain feature selection technique only 7 attributes are considered.

Table 1. Dataset Attributes for Classification.

Attribute	Length	Description
Age at diagnosis	3	Age of the patient at diagnosis
Grade	1	Specify T-cell, B-Cell involvement in lymphoma and leukemia
CS Tumor Size	3	Information on Tumour Size
CS Extension	3	Information on extension of Tumour

CS Lymph nodes	3	Information on involvement of lymph nodes.
CS Mets at DX	2	Information on distant metastasis.
Behavior code ICD-O-3	1	Describes on the nature of tumour as begin, in situ or malignant

2.1 Classification Algorithms

There are various algorithms available for classification of Breast cancer. This paper deals the with NaiveBayes and IBK classifier algorithms and evaluates performance with different parameters such as accuracy, sensitivity, specificity, entropy, ROC, PR area and so on.

2.2 IBK Algorithm

IBK implements k-nearest neighbour (KNN) classifier algorithm.KNN caches all available instances and classify new instance based on distance functions. The new instance is classified by majority vote of its neighbours. Three distance functions (Euclidean, Manhattan and Minkowsi) are used to identify the similarity among continuous variable. Hamming distance function is used for discrete variables. For most of the datasets the 'k' takes the optimal value from 3 to 10 and should be an odd number.

2.2.1 The Algorithm

- Input training data and 'K' value
- Compute the distance between the new instance predictors with the existing instances
- Identify 'K closest neighbours and based on distance to predict the target

2.3 Naive Bayes Algorithm

NaiveBayes is one of the supervised classification technique based on Bayes theorem with the assumption of predictors are independent. Naive Bayes classifier is suitable for large datasets where the presence of particular feature in a class is not related to the presence of any other feature.

Bayes theorem calculates the posterior probability of P (B/A) from P (A), P (B) and P (A/B) using below equation

$$P\left(\frac{B}{A}\right) = \frac{P\left(\frac{A}{B}\right)P(B)}{P(A)} \qquad (1)$$

Where

P (B/A) = Posterior probability of target (A) for given predictor (B)

P (A) = Prior probability of target (A)

P (B) = Prior probability of predictor (B)

P (A/B) = Likelihood which is the probability of predictor given class

2.3.1 The Algorithm

- Construct frequency table of each attribute against target(i.e., Class)
- Transform frequency table to likelihood tables
- Apply Bayesian equation to calculate the posterior probabilities for each attribute
- Attribute with highest posterior probability is the outcome

3. Simulation Results and Discussion

For this research work, NaiveBayes and IBK classifier algorithms are applied to the breast cancer dataset from Surveillance, Epidemiology, and End Results (SEER) repository. The breast cancer dataset has 769261 numbers of instances and each instance consists of 134 attributes including the class attribute. The class attribute has four values like Benign (0), uncertain benign or malignant (1), Carcinoma in situ (2) and Malignant (3). All the attributes of the data set along with their range of values are available in seer data dictionary [6].The classification algorithms are applied for the input parameters mentioned in the Table [1].The classifiers with 10 fold cross validation are analyzed and compared using WEKA software. The default configuration parameters of the classifiers are considered for classification.

In WEKA, Data pre processing has been carried out as first step and extracted 7 attributes for classification and it has been depicted in Figure [2].

The performance of the classifiers in detecting the breast cancer can be evaluated from the analysis of confusion matrix and below parameters are calculated

Accuracy is the measure of correctly classified instances for all set of instances. It can be calculated as

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

Precision is ability of the classifier to correctly classify instances for those instances that are already classified as positive and it is obtained using the equation

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

Recall is the measure of the classifier to correctly classified instances that are positive and it can be expressed as

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

F-Measure is the harmonic mean of recall and precision parameters and it can be written as

$$F\text{-}Measure = \frac{2*Recall*Precision}{Precision+Recall} \qquad (5)$$

Sensitivity is to measure the correctly classified positive instances from total number of positive instances.

$$Sensitivity = \frac{TP}{TP+FN} \qquad (6)$$

Specificity is to measure the correctly classified negative instances from total number of negative instances.

$$Specificity = \frac{TN}{TN+FP} \qquad (7)$$

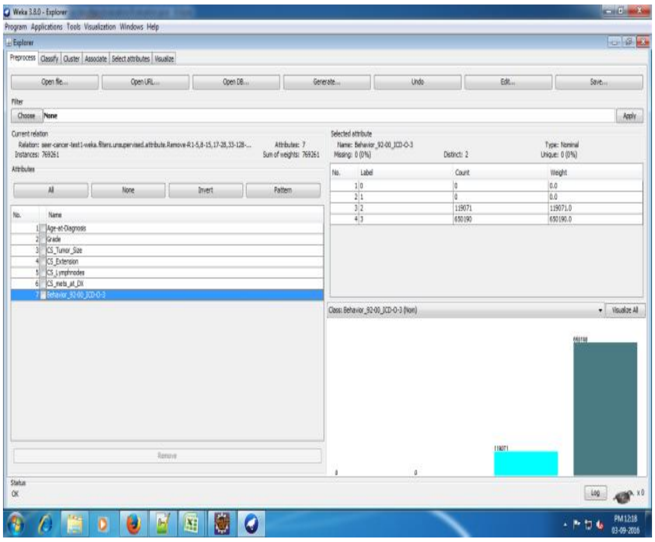
Receiver operating curve (ROC) is graphical representation of sensitivity Vs specificity.

RMSE is the difference between the actual value and predicated correct values.

Entropy: Entropy determines the average information of the attributes of instances in the dataset and it is defined as follows

$$Entropy\ H(X) = \sum_{x=1}^{n} \quad p_i \log_b p_i \qquad (8)$$

The different classifier algorithms are imposed on the pre-processed data. Figure 3 shows comparison of different performance measures of classifiers.



Filter: None

Relation: seer-cancer-test1-weka.filters.unsupervised.attribute.Remove-R1-5,8-15,17-28,33-128-...
Instances: 769261
Attributes: 7
Sum of weights: 769261

No.	Name
1	Age-at-Diagnosis
2	Grade
3	CS_Tumor_Size
4	CS_Extension
5	CS_Lymphnodes
6	CS_mets_at_DX
7	Behavior_92-00_ICD-O-3

Name: Behavior_92-00_ICD-O-3
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

No.	Label	Count	Weight
1	0	0	0.0
2	1	0	0.0
3	2	119071	119071.0
4	3	650190	650190.0

Class: Behavior_92-00_ICD-O-3 (Nom)

Status
OK

Figure 2: Data Pre-processing of selected attributes.

The simulation results of NaiveBayes and IBK classifiers are plotted here. Confusion matrix helps us to evaluate total number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) instance. With the help of TP, TN, FP and FN value, it is possible us to validate the various performance measures such as accuracy, precision , recall, F-measure, ROC, PRC, etc.

Results of IBK

Table 2. Performance parameters of IBK.

Parameters	Class (0)	Class(1)	Class(2)	Class(3)
TP Rate	0	0	0.684	0.983
FP Rate	0	0	0.017	0.316
Precision	0	0	0.883	0.944
Recall	0	0	0.684	0.983
F-Measure	0	0	0.771	0.964
ROC	0	0	0.939	0.939
PRC	0	0	0.839	0.987

Results on NaiveBayes:

Table 3. Performance parameters of NaiveBayes

Parameters	Class (0)	Class(1)	Class(2)	Class(3)
TP Rate	0	0	0.997	0.319
FP Rate	0	0	0.681	0.003
Precision	0	0	0.211	0.998
Recall	0	0	0.997	0.319
F-Measure	0	0	0.349	0.483
ROC	0	0	0.865	0.865
PRC	0	0	0.738	0.968

The simulation results of the IBK and Naivebayes Classifier for all possible values of class attribute is summarized in the Table.2 and Table.3. The weighted average of the performance parameters are captured in the Figure 3. When compared to NaiveBayes, IBK Classifier has high TP Rate of 0.937 and low FP Rate of 0.270.IBK holds good Receiver Operating Curve (ROC) of 0.939 with 0.964 Precision-Recall areas.

The comparison of information score of IBK and NaiveBayes classifier is clearly shown in Table.5 and IBK has 280504 bits per instances.

Error Results

Table 4. Error comparison of IBK and NaiveBayes.

Error	IBK	NaiveBayes
Kappa statistic	0.7349	0.1255
Mean absolute error	0.0516	0.2715
Root mean squared error	0.1607	0.5019
Relative absolute error	39.4688%	207.5248%
Root relative squared error	62.8243%	196.2435%

IBK Classifier classifies the data with minimized Mean Squared Error value of 0.1607 and various error statistics are captured in Table 4.

Table 5. Comparison of IBK and NaiveBayes.

Parameters	Naive Bayes	IBK
Correctly classified instances	326003	720822
Incorrectly classified instances	443258	48439
Accuracy	42.37%	93.7%
Specificity	31.8%	97%
Sensitivity	99.7%	70%

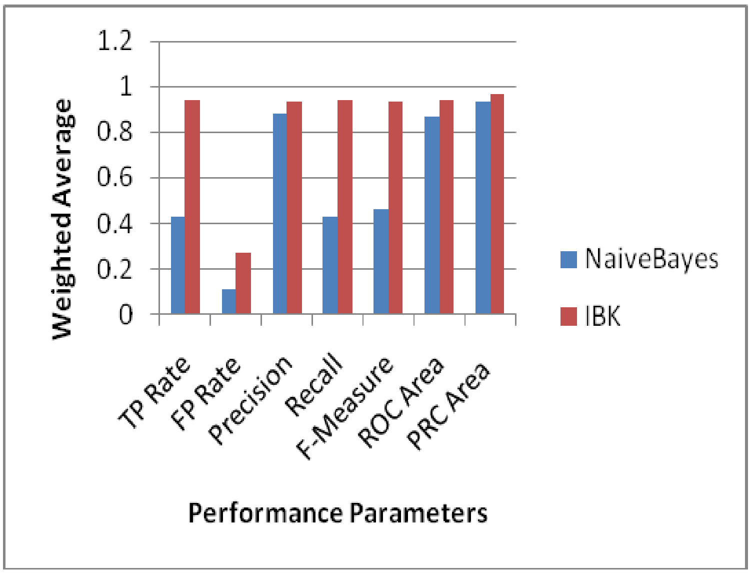


Figure 3: Comparison of Weighted average parameters.

Comparison of various parameters of Naive Bayes and IBK classifier algorithms are tabulated in the Table 5.

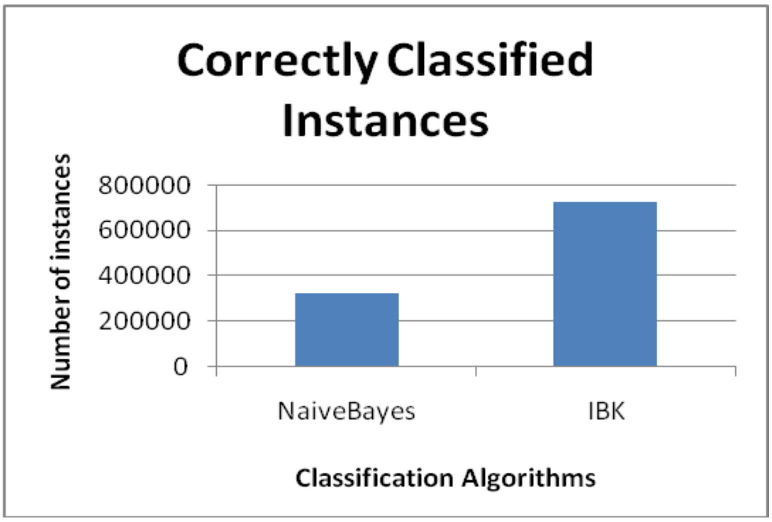


Figure 4: Comparison of Correctly classified instances.

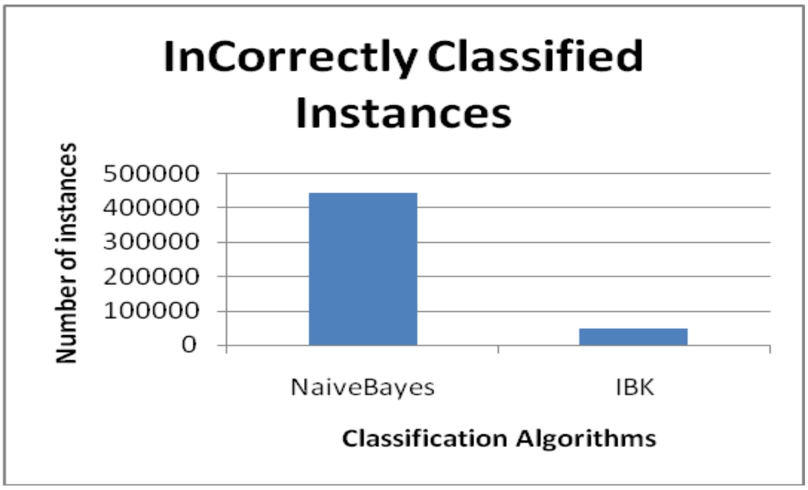


Figure 5: Comparison of Incorrectly classified instances.

IBK Classifier with performance accuracy of 93.7%, 720822 instances are correctly classified and only 48439 instances are classified as incorrect. The charts are captured in Figure 4, 5 and 6 respectively.

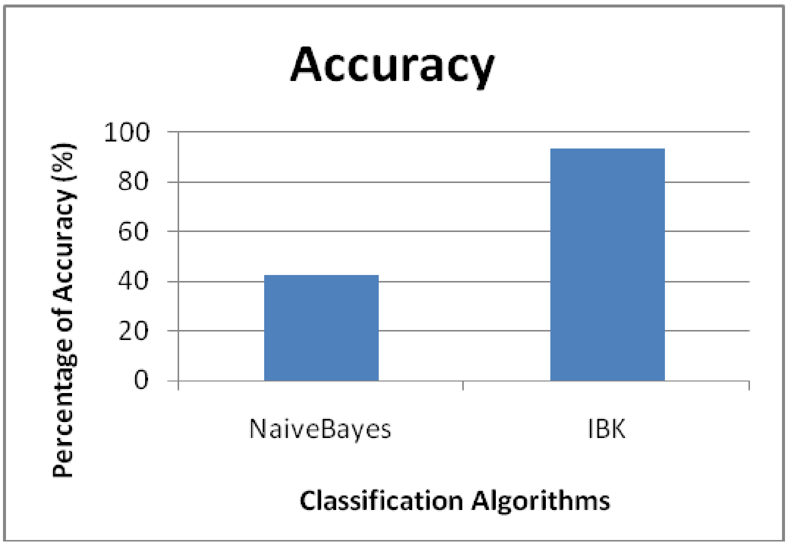


Figure 6: Comparison of performance accuracy.

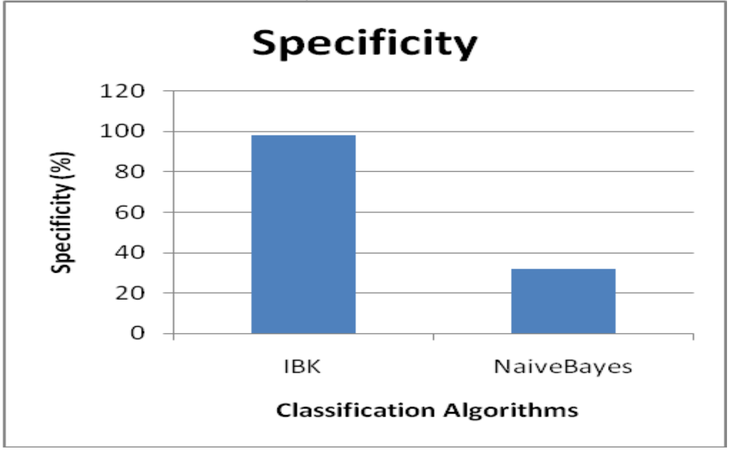


Figure 7: Comparison of Specificity.

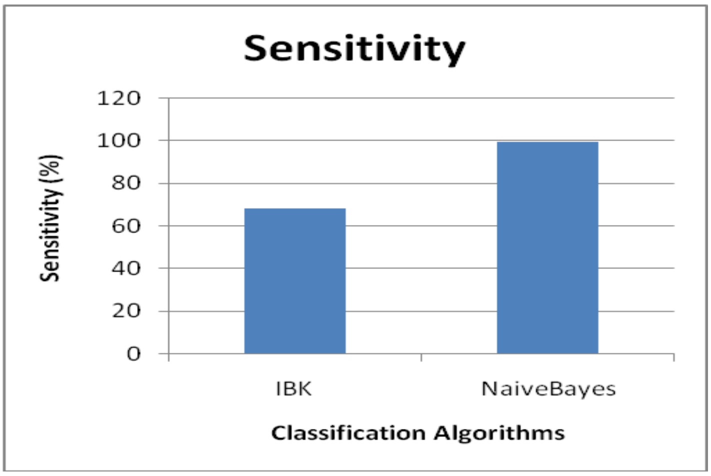


Figure 8: Comparison of Sensitivity.

IBK has specificity of 97%, sensitivity of 70% and thus it has better capability to classify the data, and graphs are depicted in the Figure 7 and Figure 8.

3. Conclusion

We have analyzed the performance of the NaiveBayes and IBK algorithms for Breast cancer classification. The simulation results shows IBK classifier classifies the data with 93% accuracy and minimum RMSE of 0.1607. IBK algorithm consumes less time and it has 0.939 ROC and 0.964 PRC values. By performance analysis and comparision, we confirm that IBK algorithm is better than Naivebayes classifier for SEER dataset Breast cancer classification.

5. References

1. Aruna S, Rajagopalan SP, Nandakishore LV. Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology. 2011; 2:37–45.
2. Vaidehi K, Subashini TS. Breast tissue characterization using combined K-NN classifier. Indian Journal of Science and Technology. 2015 Jan;8(1):23–6.

3. Williams K, Idowu PA, Balogun JA, Oluwaranti A. Breast cancer risk prediction using data mining classification techniques.Transactions on Networks and Communications. 2015; 3(2):1–11.
4. Xindog Wu, Vipin Kumar et al., "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, 14(1), 1-37 (2008).
5. R. W. Brause, "Medical analysis and diagnosis by neural networks" Lecture notes In Computer Science, vol. 2199, pp. 1-13, 2001.
6. http://seer.cancer.gov/popdata/popdic.html-SEER dictionary
7. T. M. Cover, (1965) "Geometrical and Statistical Properties of Systems of Linear with Applications in Pattern Recognition," IEEE Transactions on Electronic Computers EC-14, pp. 326-334.
8. Ramnath Takiar et al. "Projections of Number of Cancer Cases in India (2010-2020) by Cancer Groups", Asian Pacific Journal of Cancer Prevention, Vol 11, 2010.
9. Evanthia E. Tripoliti et al. "Automated Diagnosis of Diseases Based on Classification: Dynamic Determination of the Number of Trees in Random Forests Algorithm" IEEE Transactions On Information Technology In Biomedicine, Vol. 16, No. 4, July 2012.