Performance Analysis and Error Evaluation Towards the Liver Cancer Diagnosis Using Lazy Classifiers for ILPD

Manish Tiwari[1($\boxtimes$)], Prasun Chakrabarti[2($\boxtimes$)],
and Tulika Chakrabarti[3($\boxtimes$)]

[1] Department of Computer Science and Engineering,
Mewar University, Chittorgarh 312901, Rajasthan, India
immanishtiwari@gmail.com
[2] Department of Computer Science and Engineering,
ITM Universe Vadodara, Paldi 391510, Gujarat, India
dean.research@itmuniverse.ac.in
[3] Department of Chemistry, Sir Padampat Singhania University,
Udaipur 313601, Rajasthan, India
tulika.chakrabarti@spsu.ac.in

Abstract. This paper, entails the various Lazy classifiers such as IBKLG, LocalKnn algorithm, RseslibKnn algorithm used for diagnosis of the liver cancer. The results have been noted in terms of both performance and errors. The performance analyzed based on the accuracy, precision and recall and error evaluation are based on the Mean absolute error, Root mean squared error, Relative absolute error and Root relative squared error. The LocalKnn is best in terms of accuracy and recall while IBKLG indicates best precision.

Keywords: IBKLG · LocalKnn · RseslibKnn · Accuracy · Precision
Recall root mean squared error · Relative absolute error · Root relative squared

1 Introduction

Liver is the largest organ after the skin in our body. It perform many functions cleansing blood toxins, converting food into nutrients to control hormone level. The diagnosis of liver diseases at early stage can improve survival rate of patient life. Techniques are used to find pattern from the large dataset are called the data mining techniques. it have several function such as classification, association rules and clustering etc. classification is supervised learning technique used for dataset in dissimilar group of classes or in different levels. Classification method performs two steps one is dataset are used to trained to built model and in second it used for classification [1].

2 Literature Survey

In the paper [2] Indian liver patient dataset and UCLA dataset were used. Analysis was done by ANOVA and MANOVA to recognize difference among the groups. Authors took common attributes e.g. ALKPHOS, SGPT and SGOT for both datasets. Analysis of Variance (ANOVA) was done using multivariate tables. Author investigated 99% and 90% significant levels and found the good results.

The study [3] deals with two distinct feature combinations viz SGOT, SGPT, and Alkaline Phosphates of two datasets (ILPD and BUPA liver disorder). Error rate, sensitivity, prevalence and specificity were exponentially observed. The attributes like total bilirubin, direct bilirubin, albumin, gender, age and total proteins facilitate in liver cancer diagnosis.

The paper [4] indicated neural network to train adaptive activation function for extracting rules. OptaiNET, an Artificial Immune Algorithm (AIS) was used to set rules for liver disorders. Based on input attribute adaptive activation was trained to use neural network extract rules efficiently in hidden layer. ANN to performs the data coding, to classifies coding data and finally extracts rules. It correctly diagnosed 192 samples (out of 200) belonging to class 0 covering 96% and 135 samples (out of 145) belonging to class 1 covering 93%. Entire samples correctly diagnosed 94.8%.

The study [5] pointed out univariate analysis and feature selection for predicator attributes. Predictive data mining is a significant tool for researchers of medical sciences. ILPD dataset was chosen for men and women. The classification algorithms were trained to test and to perform some results for accuracy and error analysis. For men and women the SVM gave high accuracy 99.76% and 97.7% respectively.

In the survey [6] classification algorithm decision tree induction (J48 algorithm) employing dataset from the Pt. B.D. Sharma Postgraduate Institute of Medical Science, Rohtak was used. The dataset contained 150 instances (100 instances for training purpose and 50 instances for the test data), 8 attributes and 2 classes for the model using 10 fold cross validation in WEKA tool and J48 algorithms classified correctly 100% instances. The result was expressed in four categories e.g. cost/benefit of J48 for class YES = 44, cost/benefit of J48 for class NO = 56, classification accuracy for YES = 56%, classification accuracy for NO = 44%. Many other algorithms on this dataset were applied and J48 algorithms showed best results.

The publication [7] described classification using data mining approaches on ILPD. Naïve bayes, Random Forest and SVM. The algorithms were implemented using R tool and for improving the accuracy the hybrid neuro SVM that is the combination of the SVM and feedforward Neural Network (ANN) was used. Root mean square error (RMSE) and mean absolute percentage error were pointed out. This model gave 98.83% accuracy.

In the publication on [1] various decision tree algorithms were used based on the data mining concept such as AD Tree, Decision Tree, J48, Random Forest, Random Tree on the liver cancer dataset. They were used for the training purpose and pre-processing was applied for missing or noisy data. Classification algorithms were performed with feature selection and without using feature selection. Its performances were measured in terms of Accuracy, Precision, and Recall. The accuracy (71.35%) of

the decision stump was very good compared to other algorithms and J48 and random forest gave 70.66% and 70.15% accuracy respectively.

The publication on [8] indicated PSO java to execute dataset and to categorize training attributes in order to retrieve pbest and gbest. The pbest was then compared with lbest to set the best solution for attribute selection. The PSO gave gammagt 4.60, alkphos 4.49, SGPT 3.91, SGOT 3.07, drinks 1.36. The selected dataset was applied to WEKA tool to perform the classification. Then it applied the Kstar algorithm. PSO-Kstar algorithm is the best data mining technique giving accuracy up to 100%.

The paper [9] described different clustering algorithms for predication on BUPA liver disorder and ILPD dataset for performance analysis. The simple BIZ model was selected effectively. Different attribute selections were done for accuracy, such as 5, 6, 7, 8 and 9. The logistic Regression and SVM (PSO) gave best results for the BUPA liver disorder as well as ILPD dataset, with accuracy 89.14% and 89.66% respectively.

3   Methodology

In this process the Indian liver patient dataset have been taken after the preprocessing is performed in this method the missing values problem are solved after the supervised filter are used in that resample method are used then Lazy classifier such as IBKLG, LocalKnn, RseslibKnn algorithms are used in WEKA tool for classification. 10 folds cross validations are used then performance and error evaluation is performed (Fig. 1).
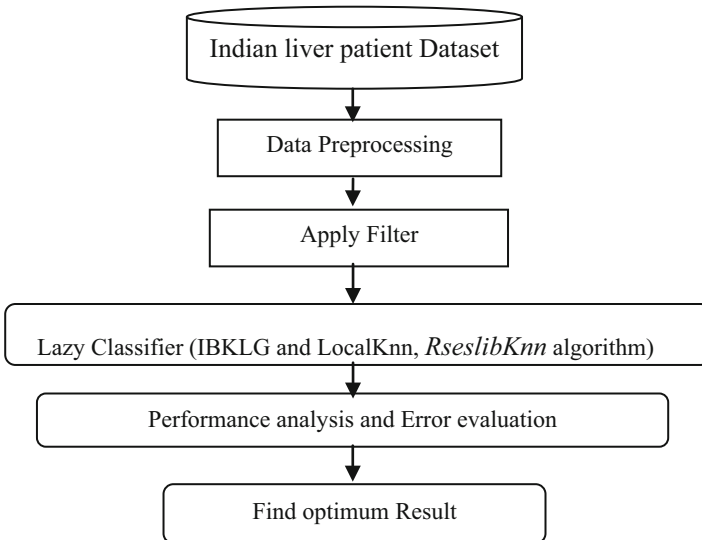

Fig. 1. Classification process

4   Result and Discussion

Lazy classifiers are used for analysis of the liver cancer disease. In this process any algorithm that gave better accuracy, precision and classified more correct instances is the good algorithm in term of early diagnosis of the liver cancer.

4.1   IBKLG Algorithm

IBKLG classifier is a part of lazy classifier. K-nearest neighbors classifier can select appropriate value of K based on cross-validation. It also performs distance weighting. It selects number of neighbor is one, The standard deviation set to 1.0, do not check capabilities to false, meanSquared value to false. It is based on nearest neighbor search algorithm using linearNNSearch algorithm. 10 folds cross validations are used for testing. It correctly classifies 573 instances (covering 98.28%) and incorrectly classifies 10 instances (covering 1.72%) out of 583 instances (Fig. 2, Tables 1 and 2).
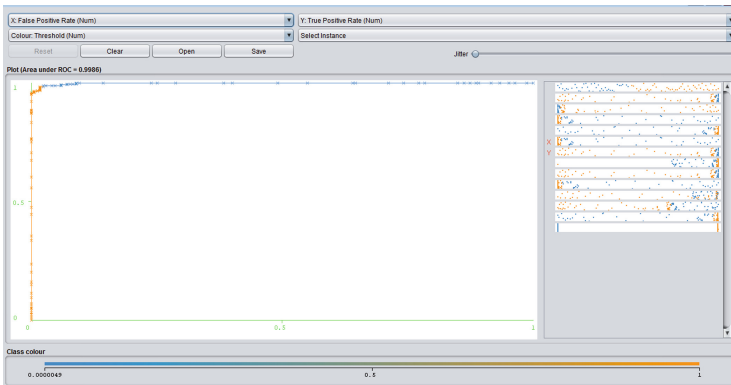

Fig. 2. Area under ROC for IBKLG algorithm with a value 0.9986

Table 1. Error evaluation for IBKLG algorithm.

Sr.No.	Type of error	Result
1	Mean absolute error	0.0172
2	Root mean squared error	0.1309
3	Relative absolute error	4.2006%
4	Root relative squared error	28.9595%

4.2   LocalKnn Algorithm

LocalKnn algorithm is based on K nearest neighbor classifier with local metric induction. It improves accuracy in relation to standard k-nn, particularly in case of data with nominal attributes. It works with reasonably 2000 + training instances. 100 batch size is selected. Do not check capabilities to set to false. Learning Optimal K values to

Table 2. Confusion matrix for IBKLG algorithm.

Performance vector:			
Confusion Matrix:			Accuracy: 98.28% (for class 1 malignant)
	M(T)	B(T)	Precision: 99.3% (for class 1 malignant)
M(P)	409	7	
B(P)	3	164	Recall: 98.3% (for class 1 malignant)
Class 1 is selected for the result because it mention positive in liver disorder			

true and number of neighbors used to vote for the decision to one, size of the local uses induce local metric to 100. The metric vicinity size for density based is 200. The voting for the decision by nearest neighbors is set to inverse square distance. It uses distance based weighting method. 10 fold cross validations are applied. It correctly classifies 576 instances (covering 98.80%) and incorrectly classifies 7 instances (covering 1.20%). Time taken to build model is 68.19 s (Fig. 3, Tables 3 and 4).
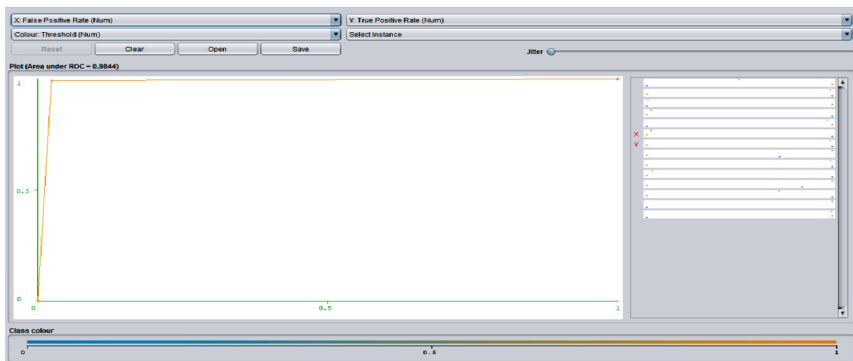

Fig. 3. Area under ROC for LocalKnn algorithm with a value 0.9844

Table 3. Error evaluation for LocalKnn Algorithm

Sr.No.	Type of error	Result
1	Mean absolute error	0.012
2	Root mean squared error	0.1096
3	Relative absolute error	2.9346%
4	Root relative squared error	24.2362%

Table 4. Confusion matrix for Local Knn Algorithm

Performance vector:			
Confusion Matrix:			Accuracy: 98.80% (for class 1 malignant)
	M(T)	B(T)	Precision 99.0% (for class 1 malignant)
M(P)	413	3	Recall 99.3% (for class 1 malignant)
B(P)	4	163	
Class 1 is selected for the result because it mention positive in liver disorder			

4.3 RseslibKnn Algorithm

RseslibKnn is a part of lazy classifier. It sets some properties defines such as batch size, learning optimal k value, do not check capabilities, cross validation, kernel setting, density based metric and so on. Time taken to building model is 1.3 s. 10 folds cross validations. It correctly classifies 571 instances (covering 97.94%) and incorrectly classifies 12 instances (covering 2.06%) out of 583 instances (Fig. 4, Tables 5 and 6).
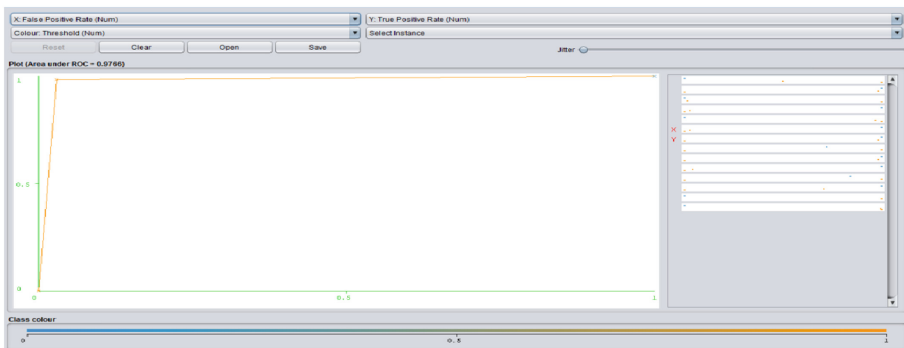

Fig. 4. Area under ROC for RseslibKnn algorithm with a value 0.9766

Table 5. Error evaluation for RseslibKnn algorithm

Sr.No.	Type of error	Result
1	Mean absolute error	0.0206
2	Root mean squared error	0.1435
3	Relative absolute error	5.0307%
4	Root relative squared error	31.7327%

Table 6. Confusion matrix for RseslibKnn algorithm

Performance vector:			
Confusion Matrix:			Accuracy: 97.94% (for class 1 malignant)
	M(T)	B(T)	Precision: 98.8% (for class 1 malignant)
M(P)	409	7	
B(P)	5	162	Recall: 98.3% (for class 1 malignant)
Class 1 is selected for the result because it mention positive in liver disorder			

4.4 Comparison of Error Evaluation and Performance Analysis of Three Lazy Classifiers (RselibKnn, IBKLG, LocalKnn) for ILPD Dataset
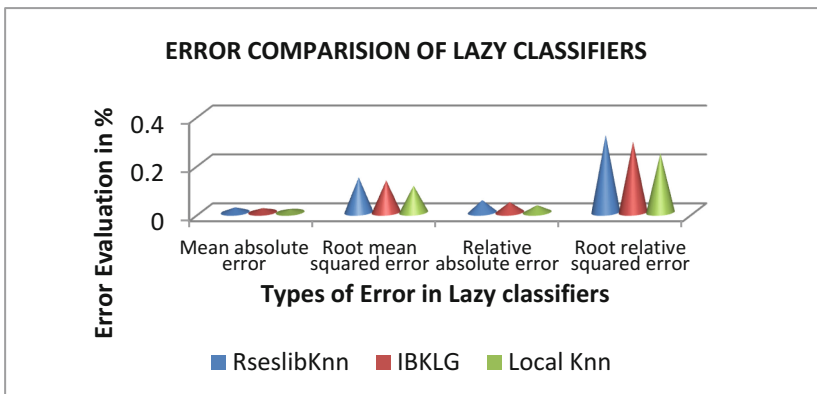
See Figs. 5 and 6.
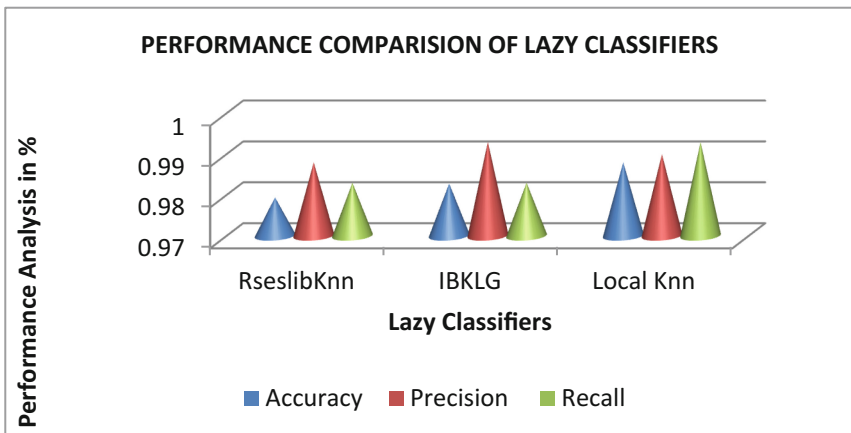

Fig. 5. Error evaluation of Lazy classifier

Fig. 6. Performance analysis of Lazy classifier

5   Conclusion and Future Perspective

A close assessment of error estimation of three Lazy classifiers (RseslibKnn, IBKLG, LocalKnn) has been performed whereby the minimum error value is achieved through LocalKnn. The LocalKnn is best in terms of accuracy and recall while IBKLG indicates best precision. It is evident that if any classification algorithm classifies instances accurately, then diagnosis of the liver cancer can be done easily and accurately in early stages.

Further research work or classifiers can be applied on different types of cancers such as Breast cancer, Prostate Cancer, Lung cancer etc. Appling these algorithms may generate better results. As an extension of this Biopsy and mammography images can be used for analysis using machine learning methods. Research can also be applied for analysis of survival rate of the patient.

References

1. Manochitra, V., Shajahaan, S.: Performance amelioration to model liver patient data using decision tree algorithms. J. Appl. Sci. Res. **11**(23), 161–167 (2015)
2. Venkata Ramana, B., Prasad Babu, M.: A critical comparative study of liver patients from USA and INDIA: an exploratory analysis. Int. J. Comput. Sci. Issues **9**(3), 506–516 (2012)
3. Hashem, E.M., Mabrouk, M.S.: A study of support vector machine algorithm for liver disease diagnosis. Am. J. Intell. Syst. **4**(1), 9–14 (2014)
4. Kahramanli, H., Allahverdi, N.: A system for detection of liver disorders based on adaptive neural networks and artificial immune system. In: Proceedings of the 8th WSEAS International Conference on Applied Computer Science, Venice, Italy, pp. 25–30 (2008)
5. Tiwari, A., Sharma, L.: Comparative study of artificial neural network based classification for liver patient. J. Inf. Eng. Appl. **3**(4), 1–5 (2013)
6. Reetu, N.K.: Medical diagnosis for liver cancer using classification techniques. Int. J. Recent Sci. Res. **6**(6), 4809–4813 (2015)
7. Nagaraj, K., Sridhar, A.: NeuroSVM: A Graphical User Interface for Identification of Liver Patients. Int. J. Comput. Sci. Inf. Technol. **5**(6), 8280–8284 (2014)
8. Thangaraju, P., Mehala, R.: Performance analysis of PSO-KStar classifier over liver diseases. Int. J. Adv. Res. Comput. Eng. Technol. **4**(7), 3132–3137 (2015)
9. Mazaheri, P., Norouzi, A.: Using algorithms to predict liver disease classification. Electron. Inf. Plan. **3**, 256–259 (2015)