Int J Pharm Bio Sci 2015 Jan; 6(1): (B) 177 - 187

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

PERFORMANCE COMPARISON FOR GEOGRAPHICALLY DISTINCT DATASETS FOR HEART DISEASE

R.KARTHIKEYAN[1], A.KUMARAVEL[2] AND A.CHANDRA SEKAR[3]

[1]Assistant Professor, Department of Computer Science and Engineering. Bharath University, Selaiyur, Chennai-600073, India
[2]Professor and Dean, Department of Computer Science and Engineering.Bharath University, Selaiyur, Chennai-600073, India,
[3] Professor,Department of Computer Science and Engineering. St.Joseph College of Engineering, Chennai, India,

ABSTRACT

Mining the data sets of different sizes or different regions many times will not yield expected maximum accuracy. Hence the data size becomes an important parameter for mining exercises. In this paper, we consider data from two different geographical regions and calculate separate performance measures. Also, we get the same for integrated data set obtained by the union of the original sets as inverse results establishing the hypothesis for integrated data set.

KEYWORDS: Heart disease data sets decision trees, decision rules, Meta classifiers, Bayes classifiers.

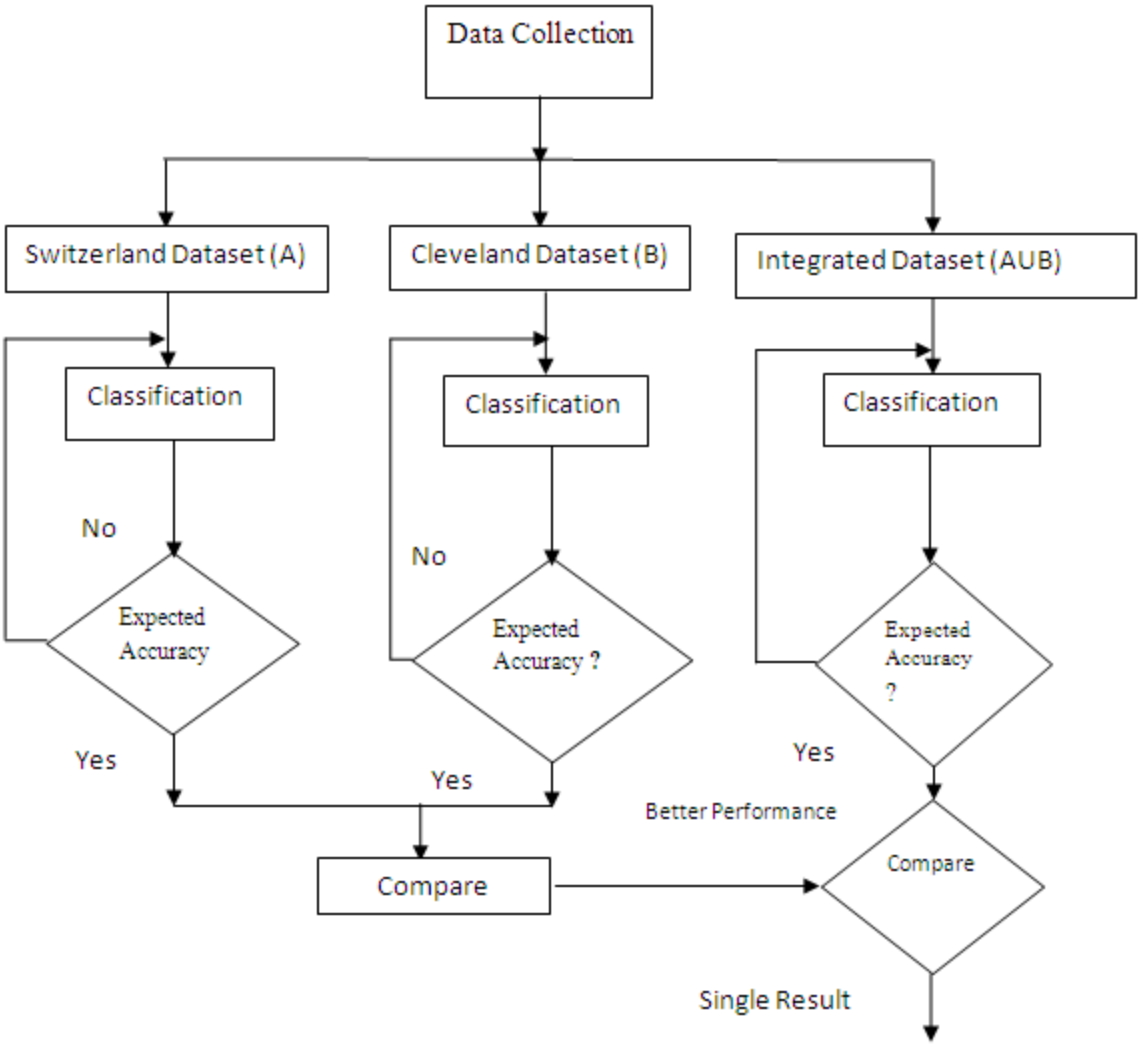
R.KARTHIKEYAN
Assistant Professor, Department of Computer Science and Engineering.
Bharath University, Selaiyur, Chennai-600073, India

***Corresponding author**

1. INTRODUCTION

The spectrum of strategies in data mining fields varies for the bio applications extensively. The main results depend on various parameters like data size, ratio of training and testing data, etc. This has been established as found in [4,5,6].The time gap in generations found in most of the time exhibits different levels of accuracy even though the learning models constructed are same. This is due to the natural evolution of changes occurs with respect to time in the data patterns which cannot be extracted easily. The same holds goods for geographically distinct distribution of datasets. In this paper, we establish such a difference as the result for the heart disease data set. We get the same for integrated data set obtained by the union of the original sets as inverse results establishing the hypothesis for integrated data set.

Figure 1
Diagram for Iterative Process involved in the Mining Process



2. DATA PREPARATION

In this section we describe the format of the data, the location of the availability of the data, the tools required for the experiments carried out etc. The diagram in fig 1 explains the flow of steps carried out in the main process. We also use weka 3.6.9, a java based tool from University of Waikato.

2.1 DATASET

The datasets has been slightly modified for these experiments in order to make it in 'attribute relation flat file (arff) format'. The data sets are downloaded from [14].

2.1.1 DATASET DESCRIPTION

The heart disease data sets describe with a set of fourteen attributes as shown in the list 2.2 below. The class attribute has discrete values {0,1,2,3,4}. The two distinct data sets obtained from two different geographical regions are Cleveland data and Switzerland data.

2.2 List of description of attributes

1. Age (In years)
2. Sex (1 = male; 0 = female)
3. Cp-chest pain type
4. Trestbps- testing blood pressure (in mm Hg on admission to the hospital)
5. Chol -serum cholestoral in mg/dl
6. Fbs -(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. Restecg- resting electrocardiographic results
8. Thalach maximum heart rate achieved
9. Exang -exercise induced angina (1 = yes; 0 = no)
10. Oldpeak ST depression induced by exercise relative to rest
11. Slope- the slope of the peak exercise ST segment
12. Ca number of major vessels (0-3) colored by flourosopy
13. Tha 3 = normal; 6 = fixed defect; 7 = reversible defect
14. Class {0,1,2,3,4 }

S.No	Attribute Name	Mean	Standard Deviation
1.	Age	54.4	9.0
2.	Sex	0.6	0.5
3.	cp	3.1	1.0
4.	Trestbps	131.6	17.6
5.	Chol	246.6	51.7
6.	Fbs	0.14	0.3
7.	Restecg	1.0	1.0
8.	Thalach	149.6	22.8
9.	Exang	0.3	0.5
10.	Oldpeak	1.0	1.2
11.	Slope	1.6	0.6
12.	Ca	0.6	0.9
13.	Tha	4.7	1.9

Table 1
Attribute table

2.3 Pie-chart for the Table 1

Dataset Name	0	1	2	3	4	Total
Cleveland	164	55	36	35	13	303
Switzerland	8	48	32	30	5	123
Integrated	51	56	41	42	10	200

Table 2
Distribution of data sets

S.No	Attribute Name	Cleveland		Sw data		Integrated	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
1.	Age	54.4	9.0	55.3	9.0	54.7	9.0
2.	Sex	0.6	0.5	0.9	0.27	0.75	0.4
3.	cp	3.1	1.0	3.7	0.69	3.31	0.9
4.	Trestbps	131.6	17.6	130.2	22.4	131.3	19.1
5.	Chol	246.6	51.7	0	0	175.5	120.13
6.	Fbs	0.14	0.3	0.10	0.19	0.13	0.31
7.	Restecg	1.0	1.0	0.36	0.6	0.8	0.94
8.	Thalach	149.6	22.8	121.5	25.9	141.5	26.95
9.	Exang	0.3	0.5	0.44	0.5	0.4	0.48
10.	Oldpeak	1.0	1.2	0.65	1.0	0.9	1.1
11.	Slope	1.6	0.6	1.8	0.6	1.65	0.6
12.	Ca	0.6	0.9	1.6	0.1	0.9	0.9
13.	Tha	4.7	1.9	5.8	1.3	5.0	1.8

Table 3
Statistical description of attributes

3. Methods Description

Here we select a standard set of methods [13] for predicting from the data set described above. We consider three types of classifiers f in our study,such as tree based, Bayes approach based, and Meta level based classifiers. The following sections describe briefly the methods for classifing and results of such methods are tabulated further. Then final results are interpreted

3.1 Tree Classifiers

Supervised Learning is performed, conducted using tree classifiers .We select four types of tree classifiers as shown below.

3.1.1 Decision Stump

A decision stump is a machine learning model consisting of a one-level decision tree. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes. A decision stump makes a prediction based on the value of just a single input feature

3.1.2 J48

The method J48 is categorized under decision tree type of classifier. A binary tree is created in this method. Once this binary tree is built, it is applied to each instance in the dataset and the class is determined. While building a tree, this method does not give significance for the missing values i.e. the attribute value can be predicted based on values in the other instances. It also allows classification via either decision trees or rules generated from them.

3.1.3 Random Forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution of all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of

features to split each node yields error rates that compares favorably to Adaboost , but are more robust with respect to noise. Internal estimates, monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

3.1.4 REP Tree

Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces

3.2 Bayes Classifiers

This type of classifiers includes probability measure for the class values and comes under supervised learning.

3.2.1 Bayes Net

Bayes Network learning using various search algorithms and quality measures. Base class for a Bayes Network classifier. Provides data structures and facilities common to Bayes Network learning algorithms like K2 and B.

3.2.2 Naïve Bayes

Class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an Updateable Classifier you need the Updateable Classifier functionality, use the Naïve Bayes Updateable classifier. The Naïve Bayes Updateable classifier will use a default precision of 0.1 for numeric attributes when build Classifier is called with zero training instances.

3.2.3 Naive BayesUpdateable

Class for a Naive Bayes classifier using estimator classes. This is the updateable version of Naïve Bayes. This classifier will use a default precision of 0.1 for numeric attributes when build Classifier is called with zero training instances.

3.2.3 NaïveBayesMultinominal

Class for building and using a multinomial Naive Bayes classifier. A Bayesian network, Bayes network, belief network, Bayesian model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

3.3 Meta Classifiers

Most of the time, the aggregation of more than one classifier has better performance. Such combinational methods are shown below.

3.3.1 Adaboost

Class for boosting a nominal class classifier using the Adaboost M1 method. Only nominal class problems can be tackled. Often dramatically improves performance, but sometimes over fits.

3.3.2 Bagging

Class for bagging a classifier to reduce variance. Can do classification and regression depending on the base learner. Generate B bootstrap samples of the training data: random sampling with replacement. Train a classifier or a regression function using each bootstrap sample For classification: majority vote on the classification results. For regression: average on the predicted values. Reduces variation. Improves performance for unstable classifiers which vary significantly with small changes in the data set, e.g., CART. Found to improve CART a lot, but not the nearest neighbor classifier.

3.3.3 Stacking

Combines several classifiers using the stacking method. Can do classification or regression.

3.3.4 Logit Boost

This classifier is for performing additive logistic regression. This class performs classification using a regression scheme as the base learner, and can handle multi-class problems. Boosting is a machine learning meta-algorithm for reducing bias in supervised learning.

3.3 Rules Classifiers

Most of the time, the aggregation of more than one classifier has better performance. Such combinational methods are shown below.

3.3.1 Zero R

Class for building and using a 0-R classifier. Predicts the mean (for a numeric class) or the mode (for a nominal class).

3.3.2 JRip

This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER),

3.3.3 OneR

Class for building and using a 1R classifier; in other words, uses the minimum-error attribute for prediction, discretizing numeric attributes. For more information

3.3.4 PART

Class for generating a PART decision list. Uses separate-and-conquer. Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule.

Table 4

Result with Rules Classifiers for Switzerland data set

Rules Classifiers	Accuracy	Mean Absolute Error	Maximum Accuracy	Minimum MAE
ZeroR	39.0244	0.2875	43.0894(JRip)	0.2707
JRip	43.0894	0.2707		
OneR	27.6423	0.2894		
PART	30.8943	0.2826		

The Maximum value obtained for the classifiers based on Rules Learning algorithms (Models of the format in the set of Rules) is 43.0894 and it is generated by JRip classifier due to minimum MAE

Table 5

Result with Tree Classifiers for Switzerland data set

Tree Classifiers	Accuracy	Mean Absolute Error	Maximum Accuracy	Minimum MAE
Decision Stump	32.5203	0.2849	42.2764(J48)	0.2437
J48	42.2764	0.2437		
Random forest	42.2764	0.2667		
REPTree	34.9593	0.2812		

The Maximum value obtained for the classifiers based on Trees Learning algorithms (Models of the format in the decision tree) is 42.2764 and it is generated by J48 classifier due to minimum MAE

Table 6

Result with Meta Classifiers for Switzerland data set

Meta Classifiers	Accuracy	Mean Absolute Error	Maximum Accuracy	Minimum MAE
AdaBoostM1	32.5203	0.2849	41.4634(Bagging)	0.2757
Bagging	41.4634	0.2757		
Stacking	39.0244	0.2875		
Logit boost	40.6504	0.2688		

The Maximum value obtained for the classifiers based on Meta Learning algorithms (Models of the format in the Composition of base classifiers) is 41.4634 and it is generated by Bagging classifier due to minimum MAE

Table 7
Result with Bayes Classifiers for Switzerland data set

Bayes Classifiers	Accuracy	Mean absolute Error	Maximum Accuracy	Minimum MAE
BayesNet	39.0244	0.2868	39.0244(Bayes net)	0.2868
NaiveBayes	37.3984	0.274		
NaiveBayesUpdatable	37.3984	0.274		
NaïveBayesMultinominal Text	39.0244	0.2875		

The Maximum value obtained for the classifiers based on Bayes Learning algorithms (Models of the format in the probability Expression for likelihood) is 39.0244and it is generated by Bayes net classifier due to minimum MAE

Table 8
Cumulative Performance for Switzerland dataset.

S.No	Types of Accuracy	Classification names	Maximum Accuracy	Minimum MAE
1	Rules	JRip	43.0894	0.2707
2	Tree	J48	42.2764	0.2437
3	Meta	Bagging	41.4634	0.2757
4	Bayes	Bayesnet	39.0244	0.2868

Table 9
Result with Rules Classifiers for with Cleveland data set.

Rules Classifiers	Accuracy	Mean Absolute Error	Maximum Accuracy	Minimum MAE
ZeroR	54.1254	0.2591	54.1254(JRip)	0.2529
JRip	54.1254	0.2529		
OneR	52.1452	0.1914		
PART	48.8449	0.2085		

The Maximum value obtained for the classifiers based on Rules Learning algorithms (Models of the format in the set of Rules) is 54.1254 and it is generated by JRip classifier due to minimum MAE

Table 10
Result with Tree Classifiers for with Cleveland data set.

Tree Classifiers	Accuracy	Mean Absolute Error	Maximum Accuracy	Minimum MAE
Decision Stump	52.8053	0.2324	58.7459(Random forest)	0.204
J48	51.4851	0.2115		
Random forest	58.7459	0.204		
REP Tree	57.7558	0.2108		

The Maximum value obtained for the classifiers based on Trees Learning algorithms (Models of the format in the decision tree) is 58.7459 and it is generated by Random forest classifier due to minimum MAE

Table 11
Result with Meta Classifiers for with Cleveland data set

Meta Classifiers	Accuracy	Mean Absolute Error	Maximum Accuracy	Minimum MAE
AdaBoostM1	52.8053	0.304	58.4158(Bagging)	0.207
Bagging	58.4158	0.207		
Stacking	39.0244	0.2591		
Logitboost	40.4504	0.1904		

The Maximum value obtained for the classifiers based on Meta Learning algorithms (Models of the format in the Composition of base classifiers) is 54.1254 and it is generated by Bagging classifier due to minimum MAE

Table 12
Result with Bayes Classifiers for with Cleveland data set.

Bayes Classifiers	Accuracy	Mean absolute Error	Maximum Accuracy	Minimum MAE
Bayes Net	57.0957	0.1872	57.0957(Bayes net)	0.1872
Naïve Bayes	56.4356	0.1843		
Naïve Bayes Updatable	56.4356	0.1843		
Naïve Bayes Multinominal Text	54.1254	0.2591		

The Maximum value obtained for the classifiers based on Bayes Learning algorithms (Models of the format in the probability Expression for likelihood) is 57.0957and it is generated by Bayesnet classifier due to minimum MAE

Table 13
Cumulative Performance for Cleveland dataset.

S.No	Types of Accuracy	Classification names	Maximum Accuracy	Minimum MAE
1	Rules	JRip	54.1254	0.2529
2	Tree	Random forest	58.7459	0.204
3	Meta	Bagging	58.4158	0.207
4	Bayes	Bayes net	57.0957	0.1872

The Maximum value obtained for the classifiers based on Rules Learning algorithms (Models of the format in the set of Rules) is48.1221 and it is generated by OneR classifier due to minimum MAE

Table 14
Result with Rules Classifiers for with Integrated data set.

Rules Classifiers	Accuracy	Mean Absolute Error	Maximum Accuracy	Minimum MAE
ZeroR	40.3756	0.2916	48.1221(OneR)	0.2075
JRip	44.8357	0.2775		
OneR	48.1221	0.2075		
PART	47.1831	0.2166		

The Maximum value obtained for the classifiers based on Rules Learning algorithms (Models of the format in the set of Rules) is48.1221 and it is generated by OneR classifier due to minimum MAE

Table 15
Result with Tree Classifiers for with Integrated data set.

Tree Classifiers	Accuracy	Mean Absolute Error	Maximum Accuracy	Minimum MAE
Decision Stump	47.4178	0.2533	49.7653(Random forest)	0.2254
J48	47.1831	0.266		
Random forest	49.7653	0.2254		
REPTree	43.8967	0.2244		

The Maximum value obtained for the classifiers based on Trees Learning algorithms (Models of the format in the decision tree) is 49.7653 and it is generated by Random forest classifier due to minimum MAE

Table 16
Result with Meta Classifiers for with Integrated data set.

MetaClassifiers	Accuracy	Mean Absolute Error	Maximum Accuracy	Minimum MAE
AdaBoostM1	47.4178	0.2533	52.3474(Bagging)	0.2276
Bagging	52.3474	0.2276		
Stacking	40.3756	0.2916		
Logitboost	50.4695	0.2173		

The Maximum value obtained for the classifiers based on Meta Learning algorithms (Models of the format in the Composition of base classifiers) is 52.3474 and it is generated by Bagging classifier due to minimum MAE.

Table 17
Result with Bayes Classifiers for with Integrated data set.

BayesClassifiers	Accuracy	Mean absolute Error	Maximum Accuracy	Minimum MAE
BayesNet	50.4693	0.2062	50.7042(NaiveBayes)	0.2009
NaiveBayes	50.7042	0.2009		
NaiveBayesUpdatable	50.7042	0.2009		
NaïveBayesMultinominal Text	40.3756	0.2916		

The Maximum value obtained for the classifiers based on Bayes Learning algorithms (Models of the format in the probability Expression for likelihood) is 50.7042and it is generated by NaiveBayes Bayes classifier due to minimum MAE.

Table 18
Cumulative Performance for Integrated dataset.

S.No	Types of Accuracy	Classifier name	Maximum Accuracy	Minimum MAE
1	Rules	OneR	48.1221	0.2075
2	Tree	Randomforest	49.7653	0.2254
3	Meta	Bagging	52.3474	0.2276
4	Bayes	NaiveBayes	50.7042	0.2009

Figure 2
For relative performance of individual and integrated heart disease datasets SW-Switzerland: CL-Clever land: In-Integrated

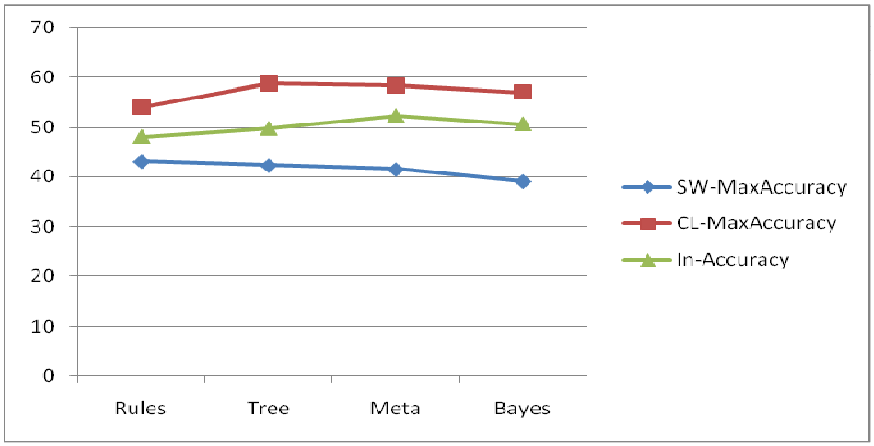


Figure3
For performance of individual heart disease datasets Integrated.

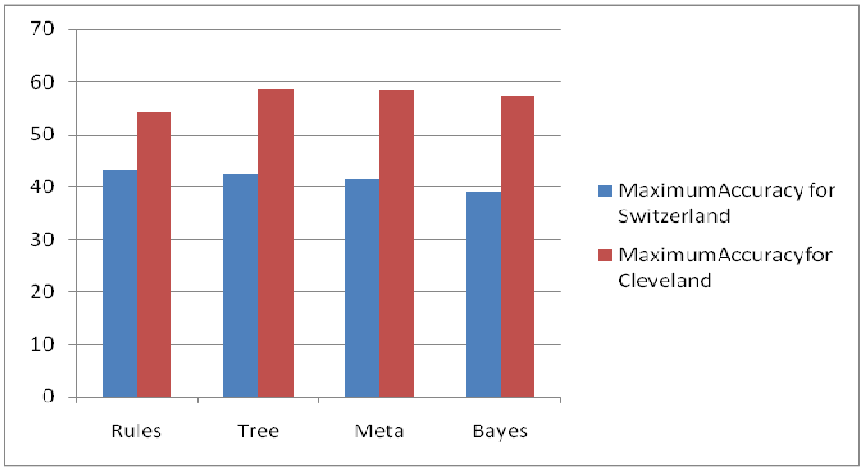
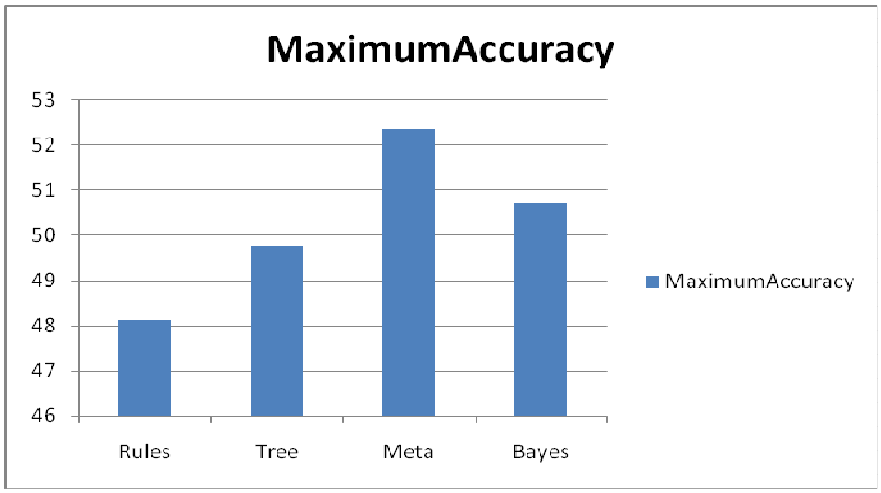


Figure 4
For performance of individual heart disease datasets Switzerland and Cleveland



4.1 RESULTS

The above chart shows the data set for Switzerland is trained with more accuracy than that of Cleveland giving around 40% for the former and around 60% in the later.

4.2 CONCLUSION

The accuracy of the integrated data set lies between the accuracies of component datasets as clearly shown in Fig 2 .Moreover the tables 7,12,17 show this behavior .In our future work this behavior can be further confirmed with more number of the component data sets.

ACKNOWLEDGEMENT

The authors would like to thank the management of Bharath University for the support and encouragement for this research work.

REFERENCES

1. https://www.waset.org/journals/waset/v68/v68-21.pdf world academy of science, engineering and technology, 2012.(Accessed on 9th july 2014)
2. Companion slides for the text by Dr..H.Dunham, *Data Mining, Introductory andAdvanced Topics*, Prentice Hall, 2002
3. Sourceboutwekahttp://www.cs.waikato.ac.nz/ml/weka/downloaded on 10 july 2014.
4. A.Kumaravel.,Pradeepa.R,Efficientmolecule reduction for drug design by intelligent search methods Int J Pharm Bio Sci, 4(2): (B) 1023 – 1029(2013)
5. A.Kumaravel., D.Udhayakumarapandian, A Novel Subset Selection For Classification Of Diabetes Dataset By Iterative Methods Int J Pharm Bio Sci ,5 (3) : (B) 1 – 8(2014)
6. A.Kumaravel., D.Udhayakumarapandian, Consruction Of Meta Classifiers For Apple Scab Infections Int J Pharm Bio Sci, 4(4): (B) 1207 – 1213(2013).
7. A.Gelman, Y. S. Su, M.Yajima, J. Hill, M.Pittau, J. Kerman, and T. Zheng, "arm:Data Analysis Using Regression and Multilevel/Hierarchical Models," R package version 1.5-02.://CRAN.Rproject.org/package=arm, (2012).
8. L.Breiman,"RandomForests,"in*MachineLearning*, vol. 45, pp. 5-32, 2001.

9. http://ipm.ncsu.edu/apple/chptr5.html(Accessed on 30th july 2013)
10. Dietterich, T. G., Jain, A., Lathrop, R.,Lozano-Perez, T.. A comparison of dynamic reposing and tangent distance for drug activity prediction.Advances in Neural Information Processing Systems, 6. SanMateo, CA: Morgan Kaufmann. 216—223, (1994)
11. A.Stensvand.,T. Amundsen.,L.Semb, D.M.Gadoury.,R.C.Seem,Ascospore release and infection of apple leaves by conidia and ascospores of Venturia inaequalis at low temperatures.Phytopathology 87:1046-1053, (1997)
12. http://weka.sourceforge.net/doc/weka/(Downloaded on 15th July 2014.
13. UCI-reprository ,https://archive.ics.uci.edu/ml/datasets.html downloaded on 10th July 2014.