Evaluation of Classifiers Performance using Resampling on Breast cancer Data

G. NAGA RAMADEVI, Dr. K. USHA RANI, Dr.D.LAVANYA

ABSTRACT: Data Mining refers to extracting or mining knowledge from large amount of data. One of the data mining techniques i.e., classification is an interesting topic to the researchers as it accurately and efficiently classifies the data for knowledge discovery. Classification is used in every field of real life. The datasets contain many irrelevant and redundant features that mislead the classifiers. Furthermore, many huge datasets have imbalanced class distribution which leads to bias over majority class in the classification process. Preprocessing techniques are helpful to handle these problems. To balance the data either Under-sampling i.e., reduces the set of examples of majority class or Over-sampling i.e., replicates minority class examples, can be used. In this paper, experiments are conducted on popular and frequently used classifiers on breast cancer datasets without-resampling and with resampling. Breast cancer datasets are considered because the breast cancer is one of the leading causes of death in women. Finally, the results are analyzed and the best classifier for each dataset is identified.

KEYWORDS: DataMining, Classification, Resampling, Breast Cancer, K-NN, SVM, C4.5, Logistic Regression, Random Forest, Statistics Measures.

1 INTRODUCTION

Machine learning is a scientific discipline that explores the construction and study of algorithms that can learn from data. It is based on known properties learned from the training data and focuses on prediction. The performance of the data is evaluated in machine learning with respect to the ability to reproduce known knowledge. Knowledge Data Discovery (KDD) is a process of deriving hidden knowledge from databases [1]. KDD consists of several methods like cleaning, integration, selection and transformation of data, data mining, and evaluation of patterns and representation of knowledge. Data mining refers to the discovery of new information in terms of patterns or rules from vast amounts of data. There are several data mining functions to find the useful patterns such as concept descriptions, association rules, classification, prediction, sequence discovery and clustering [2]. Classification is a classic data mining technique with broad applications. Medical diagnosis is regarded as an important though complicated task that needs to be executed accurately and efficiently. Data mining has the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. Supervised learning algorithms are trained on class samples (i.e., for every input where the desired output is known).

Classification wll come under supervised learning.

- *G.Naga Ramadevi is currently pursuing Ph.D in Compute Science in SPMVV, Tirupati. E-mail:* ramadeviaba@yahoo.co.in
- *Dr.K.Usha Rani is currently working as Professor in Computer Science Deprtment, SPMVV, Tirupati E-mail:* usharanikuruba@yahoo.co.in
- *Dr.D.Lavanya is currently working as Professor in Computer Science & Engineering Deprtment, SEAGI, Tirupati E-mail: lav_dlr@yahoo.co.in*

It classifies data based on the training set and constructs a model which is helpful in classifying new data [3]. The most popular classification algorithms used in our research work are K-Nearest Neighbour, Support Vector Machine, Logistic Regression, C4.5 and Random Forest. Resampling is commonly used for balanceing the data. In the small sample context, it is preferable to implement the resampling approaches for error rate estimation. Resampling adds samples to minority classes or reduces samples in majority classes in imbalanced data sets by using artificial mechanisms [4].

Random sampling consists of different techniques like simple random sampling, adaptive sampling, stratified random sampling, cluster sampling, restricted random sampling, two-stage random sampling, unequal probability sampling, double sampling, and spatially balanced sampling [5]. Non random sampling consists of different techniques like synthetic sampling, selected or targeted sampling, and haphazard sampling

Resampling methods are designed to improve classifier accuracies when used in conjunction with algorithms for training the classifiers. Resampling methods can be classified into the two groups; cross validation and bootstrap. Cross validation methods are random-sub-sampling, k-fold cross validation, leave-one-out cross validation. There are three types of Resampling methods such as Random Over-sampling and Under-sampling, Informed Under-sampling and Synthetic Sampling with data generation. Sampling methods are two types such as random sampling and non-random sampling.

Cancer is a disease characterised by uncontrolled growth and spread of abnormal cells and the capacity to invade other tissues that can be caused by both external factors like radiation, chemicals, tobacco etc., and internal factors like inherited mutations, hormones, immune conditions etc. Most of the cancers are named after the organ type or type of cell in which they appear e.g., Melanoma, Colon Cancer, Breast Cancer, Lung Cancer, Leukaemia cancer etc. Extra cells may form a mass of tissue called a tumor. Tumors can be either belign or malignant. [6].

2 Related Work

A few appropriate studies on various classifiers related to this research work is presented. Nithya et.al [7] performed a work on deriving classification rules for Indian rice diseases. The decision tree C4.5 algorithm was used to classify the disease of rice based on the symptoms. Orlando Anunciacao et.al [8] applied decision trees to detect high risk breast cancer groups over the data set produced by department of genetics of faculty of medical sciences. Dr. Medhat Mohamed Ahmed Abdelaal et.al [9] investigated the capability of the classifier SVM with Tree Boost and Tree Forest in analyzing the DDSM dataset for the extraction of the mammographic mass features along with age that discriminates true and false cases. Robert Y, J.Lee [10] performed a work to analyze whether chemotherapy could prolong survival time of breast cancer patients using data mining technique. Three nonlinear smooth support vector machines (SVM) are used for classifying breast cancer patients into the three prognostic groups i.e. Good, Poor and Intermediate. Delen et. al [11] performed a work with ANN, decision tree and logistic regression techniques for breast cancer survival analysis. They used the SEER (Surveillance Epidemiology and End Results) data's twenty variables in the prediction models. J. Padmavati [12] performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. Mythili T et.al [13] performed a work on a heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL) on the Cleveland Heart Disease. Delen Dursun et.al [14] performed a comparative study of multiple prediction models for breast cancer survivability using a large dataset with three different classification models: artificial neural networks, decision trees, and logistic regression have been used in the experiment with 10-fold-cross-validation. Shelly Gupta et.al [15] proposed a work on the performance analysis of several data mining classification techniques using three different machine learning tools over the healthcare datasets. Chao Chen [16] proposed a work on the imbalanced data classification problem using random forest. Anne-Laure Boulesteix et.al [17] performed a work on overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Vrushali YKulkarni et.al [18] performed a work on effective and classification using random forest algorithm on benchmarking datasets.

3 Materials And Methods

3.1. Data Sets:

In this study for the analysis of classifiers performance only breast cancer data sets are considered as breast cancer is a leading cause of death in Women in the world as well as in our country. Four different breast cancer data sets Breast cancer (BC), Wisconsin Diagnostic Breast Cancer (WDBC), Wisconsin Breast Cancer (WBC) and Wisconsin Prognostic Breast Cancer (WPBC) are considered which are publicly available at UCI Machine learning Repository [19]. The description of the data sets is given in Table 1. In this study for the analysis of classifiers performance on breast cancer datasets experiments are conducted with 10-fold-cross-validation using open-source data mining tool WEKA (Version 3.6.11). Preprocessing technique, Resampleing ie., random sub sampling with no replacement is also applied.

Table1: Description of Breast Cancer Datasets

Data Sets	No.of Instances	No.of Attributes	%of Major Class (-ve)	%of Minor Class (+ve)
BC	286	10	70.28	29.72
WDBC	569	32	62.74	37.26
WBC	699	10	65.52	34.48
WPDC	198	34	76.26	23.73

3.2 Classifiers Used

As per the survey stated in section 2 the popular and frequently used five classification algorithms K-Nearest neighbor (K-NN), Support Vector Machine (SVM), Logistic Regression (LR), C4.5 and Random forest (RF) are experimented in this study. The following section gives a brief description about each of these algorithms.

3.2.1. K-Nearest Neighbor (K-NN)

K-NN is lazy learning or instance-based learning where the function is approximated locally and all computations are deferred until classification. The K-NN algorithm is the simplest machine learning algorithms [20]. The K-NN algorithm is a non-parametric method used for classification and regression in pattern recognition. The output depends on whether KNN is used for classification or regression and in K-NN classification, the output is a class membership.

3.2.2. Support Vector Machine (SVM)

Support Vector Machine finds an optimal solution by Maximizing the distance between the hyper plane and the difficult points close to decision boundary. SVM is a classification method for both linear and nonlinear data [21]. Kernel and cost are the two parameters to select very good accuracy in typical domains and which are extremely robust. It is used both for classification and prediction. SVM is widely used in different areas like object detection and recognition, content-based image retrieval, text recognition, biometrics, Speech recognition and benchmarking time-series prediction tests, etc.

3.2.3. Logistic Regression

In statistics, Logistic regression is a probabilistic statistical classification model [22]. The Binary response predicted from a binary predictor by using logistic regression. It is also used to predicting the outcome of a categorical dependent variable (i.e., a class label) based on one or more predictor variables (features). It is also used in estimating

the parameters of a qualitative response model. It measures the relationship between a categorical dependent variable and one or more independent variables. Logistic regression is used in different areas such as medicine, marketing, engineering and economics.

3.2.4. Decision Induction Tree: C4.5

Decision tree is a flow-chart-like tree structure in which internal node represents a test on anattribute, branch represents an outcome of the test and leaf nodes represent class labels. Decision tree classification has been used for predicting medical diagnoses [23]. In Pruning phase the sub trees are eliminated which may over fit the data. This enhances the accuracy of a classification tree. It can handle continuous and discrete attributes. Decision tree classifiers provide human readable rules of classification, easy interpretation, faster decision tree construction and yields better accuracy.

3.2.5. Random Forest

Random forest is most accurate learning algorithm. It runs efficiently on large databases. It can handle thousands of input variables without variable deletion. It can estimate importance of the variable and an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing [24]. It has methods for balancing error in class population unbalanced data sets. Generated forests can be saved for future use on other data. Prototypes are computed that give information about the relation between the variables and the classification.

3.3. Measures Considered

Most of the performance measures for two-class problems are built over a 2x2 confusion matrix as illustrated in Table 2. Confusion matrix is a visualization tool which is commonly used to present the accuracy of the classifiers in classification. It is used to show the relationships between outcomes and predicted classes. A classifier is evaluated by a confusion matrix, the columns show the predicted class and the rows show the actual class. The entries in the confusion matrix are as follows:

TP is true positive, the number of positive cases that are correctly identified as positive; FN is false negative, the number of positive cases that are misclassified as negative cases; FP is false positive, the number of negative cases that are incorrectly identified as positive cases; TN is true negative, the number of negative cases that are correctly identified as negative cases [25].

TABLE 2: Confusion Matrix for a two-class Problem

	Positive Prediction	Negative Prediction
Active Positive Class	TP	FN
Active Negative Class	FP	TN

The most frequently used metrics for measuring the performance of learning systems are the error rate and the accuracy. In environments with imbalanced data, alternative metrics that measure the classification performance on positive and negative classes independently are needed. Table 3 presents the most well known the fundamental evaluation metrics.

TABLE 3: Fundamental Evaluation Metrics.

Measure	Formula	Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall Effectiveness of the algorithm by estimating the probability of the true value of the class label
Errorrate= 1-accuracy	$\frac{FP+FN}{TP+TN+FP+FN}$	Estimation of misclassification probability according to model prediction
Sensitivity (or) Recall	$\frac{TP}{TP+FN}$	Accuracy of Positive Samples or a measure of completeness
Specificity	$\frac{TN}{TN+FP}$	Accuracy of Negative examples
Precision	$\frac{TP}{TP+FP}$	Measure of correctness(i.e., out of positive labeled examples, how many are really a positive examples)

Statstical Measures

Along with accuracy other measures of performance have been considered. Statistical Measures are used to find the efficiency of the classification algorithms. Those are kappa statistics; Mean Absolute Error [MAE]; Root Mean Squared Error [RMSE]. One of the most familiar measures is Kappa statistics.

Kappa Statistics

Kappa is intended to give the reader a quantitative measure of the magnitude of agreement between observers. The calculation is based on the difference between how much agreement is actually present ("observed" agreement-Po) compared to how much agreement would be expected to be present by chance alone ("expected" agreement-Pe). The kappa value ranges from 0.1 to a maximum of 1.00 with poor agreement to very good agreement respectively [26]. It is calculated as: Kappa = (Po-Pe) / (1-Pe)

Mean Absolute Error (MAE)

It is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. It is a linear score which means that all the individual differences are weighted equally in the average.

MAE= $\sum(|f(xi) - yi|)/N$, where xi is an actual known value and yi is a predicted value.

Root Mean Squared Error (RMSE)

The RMSE is a quadratic scoring rule which measures the average magnitude of the error. Expressing the formula in words, the difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable.The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude .Both the MAE and RMSE can range from 0 to ∞. They are negatively-oriented scores: Lower values are better [27].

RMSE = $\sqrt{MSE} = \sqrt{\sum(f(xi) - yi)2/N}$, where xi is an actual known value and yi is a predicted value.

4. EXPERIMENTS & ANALYSIS

The experimental results of the five classifiers K-NN, SVM, Logistic Regression, C 4.5 and Random Forest on original four breast cancer datasets (i.e, without resampling) are tabulated in Table 4. Through the results it is observed that C4.5 has better accuracy rate ie., 75.17% over other four classifiers for BC dataset. SVM has produced higher accuracy of classification on both WDBC and WBC datasets such as 97.72% and 97.00% respectively than other classifiers. Logistic regression classifier has better accuracy i.e., 79.80% of classification on WPBC dataset than the other four classifiers which we have considered in this study. To improve the classifiers performance a pre-processing technique i.e, resampling is applied on the datasets and then by applying the classiifers the obtained results are tabulated in Table 5.

By comparing the results it is observed that by using resampling technique it is observed that all classifiers accuracy rate is increased on all the four datasets which we considered. To verify this fact, the accuracies of all the classifiers on all data sets without resampling and with resampling are tabulated in Table 6.

From the Tble 6, one can clearly observe that by applying resampling technique the classifiers performance is improved than without resampling. Further it is observed that for all the four datasets i.e, BC, WDBC, WBC and WPDC both K-NN and Random Forest are having better classifying accuracy rates than the other three classifiers i.e., SVM, Logistic regression and C4.5. The same graphically is represented in Figure 1.

From the Table 7 it is observed that kappa statistic also improved by applying classifiers with resampling than without resampling. For all the data sets Kappa values are increased after applying resampleing technique and then performing classification. It's a good sign because it is already mentioned that kappa value towards 1.00 is a good agreement between observed and expected accuracy. Through the results one can observe that by applying resampling on the data sets not only the classifiers performance improved the kappa values are also improved. Especially with K-NN and Randaom Forest classifiers there is a great improvement in the kappa values. Graphically also one can observe in Figure 2.

The results show that not only the accuries improved after applying resample technique. Further the MAE and RMSE rates are also decreased which indicates the fine performance of the classifiers.

5. CONCLUSION

In this study the popular classification algorithms: K-Nearest Neighbour, Support Vector Machine, Logistic Regression, C4.5 and Random Forest are considered to examine their performace on four Breast Cancer data sets: Breast cancer (BC), Wisconsin Diagnostic Breast Cancer (WDBC), Wisconsin Breast Cancer (WBC) and Wisconsin Prognostic Breast Cancer (WPBC). First the classidiers performance on original data sets ie., without resampling is studied and then the classifiers performance with resampling on data sets are also experimented and the results are compared. By this it is clear that after resmapling all the classifiers performance on all the data sets is improved. Not only calssfication accuracies improved the kappa value also increased. Moreover the MAE and RMSE rates also decreased. Over all to diagonosis the breast cancer on the four breast cancer data sets with resmapling, K-NN and Random forest classifiers are preferable. To improve the accuracy of the classifiers further study will be conducted by using feature extraction and feature selection techniques on the same data sets.

TABLE 4: Classfiers Performance on Data Sets (Without Resampling)

Classifiers	Accuracy	Kappa statistics	MAE	RMSE	Sensitivity	specificity	Time (Sec)
BC DATASET							
K-NN	67.83%	0.2029	0.3312	0.535	0.4	0.796	0
SVM	68.18%	0.1488	0.3182	0.5641	0.282	0.851	0.24
Logistic Regression	68.18%	0.1677	0.3795	0.4677	0.318	0.836	0.4
C.4.5	75.17%	0.2872	0.36	0.4329	0.294	0.945	0.07
Random Forest	65.73%	0.1326	0.3792	0.4791	0.329	0.796	0.12
WDBC DATASET							
K-NN	95.96%	0.9135	0.0422	0.2007	0.943	0.969	0
SVM	97.72%	0.9507	0.0228	0.1512	0.948	0.994	0.02
Logistic Regression	93.50%	0.8618	0.0657	0.2549	0.929	0.938	0.07
C.4.5	93.15%	0.8544	0.0741	0.2579	0.925	0.936	0.03
Random Forest	95.25%	0.8986	0.0742	0.1842	0.939	0.961	0.05
WBC DATASET							
K-NN	95.28%	0.8948	0.0473	0.2128	0.917	0.972	0
SVM	97.00%	0.9337	0.03	0.1733	0.963	0.974	0.09
Logistic Regression	96.57%	0.924	0.0486	0.1667	0.95	0.974	0.2
C.4.5	95.14%	0.893	0.0637	0.2142	0.942	0.956	0.1
Random Forest	96.14%	0.9143	0.0558	0.1686	0.938	0.974	0.18
WPBC DATASET							
K-NN	72.73%	0.2467	0.2752	0.5194	0.426	0.821	0
SVM	75.76%	0.0528	0.2424	0.4924	0.064	0.974	0.18
Logistic Regression	79.80%	0.4252	0.2359	0.4237	0.532	0.881	0.26
C.4.5	74.75%	0.2704	0.2859	0.4762	0.404	0.854	0.14
Random Forest	77.78%	0.2133	0.3263	0.4221	0.213	0.954	0.11

TABLE 5: Classfiers Performance on Data Sets (With Resampling)

Classifiers	Accuracy	Kappa statistics	MAE	RMSE	Sensitivity	Specificity	Time (Sec)
BC DATASET							
K-NN	85.66%	0.6646	0.1431	0.3324	0.764	0.898	0
SVM	69.58%	0.2363	0.3042	0.5515	0.382	0.838	0.24
Logistic Regression	70.98%	0.281	0.3512	0.4498	0.427	0.838	0.4
C.4.5	79.02%	0.4419	0.3108	0.4156	0.438	0.949	0.07
Random Forest	85.31%	0.6488	0.2077	0.3215	0.719	0.914	0.12
WDBC DATASET							
K-NN	98.42%	0.9674	0.0173	0.1256	0.979	0.988	0
SVM	97.54%	0.9489	0.0246	0.1569	0.945	0.997	0.02
Logistic Regression	97.72%	0.9528	0.0234	0.1502	0.966	0.985	0.03
C.4.5	97.01%	0.9382	0.0337	0.1724	0.953	0.982	0.01
Random Forest	98.07%	0.9601	0.0464	0.1296	0.97	0.988	0.04
WBC DATASET							
K-NN	97.71%	0.9483	0.0215	0.1401	0.953	0.989	0
SVM	96.71%	0.9273	0.0329	0.1814	0.983	0.959	0.02
Logistic Regression	96.14%	0.9137	0.0509	0.1683	0.953	0.966	0.03
C.4.5	96.28%	0.9168	0.0437	0.1902	0.953	0.968	0.01
Random Forest	98.00%	0.9551	0.0386	0.1288	0.974	0.983	0.03
WPBC DATASET							
K-NN	87.88%	0.679	0.1245	0.3465	0.776	0.913	0
SVM	86.36%	0.5791	0.1364	0.3693	0.531	0.973	0.02
Logistic Regression	87.37%	0.6633	0.1371	0.3532	0.755	0.913	0.1
C.4.5	85.86%	0.6255	0.1577	0.3612	0.735	0.899	0.03
Random Forest	87.88%	0.6401	0.1934	0.2934	0.612	0.966	0.03

TABLE 6: Classifiers Accuracy on Data sets with and without resampling

DATASETS	BREAST CANCER		WDBC		WBC		WPDC	
	Accuracy		Accuracy		Accuracy		Accuracy	
CLASSIFIERS	Without Resample	With Resample	Without Resample	With Resample	Without Resample	With Resample	Without Resample	With Resample
K-NN K=1 E	67.83%	85.66%	95.96%	98.42%	72.73%	97.71%	72.73%	87.88%
SVM	68.18%	69.58%	97.72%	97.54%	75.76%	96.71%	75.76%	86.36%
LOGISTIC	68.18%	70.98%	93.50%	97.72%	79.80%	96.14%	79.80%	87.37%
J48	75.17%	79.02%	93.15%	97.01%	74.75%	96.28%	74.75%	85.86%
RANDOM FOR-EST	65.73%	85.31%	95.25%	98.07%	77.78%	98.00%	77.78%	87.88%

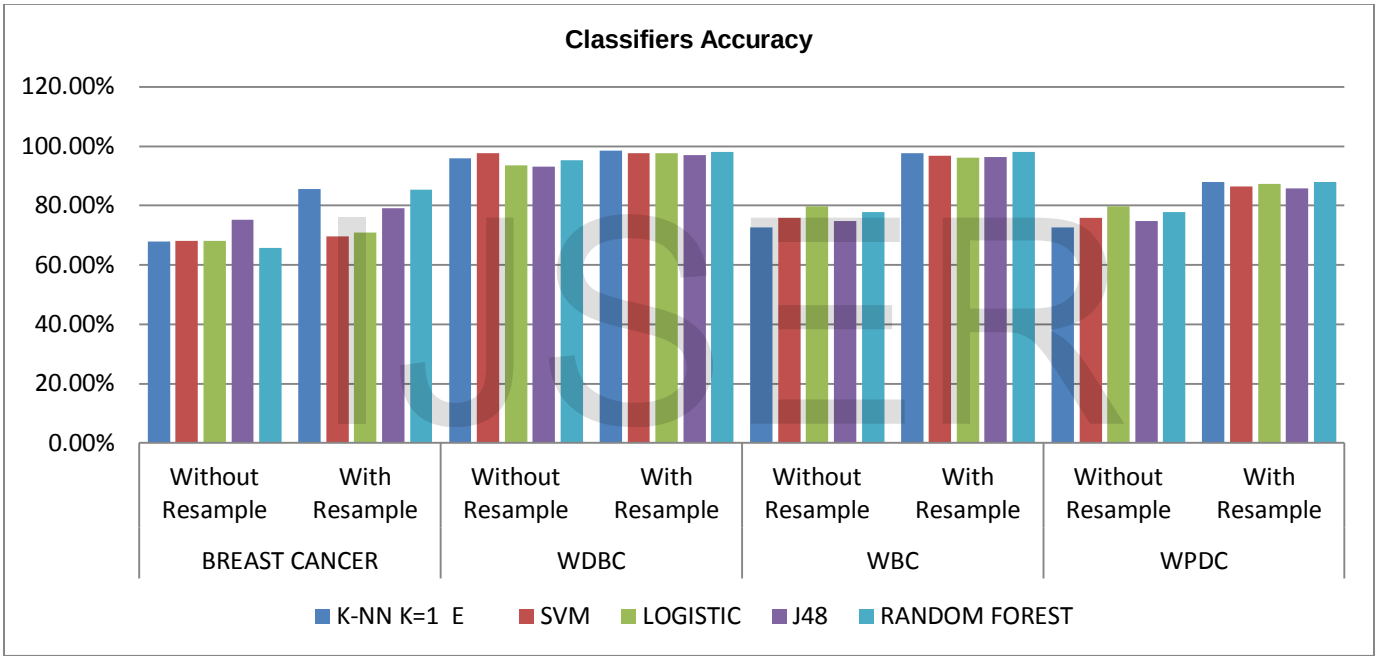

Fig1. Classifiers Accuracy

Table 7: Kappa Statistics

DATASETS	BREAST CANCER		WDBC		WBC		WPDC	
CLASSIFIERS	Without resample	With resample	Without resample	With resample	Without resample	With resample	Without resample	With resample
	KAPPA STATISTICS	KAPPA STATISTICS	KAPPA STATISTICS	KAPPA STATISTICS	KAPPA STATISTICS	KAPPA STATISTICS	KAPPA STATISTICS	KAPPA STATISTICS
K-NN K=1 E	0.2029	0.6646	0.9135	0.9674	0.8948	0.9483	0.2467	0.679
SVM	0.1488	0.2363	0.9507	0.9489	0.9337	0.9273	0.0528	0.5791
LOGISTIC REGRESSION	0.1677	0.281	0.8618	0.9528	0.924	0.9137	0.4252	0.6633
C4.5	0.2872	0.4419	0.8544	0.9382	0.893	0.9168	0.2704	0.6255
RANDOM FOREST	0.1326	0.6488	0.8986	0.9601	0.9143	0.9551	0.2133	0.6401

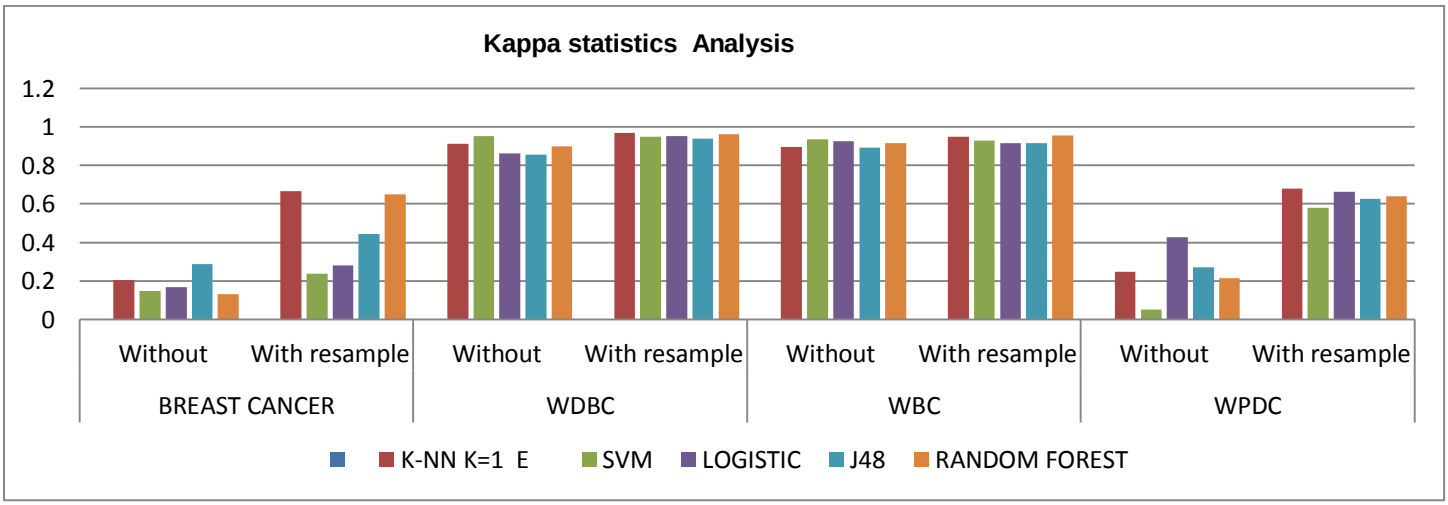

Fig 2. Kappa Statistics of the Classifiers

REFERENCES

[1] J. Han and M. Kamber," Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2000.

[2] Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of Cancer Disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012, pp. 55-66.

[3] Maaten L.J.P., Postma E.O. and Herik H.J. van den, 2007. Dimensionality reduction: "A comparative review", Tech. rep.University of Maastricht.

[4] Frawley, W., Batheus, C., 1991. Knowledge Discovery in Databases: An Overview. In Piatetsky-Shapiro, G. and Frawley, W. (Eds.), Knowledge Discovery in Databases, MIT Press,Cambridge, MA, pp1-27.

[5] V.Garcia et al., [2007] "The class imbalance problem in pattern classification and learning" , Martin Sewell 2007,"Ensemble Methods"

[6] P.Ramachandran, N.Girija and T.Bhuvaneswari, "Health care Service Sector: Classifying and finding Cancer spread pattern in Southern india using data Mining techniques", International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 05 May 2012, pp. 682-687.

[7] Nithya et.al [2011] "deriving classification rules for Indian rice diseases", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011.

[8] Orlando Anunciacao et.al[2010]," A Data Mining Approach for the Detection of High-Risk Breast Cancer Groups."

[9] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed, "Using datamining for assessing diagnosis of breast cancer," in Proc. International multi conference on computer science and information Technology, 2010, pp. 11-17.

[10] Robert Y,J. Lee , O. L. Mangasarian y& W. H. Wolberg, "Survival-Time Classification of Breast Cancer Patients"

[11] Delen Dursun , Walker Glenn and Kadam Amit , "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine ,vol. 34, pp. 113-127 , June 2005.

[12] Padmavati J., "A Comparative study on Breast Cancer Prediction Using RBF and MLP," International Journal of Scientific & Engineering Research, vol. 2, Jan. 2011.

[13] Mythili T et.al [2013] " A Heart Disease Prediction Model Using SVM-Decision Trees-Logistic Regression (SDL)", International Journal of Computer Applications (0975 – 8887), Volume 68– No.16, April 2013

[14] Dursun Delen*, Glenn Walker, Amit Kadam, Predicting breast cancer survivability: a comparison of three data mining methods Artificial Intelligence in Medicine (2004).

[15] Shelly Gupta, Dharminder Kumar and Anand Sharma , "Al Performance Analysis Of Various Data Mining Classification Techniques On

Healthcare Data ",International Journal of Computer Science & Information Technology (IJCSIT) ,Vol 3, No 4, August 2011.

[16] Chao chen etal., "using random forest to learn imbalanced data", Biometrics reaserch, Merk research labs.and Dept of statistics, UC Berkely.

[17] Anne-Laure Boulesteix et.al [2012] Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics, WIRE.

[18] Vrushali Y Kulkarni et.al [2014] Effective Learning and Classification using Random Forest Algorithm, IJEIT, Volume 3, Issue 11, May 2014.

[19] UCI Machine Learning Repository.[Online]. Availble: http://archive.ics.uci.edu/ml/datasets.html

[20] J. C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery", 2(2): 121-168, 1998.

[21] Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth International Group, 1984.

[22] Chao chen et.al.[1997], "using random forest to learn imbalanced data", Biometrics reaserch, Merk research labs.and Dept of statistics, UC Berkely.

[23] Gouda I.salama et al, " Breast cancer diagnosis on three different datasets using multi-classifiers", International journal of computer and information technology (2277-0764) ,volume 01-issue01,September 2012.

[24] Gopalakrishna Murthy at el., "Performance analysis and evaluation of different data mining algorithms used for cancer classification" ,IJARAI,Vol ,No.5, 2013

[25] Mohamed Bekkar et al., "Evaluation Measures for ModelsAssessment over Imbalanced Datasets",Iournal Of Information Engineering and Applications,Vol 3 ,No 10,2013.

[26] Anthony J. Viera, MD; Joanne M. Garrett, PhD, Understanding Interobserver Agreement: The Kappa Statistic ,Research series ,Family Machines,May2005

[27] http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/msg/ver_cont_var/uos3/uos3_ko1.htm

[28] Kung-Min Wang, Bunjira Makond, Wei-Li Wu, K-J Wang and Y.S.Lin, "Optimal data mining method for predicting breast cancer survivability", International Journal of Innovative Management, Information & Production, Volume 3, Number 2, June 2012, pp.28-33