Comparison of Machine learning Algorithm for Optimal Classification Based on Liver function Test

Dr. M. Mayilvaganan[1], R.Rajamani[2]
Associate Professor[1], Assistant Professor[2],
[1]Department of Computer Science, [2]PSG College of Arts and Science,
[12]Coimbatore.

Abstract- The liver function tests can determine the nature of problem. Along with this test to find the normal range values of different parameters based on Child-Turcotte- Pugh Classification for severity. In this paper, the work is focuses on to classify the parameter of Bilirubin, Albumin, Prothrombin time (INR), Ascites, Encephalopathy which obtaining the total score by adding score of each parameter for measuring the severity of liver disease based on rule based classification and apply different machine learning algorithm for analysing the efficiency of machine learning algorithm which one of the algorithm is better to analysis.

Keyword- bagging with C4.5 algorithm; AdaBoost with C4.5 algorithm; C4.5 with 10 fold- cross validation; bagging with random forest; attribute selection with C4.5 algorithm

I. INTRODUCTION

Machine learning is programming computers to optimize a performance criterion with respect to the given parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data. In this work, a performance can be analysed for diagnosis a liver function test by learning the level of fluctuation of parameter as per the severity found. In this research, to compare supervised learning machine learning techniques can be applying such as bagging with C4.5 algorithm, AdaBoost with C4.5 algorithm, C4.5 with 10 fold- cross validation and bagging with random forest and attribute selection with C4.5 algorithm. The result and experimental performances can be analysing in this research[1][2][3].

II. DATA PREPARATION

Table 1 can be representing as the collection of parameter which has to measure the severity score of liver function test. To obtaining the total score by adding score of each parameter for measuring the severity of liver disease as points under Child-Turcotte-Pugh Score[19][20][21][22].

Table 1 List of parameter follow for measuring the Liver function test

Encephalopathy (E)	
None	(1 point)
Grade 1: Altered mood/confusion	(2 points)
Grade 2: Inappropriate behavior, impending stupor, somnolence	(2 points)
Grade 3: Markedly confused, stuporous but arousable	(3 points)
Grade 4: Comatose/unresponsive	(3 points)
Bilirubin (B)	
<2 mg/dL	(1 point)
2-3 mg/dL	(2 points)
>3 mg/dL	(3 points)
Albumin (A)	
>3.5 g/dL	(1 point)
2.8-3.5 g/dL	(2 points)
<2.8 g/dL	(3 points)
Prothrombin time prolongation (P)	
Less than 4 seconds above control/INR <1.7	(1 point)
Prothrombin time prolongation (P)	
4-6 seconds above control/INR 1.7-2.3	(2 points)
More than 6 seconds above control/INR >2.3	(3 points)
Ascites (AS)	
Absent	(1 point)
Slight	(2 points)
Moderate	(3 points)

CTP Score Interpretation (CTPScore)	
5 to 6 points	Child Class A
7 to 9 Points	Child Class B
10 to 15 Points	Child Class C

Table 2 Sample Training data

E	B	A	P	AS	CTPScore
1	1	1	1	1	Class A
3	3	2	2	3	Class C
2	2	1	1	1	Class B
1	2	2	1	2	Class B
1	2	2	1	1	Class B
3	3	2	2	3	Class C
1	1	1	1	1	Class A
3	3	3	3	3	Class C
2	1	2	2	1	Class B
1	3	3	2	2	Class C
1	1	1	1	1	Class A
1	1	1	1	1	Class A
1	1	1	1	1	Class A
3	3	3	2	2	Class C
1	2	1	1	1	Class A
3	3	3	3	3	Class C
1	1	1	2	1	Class A
1	1	1	1	2	Class A

Table 2 represents the patient's liver function test, 500 training data for analysing the severity of the liver disease diagnosis in early for affected person to avoid the severe problem. In this research, to compare bagging with random forest, bagging with C4.5, AdaBoost with C4.5 algorithm, C4.5 with 10 fold-cross validation and bagging with random forest and attribute selection with C4.5 algorithm can be implement for analysing the performance of efficiency of algorithm.

III. PROPOSED RESEARCH MEHODOLOGY

Random forest classifier is a meta-estimator that fits a number of decision trees on various sub- samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting[9][10][11][12]. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement[18][19][20].

In bagging, the same approach is used, but instead for estimating entire statistical models, most commonly decision trees [5]. Multiple samples of your training data are taken then models are constructed for each data sample. When to make a prediction for new data, each model makes a prediction and the predictions are averaged to give a better estimate of the true output value.

In bagging with C4.5 algorithm, to improves generalization error by reducing the variance of the base classifiers [6]. The performance of bagging depends on the stability of the base classifier. After training the classifiers of CTPscore, a test instance is assigned to the class that receives the highest number of votes.

AdaBoost was the first really successful boosting algorithm developed for binary classification. It is the best starting point for understanding boosting. Modern boosting methods build on AdaBoost, most notably stochastic gradient boosting machines [7]. Stochastic gradient descent is a simple and very efficient approach to fit linear models. It is particularly useful when the number of samples is very large. It supports different loss functions and penalties for classification.

AdaBoost is used with short decision trees. After the first tree is created, the performance of the tree on each training instance is used to weight how much attention the next tree that is created should pay attention to each training instance. Training data that is hard to predict is given more weight, whereas easy to predict instances are given less weight [8][13]. Models are created sequentially one after the other, each updating the weights on the training instances that affect the learning performed by the next tree in the sequence[14][15][16][17]. After all the trees are built, predictions are made for new data, and the performance of each tree is weighted by how accurate it was on training data.

IV. EXPERIMENTAL ANALYSIS

In experiment analysis, the True Positive rate (TP rate), False Positive rate (FP), Precision, Recall, F-Measure, ROC Area classified based on Class can measured for all algorithm. In bagging with C4.5 by pruned data and its detailed accuracy 97.4% as shown in table 3 and confusion matrix can shown the relevant diagnosis of patient's severity in class wise as shown in table 4.

Table 3 Detailed Accuracy by Class of Bagging with C4.5

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.932	0.005	0.972	0.932	0.952	0.944	0.999	0.995	Class A
	1.000	0.065	0.986	1.000	0.993	0.960	0.999	1.000	Class C
	0.556	0.010	0.667	0.556	0.606	0.595	0.983	0.695	Class B
Weighted Avg.	0.974	0.054	0.972	0.974	0.973	0.944	0.999	0.988	

Table 4 Confusion matrix

a	b	c	<-- classified as
69	0	5	a = Class A
0	408	0	b = Class C
2	6	10	c = Class B

When experiment in AdaBoost with C4.5 by pruned data, the number of iteration is 10. The detailed accuracy of AdaBoost is 98.6% as shown in table 5 and confusion matrix can shown the relevant diagnosis of patient's severity in class wise as shown in table 6.

Table 5 Detailed Accuracy by Class of AdaBoost with C4.5

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.986	0.002	0.986	0.986	0.986	0.984	1.000	0.999	Class A
	0.998	0.043	0.990	0.998	0.994	0.966	0.999	1.000	Class C
	0.722	0.004	0.867	0.722	0.788	0.784	0.993	0.831	Class B
Weighted Avg.	0.986	0.036	0.985	0.986	0.985	0.963	0.999	0.994	

Table 6 Confusion Matrix

a	b	c	<-- classified as
73	0	1	a = Class A
0	407	1	b = Class C
1	4	13	c = Class B

When experiment in Attribute Selection with C4.5 and detailed accuracy is 96.6% as shown in table 7 and confusion matrix can shown the relevant diagnosis of patient's severity in class wise as shown in table 8.

Table 7 Detailed Accuracy by Class of Attribute Selection with C4.5

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.959	0.000	1.000	0.959	0.979	0.976	0.993	0.986	Class A
	0.985	0.087	0.980	0.985	0.983	0.906	0.938	0.966	Class C
	0.556	0.019	0.526	0.556	0.541	0.523	0.676	0.306	Class B
Weighted Avg.	0.966	0.072	0.967	0.966	0.966	0.903	0.937	0.945	

Table 8 Confusion Matrix

a	b	c	<-- classified as
71	0	3	a = Class A
0	402	6	b = Class C
0	8	10	c = Class B

When experiment in bagging with random forest and detailed accuracy is 98% as shown in table 9 and confusion matrix can shown the relevant diagnosis of patient's severity in class wise as shown in table 10.

Table 9 Detailed Accuracy by bagging with random forest

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.986	0.007	0.961	0.986	0.973	0.969	1.000	0.999	Class A
	0.998	0.054	0.988	0.998	0.993	0.960	1.000	1.000	Class C
	0.556	0.004	0.833	0.556	0.667	0.671	0.995	0.846	Class B
Weighted Avg.	0.980	0.046	0.978	0.980	0.978	0.951	1.000	0.994	

Table 10 Confusion Matrix

a	b	c	<-- classified as
73	0	1	a = Class A
0	407	1	b = Class C
3	5	10	c = Class B

When experiment in C4.5 based on the 10-fold validation test mode and its detailed accuracy is 96.6% as shown in table 11 and confusion matrix can shown the relevant diagnosis of patient's severity in class wise as shown in table 12.

Table 11 Detailed Accuracy by C4.5 based on the 10-fold validation test mode

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.959	0.000	1.000	0.959	0.979	0.976	0.993	0.986	Class A
	0.985	0.087	0.980	0.985	0.983	0.906	0.938	0.966	Class C
	0.556	0.019	0.526	0.556	0.541	0.523	0.676	0.306	Class B
Weighted Avg.	0.966	0.072	0.967	0.966	0.966	0.903	0.937	0.945	

Table 12 Confusion Matrix

a	b	c	<-- classified as
71	0	3	a = Class A
0	402	6	b = Class C
0	8	10	c = Class B

Table 13 Time efficiency

Algorithm	Time taken in seconds
Bagging with Random Forest	0.8
AdaBoost with Random Forest	0.79
C4.5 with 10- fold validation	0.02
Attribute Selection with C4.5	0.03
AdaBoost with C4.5	0.06
Bagging with C4.5 Out of Bagging	0.09

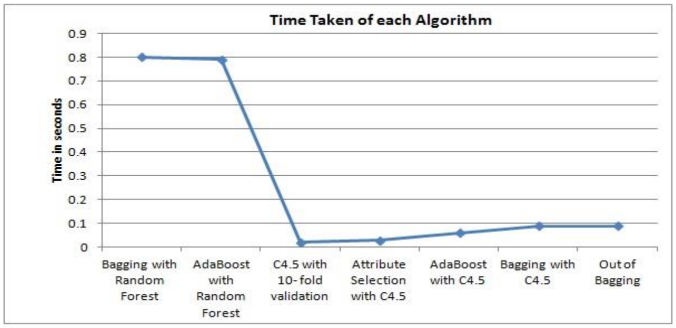

Fig.1: Time taken of each Algorithm

Fig.1 represents the time taken of each algorithm during execution of run time. From this analysis, Bagging with Random forest algorithm takes 0.8 seconds, AdaBoost with Random Forest can takes 0.79 seconds, C4.5 with 10- fold validation can takes 0.02 seconds but its accuracy is lower than AdaBoost C4.5 algorithm, Attribute Selection with C4.5 can takes 0.03 seconds, AdaBoost with C4.5 takes 0.06 seconds and Bagging with C4.5 and Out of Bagging can takes 0.09 seconds as shown in table 13.

Table 14 : Accuracy Comparison of Different Ensemble methods

Algorithm	Classified Instance Accuracy	In classified Instance Accuracy	Mean absolute error	Relative absolute error	Root relative squared error	Kappa statistic	Root mean squared error
Bagging with Random Forest AdaBoost with Random Forest	98%	2%	0.0172	8.24 %	27.34 %	0.9343	0.088
C4.5 with 10- fold validation	96.6%	3.4%	0.0295	14.11 %	44.91 %	0.8898	0.1446
Attribute Selection with C4.5	96.6%	3.4%	0.0295	14.11 %	44.91 %	0.8898	0.1446
AdaBoost with C4.5	98.6%	1.4%	0.0105	5.04 %	30.51 %	0.9543	0.0982
Bagging with C4.5	97.4%	2.6%	0.0268	12.84 %	34.01 %	0.914	0.1095
Out of Bagging	96.356%	3.644%	0.0323	15.46 %	42.46%	0.8805	0.1368

From the analysis an accuracy of different machine learning can be experimented by the input of different parameter in liver function test. Here, an accuracy of Bagging with Random Forest and AdaBoost with Random Forest can achieved 98%, C4.5 with 10- fold validation and Attribute Selection with C4.5 can achieved 96.6%, AdaBoost with C4.5 can achieved 98.6%, Bagging with C4.5 can achieved 97.4% and Out of Bagging of C4.5 can achieved 96.356% as shown in fig.2. Finally, Mean absolute error, Relative absolute error, Root relative squared error, Kappa statistic, Root mean squared error can be measured in different algorithm s shown in table 14.

From this analysis, when to compared with different algorithm, AdaBoost with C4.5 can give better result of classification slightly than bagging and AdaBoost algorithm and then more vary than other algorithm.

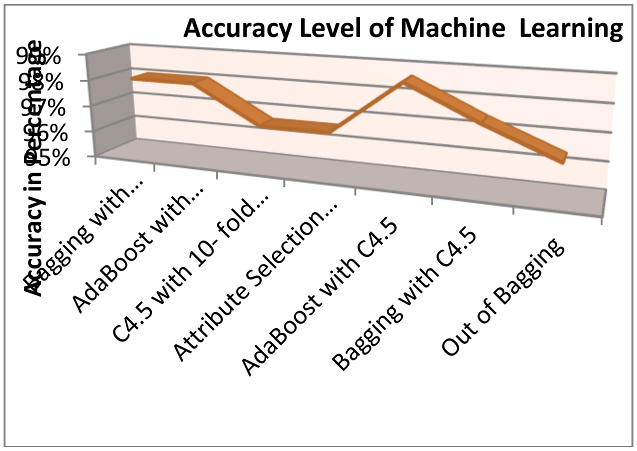

Fig.2: Accuracy Level of Machine Learning

V. CONCLUSION

In machine learning techniques, the execution of the run program to optimize the parameters of the model using the training data liver function test. In this work, a performance can be analysed for diagnosis a liver function test by learning the level of fluctuation of parameter as per the severity found by using an algorithm of Bagging with Random Forest and AdaBoost with Random Forest can achieved 98%, C4.5 with 10- fold validation and Attribute Selection with C4.5 can achieved 96.6%, AdaBoost with C4.5 can achieved 98.6%, Bagging with C4.5 can achieved 97.4% and Out of Bagging of C4.5 can achieved 96.356%. From this experiment AdaBoost C4.5 can classify the data in binary format and easily achieved in 98.6% as an accuracy of classification. Which better result when compared with other.

VI.REFERENCES

[1]. Hall, Mark, et al. "The WEKA data mining software: an update." *ACM SIGKDD* Explorations Newsletter 11.1 (2009): 10-18.

[2]. T.F. Gonzales. Clustering to minimize the maximum inter cluster distance. Theoretical Computer Science,1985,38(2-3):293-306.

[3]. Kannan, M., S. Prabhakaran, and P. Ramachandran. "Rainfall forecasting using data mining technique."(2010)

[4]. Arun K Pujari, 2003, "Data mining techniques", University Press (India).

[5]. Jiawei Han Micheline Kamber, 2006, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publisher an imprint of Elsevier

[6]. L. Breiman, J. Friedman, R. Olshen and C. Stone. "Classification and Regression Trees", Wadsworth International Group, Belmont, CA, 1984.

[7]. Quinlan, J.R. (2003). "C5.0 Online Tutorial", http://www.rulequest.com.

[8]. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z., Steinbach,M., Hand, D. J and Steinberg, D (2008). "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, 14 (1): 1-37.

[9]. Wolpert, D. (1992). "Stacked generalization", Neural Networks, 5: 241-259.

[10]. Schapire, R. (1990). "The strength of weak learnability", Machine Learning, 5(2): 197-227. [11]. Breiman, L. (1996a). "Bagging Predictors", Machine Learning, 24(2): 123-140.

[11]. Breiman, L (2001). "Random Forests". Machine Learning 45 (1): 5–32.

[12]. Freund, Y. Schapire, R. (1996). "Experiments with a new boosting algorithm", In Proceedings of the Thirteenth International Conference on Machine Learning, 148-156 Bari, Italy.

[13]. Dietterich, T. G. (2000). "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization". Machine learning, 40: 139-157.

[14]. Kotsiantis, S and Pintelas, P. (2004). "Local Boosting of Weak Classifiers", Proceedings of Intelligent Systems Design and Applications (ISDA 2004), August 26-28, Budapest, Hungary.

[15]. Opitz, D and Maclin, R (1999) "Popular Ensemble Methods: An Empirical Study", 11: 169-198.

[16]. Chap T. Le (1997). "Applied survival analysis", Wiley, New York.

[17]. Quinlan, J. R. (1996) "Bagging, Boosting and C4.5", AAAI/IAAI, 1: 725-730.

[18]. Endo, A, Shibata, T and Tanaka, H (2008) "Comparison of Seven Algorithms to Predict Breast Cancer Survival", Biomedical Soft Computing and Human Sciences, 13(2), pp.11-16.

[19]. Banfield, R.E, Hall, L.O, Bowyer, K.W and Kegeimeyer, W. P. "A comparison of decision tree ensemble creation techniques" (2007), IEEE Transactions on pattern analysis and machine intelligence, 29: 173-180.

[20]. Pugh RN, Murray-Lyon IM, Dawson JL, et. al. Transection of the oesophagus for bleeding oesophageal varices. *Br J Surg*. 1973; 60:646. PMID: 4541913

[21]. Child CG, Turcotte JG. The Liver and Portal Hypertension. Philadelphia, WB Saunders Co. 1964. NLMN: 46218

[22]. Trey C, Burns DG, Saunders SJ. Treatment of hepatic coma by exchange blood transfusion. *NEJM*. 1966; 274:473. PMID: 5904286