Predicting Breast Cancer Recurrence using Data Mining Techniques

Siddhant Kulkarni

Mangesh Bhagwat

ABSTRACT
Breast Cancer is among the leading causes of cancer death in women. In recent times, the occurrence of breast cancer has increased significantly and a lot of organizations are taking up the cause of spreading awareness about breast cancer. With early detection and treatment it is possible that this type of cancer will go into remission. In such a case, the worse fear of a cancer patient is the recurrence of the cancer. This paper evaluates various data mining techniques and their ability to predict whether any particular patient will face a recurrence. Experimental results will show the accuracy of various classifiers when applied on the Breast Cancer Dataset[1].

Keywords
Breast Cancer, Data Mining, Data pre-processing, Classifiers

1. INTRODUCTION
Widespread use of machine learning in medical application is becoming mainstream extremely quickly. The analysis of existing medical records enables machine learning algorithms to make predictions about the health of a patient to a certain degree of certainty. This paper focuses on using Classification a.k.a. Supervised Learning to predict whether a particular patient will face recurrence of cancer.

Recurrence of cancer refers to the reoccurrence of cancer in a patient whose previous cancer has gone into remission. Remission is usually the result of chemotherapy and regular treatment by oncologists. Recurrence of cancer is one of the biggest fears in the life of a cancer patient and thus one of the issues that affect their quality of life. This paper attempts to evaluate the accuracy of various classifiers with the help of different pre-processing techniques which reduce errors in the input data.

The dataset used for experimentation has a total of 10 attributes including the class attribute. The features considered in this dataset include various aspects such as age, location of tumor, etc. The accuracy of classifiers is evaluated on the data preprocessed using 6 different data preprocessing techniques and also without using any preprocessing technique and providing the input data to classifier as is.

Rest of the paper is organized as follows: Section 2 provides an overview of the related work done by researchers. Section 3 lists the various classifiers used during experimentation whereas section 4 lists the various preprocessing techniques used during experimentation. Section 5, 6 and 7 elaborate the experimental results, conclusions and references respectively.

2. RELATED WORK
Skevofilakas et al. [2] have developed a decision support system for treatment of breast cancer using various data mining techniques for improving the provision and visualization of clinical data that is specific to a particular disease.

Menolascina et al. [3] have presented a comparison between J48, Naive Bayesian Tree, Ant Miner and Gene Expression Programming to perform aCGH based breast cancer subtype profiling.

Yang et al. [4] have proposed the use of Beier-Neely field morphing along with decision trees to analyze the parameters identified from parametric analysis for diagnosis of Breast cancer using thermographs in gray scale.

Sarvestani et al. [5] have proposed the use of statistical neural network structures such as SOM, RBF, GRNN, PNN on the WBCD and NHBCD data sets to test accuracy for predicting breast cancer survivability.

Salama et al. [6] have presented a comparison between classification accuracies of Decision tree, Multi Layer Perception, Naive Bayesian, Sequential Minimal Optimization and Instance based for k-Nearest Neighbor on the Wisconsin Breast Cancer, Wisconsin Diagnosis Breast Cancer and Wisconsin Prognosis Breast Cancer datasets.

3. CLASSIFIERS
Classifiers are Supervised Learning algorithms which rely on annotated training data with accurate class labels to build a learning model (statistical, tree based, etc.) which these algorithms then apply on the testing data to predict the class label of this data. This paper has evaluated accuracy of 19 different classifiers to predict the recurrence of breast cancer. These classifiers use different learning paradigms and thus this paper presents a comprehensive set of results. Following classifiers have been tested in this paper: Bayesian Network, IBK, Naive Bayesian, K-Star, Naive Bayesian for Multinomial Text, K-Star, LWL, Input Mapped Classifier, Decision Table, jRip Classifier, OneR Classifier, PART, ZeroR Classifier, Decision Stump, j48, LMT Classifier, Random Forest, Random Tree, Reduced Error Pruning Tree and k-Nearest Neighbor Classifier with Hamming Distance as a similarity measure.

4. DATA PRE-PROCESSING
Any data captured in real life is going to have one form of errors or the other. Data preprocessing techniques are utilized in various processes to remove errors from input data to avoid corruption of the experimental results. Applying data preprocessing as the first stages avoids the ripple effect corrupted data tends to have on the rest of the classification process. This paper uses various preprocessing techniques such as Standardization, Discretization, PKIDiscretization, Numeric to Binary, Normalization, String to Nominal on the training as well as testing data before it is provided to the classifier.

5. EXPERIMENTAL RESUTLS

Intel Core-i5 (1.70 GHz) machine having 4 GB Ram is used to perform experiments. Performance of various classifiers in combination with data pre-processing methods is evaluated using the Breast Cancer dataset. All the experiments except for the ones with kNN classifier were performed using Weka tool (Version 3.7.9) [7]. Throughout these experiments, the heap size allocated to Weka tool was 1408 MB using the Java –Xmx command. The k Nearest Neighbor classifer using Hamming Distance was developed entirely in Java using JDK1.7.75.

Table 1. Abbreviations used while presenting Experimental Results

Abbreviation	Meaning
None	No Preprocessing
Str2Nom	String to Nominal
PKID	PKI Discretization
N2B	Numeric To Binary
Std.	Standardization
Norm.	Normalization
Disc.	Discretization

Figure 1 shows the accuracy of Bayes Net classifier with various data pre-processing techniques. It can be seen that Bayes Net provides the maximum accuracy of 47.67%.

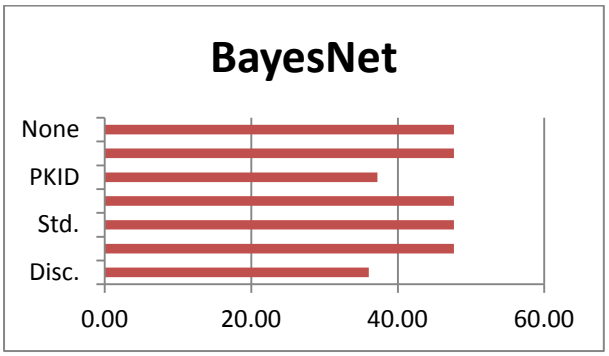

Fig.1 Accuracy of Bayes Net Classifier for Predicting Breast Cancer Recurrence

Figure 2 shows the accuracy of Naive Bayesian classifier with various data pre-processing techniques. It can be seen that Naive Bayesian provides the maximum accuracy of 46.51%.

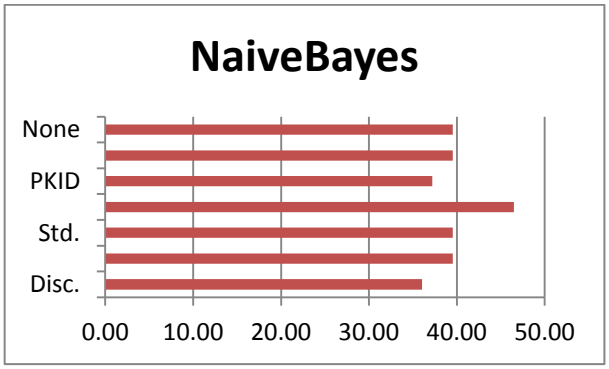

Fig.2 Accuracy of Naive Bayesian Classifier for Predicting Breast Cancer Recurrence

Figure 3 shows the accuracy of Naive Bayesian Multinomial Text classifier with various data pre-processing techniques. It can be seen that Naive Bayesian Multinomial Text provides the maximum accuracy of 25.58%.

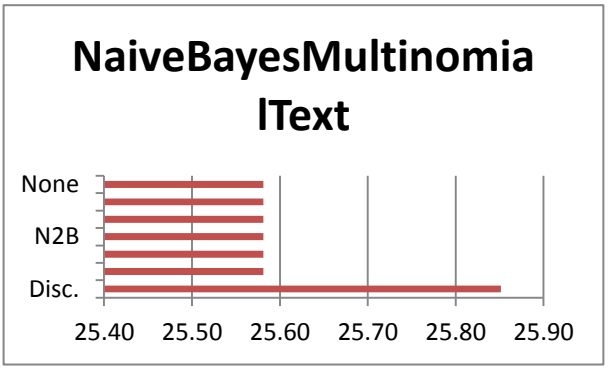

Fig.3 Accuracy of Naive Bayesian Multinomial Text Classifier for Predicting Breast Cancer Recurrence

Figure 4 shows the accuracy of IBK classifier with various data pre-processing techniques. It can be seen that IBK provides the maximum accuracy of 45.35%.

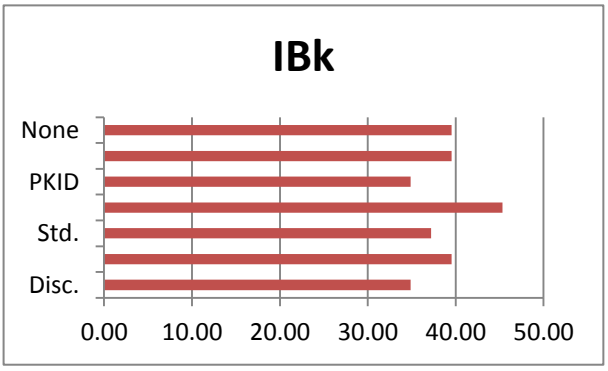

Fig.4 Accuracy of IBK Classifier for Predicting Breast Cancer Recurrence

Figure 5 shows the accuracy of K-star classifier with various data pre-processing techniques. It can be seen that K-star provides the maximum accuracy of 31.50%.

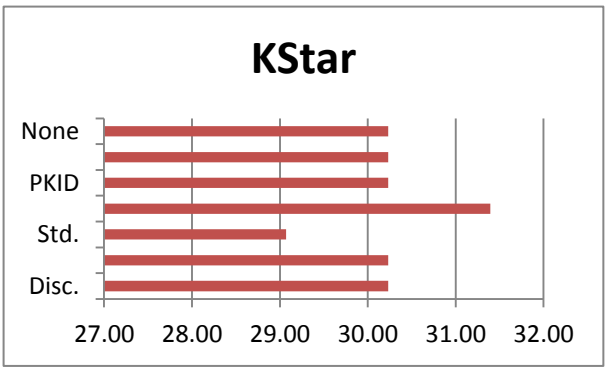

Fig.5 Accuracy of K-star Classifier for Predicting Breast Cancer Recurrence

Figure 6 shows the accuracy of LWL classifier with various data pre-processing techniques. It can be seen that Naive Bayesian provides the maximum accuracy of 34.88%.

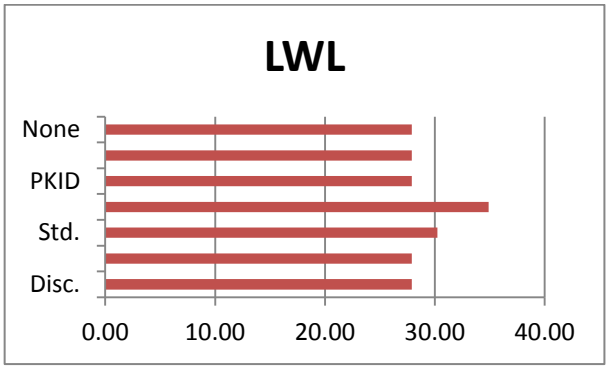

Fig.6 Accuracy of LWL Classifier for Predicting Breast Cancer Recurrence

Figure 7 shows the accuracy of Input Mapped classifier with various data pre-processing techniques. It can be seen that Input Mapped Classifier provides the consistent accuracy of 25.58%.

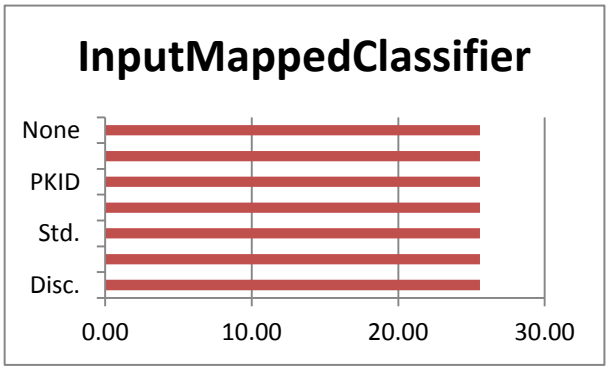

Fig.7 Accuracy of Input Mapped Classifier for Predicting Breast Cancer Recurrence

Figure 8 shows the accuracy of Decision Table classifier with various data pre-processing techniques. It can be seen that Decision Table provides the consistent accuracy of 44.19%.

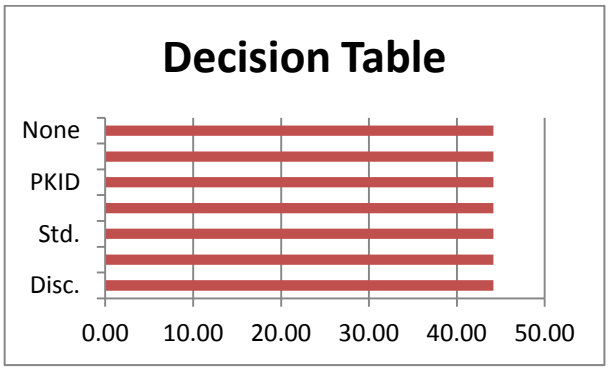

Fig.8 Accuracy of Decision Table Classifier for Predicting Breast Cancer Recurrence

Figure 9 shows the accuracy of jRip classifier with various data pre-processing techniques. It can be seen that jRip provides the maximum accuracy of 74.42%.

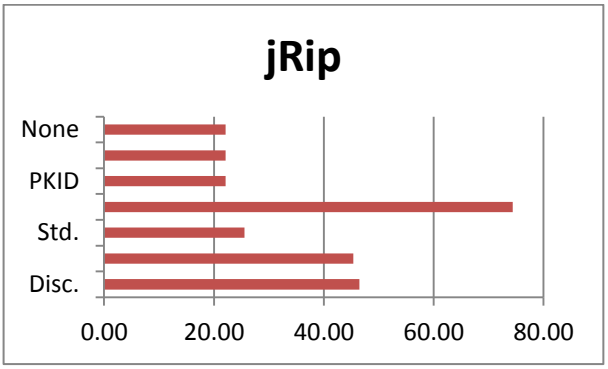

Fig.9 Accuracy of jRip Classifier for Predicting Breast Cancer Recurrence

Figure 10 shows the accuracy of OneR classifier with various data pre-processing techniques. It can be seen that OneR provides the consistent accuracy of 27.91%.

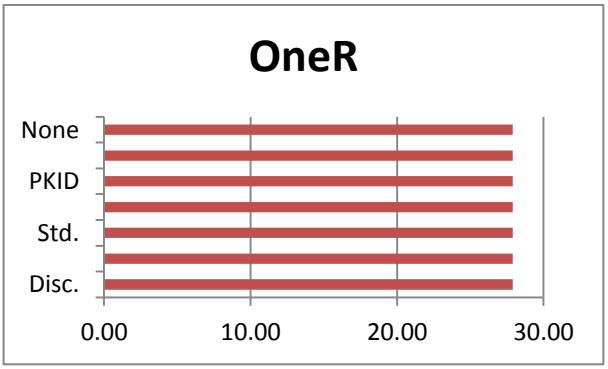

Fig.10 Accuracy of OneR Classifier for Predicting Breast Cancer Recurrence

Figure 11 shows the accuracy of PART classifier with various data pre-processing techniques. It can be seen that PART provides the maximum accuracy of 41.86%.

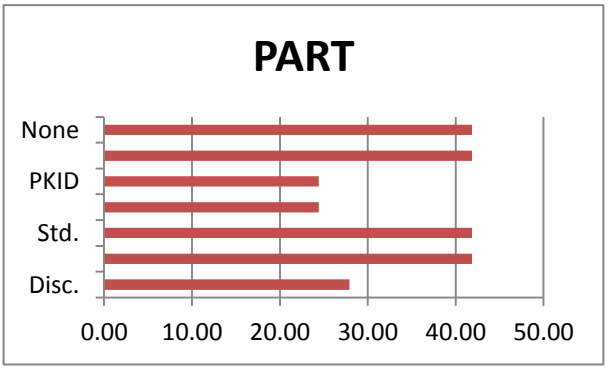

Fig.11 Accuracy of PART Classifier for Predicting Breast Cancer Recurrence

Figure 12 shows the accuracy of ZeroR classifier with various data pre-processing techniques. It can be seen that ZeroR provides the consistent accuracy of 25.58%.

Fig.12 Accuracy of ZeroR Classifier for Predicting Breast Cancer Recurrence

Figure 13 shows the accuracy of Decision Stump classifier with various data pre-processing techniques. It can be seen that Decision Stump provides the maximum accuracy of 44.19%.

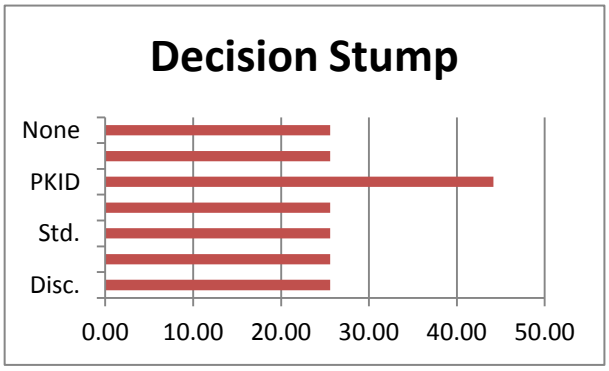

Fig.13 Accuracy of Decision Stump Classifier for Predicting Breast Cancer Recurrence

Figure 14 shows the accuracy of J48 classifier with various data pre-processing techniques. It can be seen that J48 provides the maximum accuracy of 25.58%.

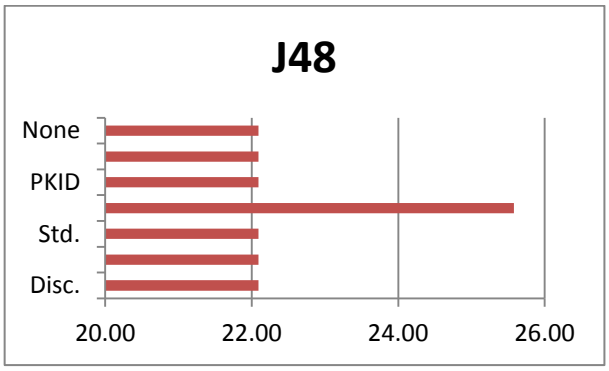

Fig.14 Accuracy of J48 Classifier for Predicting Breast Cancer Recurrence

Figure 15 shows the accuracy of LMT classifier with various data pre-processing techniques. It can be seen that LMT provides the consistent accuracy of 25.58%.

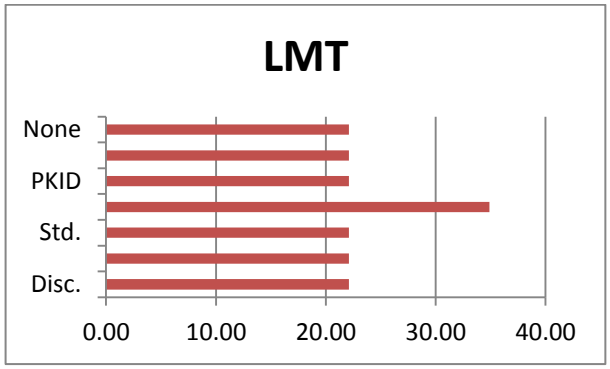

Fig.15 Accuracy of LMT Classifier for Predicting Breast Cancer Recurrence

Figure 16 shows the accuracy of Random Forest classifier with various data pre-processing techniques. It can be seen that Random Forest provides the maximum accuracy of 45.35%.

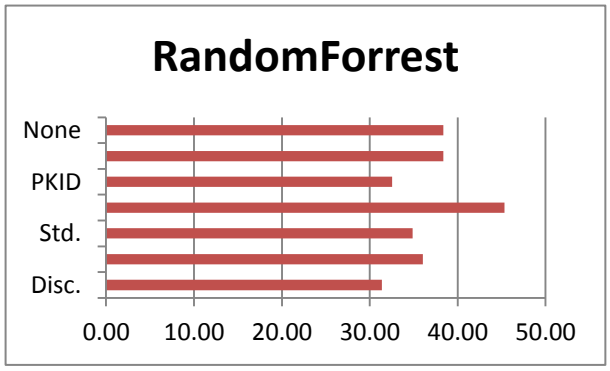

Fig.16 Accuracy of Random Forest Classifier for Predicting Breast Cancer Recurrence

Figure 17 shows the accuracy of Random Tree classifier with various data pre-processing techniques. It can be seen that Random Tree provides the maximum accuracy of 46.51%.

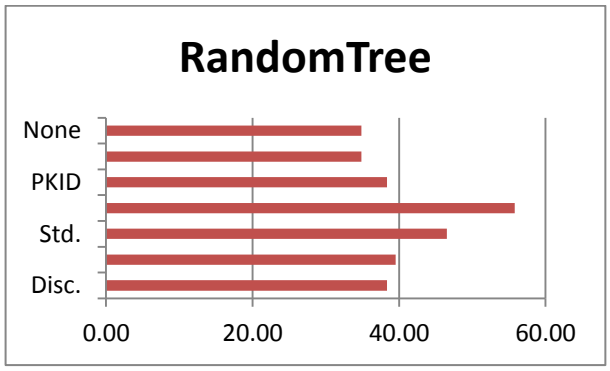

Fig.17 Accuracy of Random Tree Classifier for Predicting Breast Cancer Recurrence

Figure 18 shows the accuracy of REPTree classifier with various data pre-processing techniques. It can be seen that REPTree provides the consistent accuracy of 25.58%.

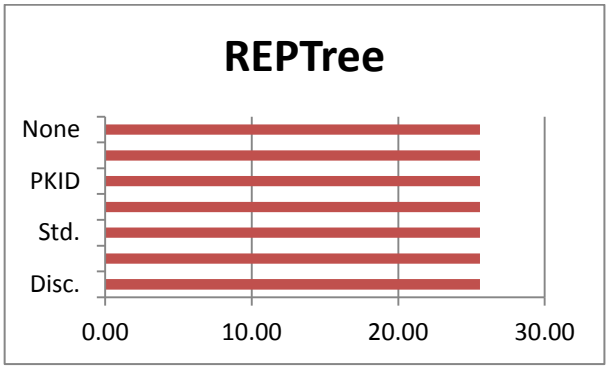

Fig.18 Accuracy of REPTree Classifier for Predicting Breast Cancer Recurrence

Figure 19 shows the accuracy of k-Nearest Neighbor classifier(with Hamming Distance) with various data pre-processing techniques. It can be seen that k-Nearest Neighbor classifier(with Hamming Distance) provides the maximum accuracy of 74.42%.

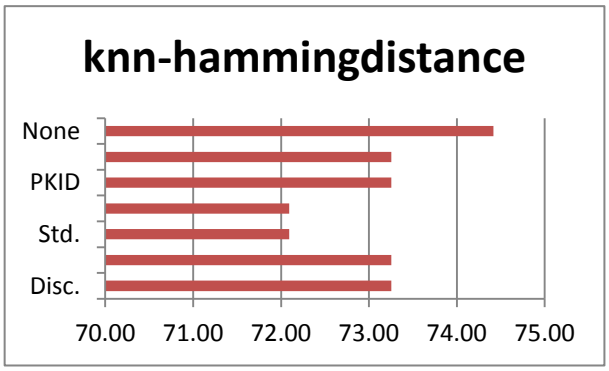

Fig.19 Accuracy of K-NN Classifier for Predicting Breast Cancer Recurrence

6. CONCLUSION

This paper evaluates the accuracy of various classifiers for predicting recurrence of breast cancer based on the attributes provided in the data set [1]. In order to support the task of classification, this paper uses various data pre-processing techniques and presents the results accordingly.

Figure 20 shows the maximum accuracy provided by each classifier in order to compare them.

Maximum accuracy of 74.42% can be achieved by using jRip Classifier with Numeric to Binary data pre-processing and by using k-Nearest Neighbor Classifier with Hamming Distance as a similarity measure.

According to the results presented in Figure 9 and 19, it is clear the k-Nearest Neighbor provides a consistently high accuracy of predicting Breast Cancer recurrence as compared to the jRip classifier.

7. FUTURE SCOPE

This paper attempts to predict recurrence of breast cancer using different classifiers. Even though many classifiers have been covered, there is still a lot more that can be done.

Extensive research can be carried out on this application using different Machine Learning techniques. In addition, a more detailed data set can be developed which covers additional attributes related to a patient's medical history. Unsupervised machine learning can be applied on such a data sets to identify clusters of patient records and the similarity between them.

8. REFERENCES

[1] Breast Cancer Data Set. <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer> Last visited March 2015.

[2] Marios Skevofilakas, Konstantina Nikita, Panagiotis Templaleksis, K. Birbas, I. Kaklamanos, G. Bonatsos, "A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines", pp. 2429 - 2432, Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, China, Sept. 2005.

[3] Menolascina F, Tommasi S, Paradiso A, Cortellino M, Bevilacqua V, Mastronardi G, "Novel Data Mining Techniques in aCGH based Breast Cancer Subtypes Profiling: the Biological Perspective", pp. 9-16, Proceedings of the 2007 IEEE symposium on computational Intelligence in Bioinformatics and Computational Biology, 2007.

[4] Chi-Shih Yang, Ming-Yih Lee, "Parametric Data Mining and Diagnosis Rules for Digital Thermographs in Breast Cancer", pp. 98-101, 30th Annual International IEEE Conference Vancover, Canada, August 2008.

[5] A. Soltani Sarvestani, A. A. Safavi, N. M. Parandeh, M. Salehi, "Predicting Breast Cancer Survivability using data mining techniques", pp. V2-227 - V2-231, 2nd International Conference on Software Technology and Engineering, 2010.

[6] Gouda Salama, M.B. Abdelhalim, Magdy Zeid, "Experimental Comparison of Classifiers for Breast Cancer Diagnosis", Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on , vol., no., pp.180,185, 27-29 Nov. 2012.

[7] Weka 3: Data Mining Software in Java. <http://www.cs.waikato.ac.nz/ml/weka/> (accessed August 2013)

9. APPENDIX

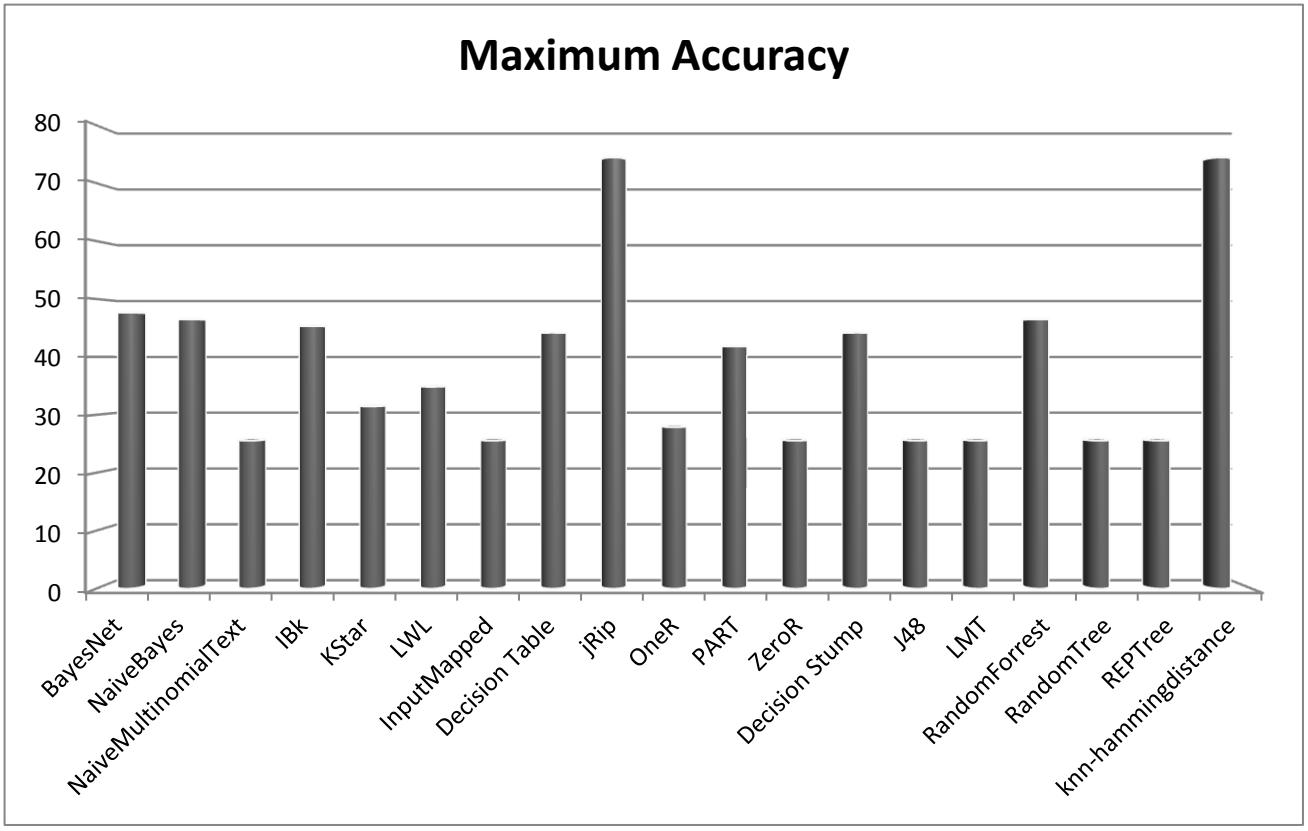

Fig. 20 Maximum accuracy provided by each classifier while predicting recurrence of Breast Cancer