Indian Journal of Science and Technology, *Vol 10(22), DOI: 10.17485/ijst/2017/v10i22/112117, June 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Optimizing Decision Tree Through Attributes Generation Using Genetic Programming for Clinical Data

Narander Kumar[1] and Sabita Khatri[2]

[1]Department of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Vidya Vihar, Raebareli Road, Lucknow – 226025, Uttar Pradesh, India; nk_iet@yahoo.co.in
[2]Department of Computer Engineering, University Institute of Engineering and Technology (UIET), Babasaheb Bhimrao Ambedkar University (A Central University), Vidya Vihar, Raebareli Road, Lucknow – 226025, Uttar Pradesh, India; sabitaarora@gmail.com

Abstract

Objective: To intend towards increasing classification efficiency of J48 classifier by introducing attribute set on the basis of applied genetic programming. The constructed set of attribute not only enhances the data classification capabilities of J48 but also increased the data space for the algorithm towards giving more accurate results. **Methods/Analysis:** The datasets related to heart and liver disease were selected from the UCI machine learning repositories. The experiment has been conducted with the help of WEKA tool, which is an open source tool for data mining. **Finding:** After experimentation it is found that the efficiency of J48 is giving better classification accuracy with reduced error rate when applied with datasets after inclusion of newly generated attributes by genetic programming. After adding attributes induced by genetic programming, significant efficiency boost can be seen in classification capabilities of J48 by 74% to 83% and 68% to 72% for heart and liver disease datasets respectively. **Improvement:** We obtained better results when compared to the existing literature for the chosen clinical datasets.

Keywords: Classification, Clinical Data, Decision Tree, Data Mining, Genetic Attribute

1. Introduction

Knowledge discovery in database is a goal of data mining as it involves required steps in order to identify some patterns that can turn into knowledge by applying on flood of data[1]. Various data mining algorithms are used to generate knowledge, out of which decision tree is most valuable algorithm which represents the data into hierarchy form and automatic construction of decision tree generates for multi disciplinary data is well defined in the literature[2]. The approach of genetic programming to improve classification quality has been tested in four different domain namely Synthetic Domain, Riplet Data Set, Pima Indian Diabetes, EEG Signal Classification[3]. By using the fitness value on heart disease dataset different classifier like Naïve Bayes, Support Vector Machine, Decision tree, Artificial Neural Network has been tested on different efficiency parameters like sensitivity, specificity, accuracy and error rate[4]. Combining fuzzy logic and genetic algorithm is called GAFL genetic algorithm fuzzy logic, which is an effective model for heart disease explained by Santhanam[5].

*Author for correspondence

Various approaches are defined for classification purpose, in which patients are considered in tree class as pre diabetic, diabetic and no diabetic with the help of multi class genetic programming[6]. Classification approach for heart disease by considering the highly contributed features are well defined and very close for prediction by using feature selection method with improved performance[7]. Genetic Programming has proven application capabilities in data mining domain, it was well proposed[8]. The concept of genetic programming based on natural selection has already been established, with the inclusion of the concept of survival of fittest[9]. The best set of attribute subset by wrapper approach based on multi objective genetic algorithm for decision tree classification approach has been well defined[10]. New attribute construction can be performed from original attribute by genetic programming for gaining more accurate result. Attribute construction method can be divided into hypothesis, data driven approach as well as preprocessing and interleaving approach. The data driven, preprocessing and GPCI, which involves genetic programming for attribute construction has been defined by different researchers[11]. Use of machine learning algorithm is nowadays become a curious approach towards finding hidden pattern from raw data. Also the application of genetic programming and genetic algorithm to pre process the data before its classification by decision tree can easily be noticed[12]. Classifiers can further be designed by the use of genetic programming with the use of chromosome operations and mutation operations to reduce the diverse effect of genetic programming and O-Ring chromosome[13]. The applicability of machine learning algorithms with genetic programming is well defined and accepted by many scientific organizations[14]. A comparative partner selection of genetic program mining can also be defined in further two stage concept in which first stage new attributes are generated from available ones and in next is used by different classifier to provide more accurate output[15]. Other Classification algorithm beside decision tree like Naïve Bayes has also given fruitful result for the prediction of disease by medical data set. The enhanced probability of correctly classified instances can be increased with Naïve Bayes algorithm for heart disease classification using genetic programming[16]. WEKA is an open source and well known data mining tool which is widely accepted for research in various domains of data mining[17].

2. Proposed Work of Genetic Attribute Construction and Implementation

Data Mining has vast applied and proven capabilities in various research domains, out of which predictive approach is most significant. Every predictive approach is based on some facts or attributes to define the predictive class, so attribute selection is mile stone for some prediction. The input data space is one of the dominant key factors and can be deterministic one for increasing efficiency of any classifier for a given dataset. Quality of prediction is firmly depends on such factors, which are highly involved toward affecting the performance of classifier algorithms. Preprocessing of data that includes consistency, noise removal with adequate attributes facts can also enhanced the predictive performance of classifier. Addition of new attributes with original one can also enhance the potential of prediction and generate the hidden pattern for raw data with more accuracy. In this paper authors has focused to define genetic programming algorithm to construct new attributes along with original one for given dataset. After that it is passed to subsequent data mining algorithm in order to obtain more correctly classified instances. Majority of cases, mining of raw data with any data mining technique is applied only on original attribute which limits the power of performance. So later, the construction of new attribute from existing one comes in consideration by koza[8], which is proved as a good enhancer of accuracy. In most of the prediction approach based on data mining algorithm has considered only original attributes. Dependency only on original attributes limits the power of performance, so to overcome this and to enhance the potential of performance new attribute are constructed. Attribute construction method is basically defined in two groups first is hypothesis based and data driven based methods. In hypothesis based method new attributes are generated by previously based according to their hypothesis until a satisfactory extended attribute set is not formed however in contrary to it in data driven based method it is not depended on hypothesis, construction of new attribute from original ones are based on relationship in data. Example of data driven method is GALA and GPCI, in both of the method first original attribute is converted in Boolean attributes then new attributes are generated by the combination of Boolean attributes. Both of example follow same basic concept only difference is GPCI used additionally the genetic

programming for the formation of new attribute. Attribute construction can also define in two groups like interleaving approach and pre processing approach. In interleaving approach construction of new attributes include the mining algorithm but preprocessing approach is independent to it. In preprocessing approach first preprocessing of data is done then new attribute is produce and then passes it to mining algorithms for pattern prediction.

The attribute construction involves all important phases of genetic programming, where individuals are defined in LISP tree like structure. All possible tree base structured are formed by function symbol and terminal symbol, the function in function set may be arithmetic symbol like +,-,*, /, ^, mathematical functions like sin, cos, exp, sqrt, Boolean function such as AND, OR, NOT, conditional function, iterative operation and other functions are used. All the features in given dataset are represented as floating point number and potential solution is find by using these functions(+, -, *, /, ^, sin, cos, In, Exp, Sqrt) and terminal set, selection operator, crossover mutation are combined used to solve the problems by genetic programming. Genetic programming further involves the generation of initial population. For this task grow method, full method and ramped half and half methods are used. In this research work author have chosen the ramped half and half method for their experimental purpose. The selection of next fittest population is one of most crucial task performed in genetic approach. The most popular selection algorithm are fitness per portion and tournament selection, both of the algorithm the goal is the individual with better fitness has more probability for being selected in comparison to one with having the worse fitness. Tournament selection algorithm is most widely used in genetic programming. For space generation crossover is used in genetic programming. Crossover is essentially a genetic operator, in machine learning algorithms the computer programs are define as genes which are altered to by genetic programming where the space of solution is generated by computer programming, so the performance can be improved. The mutation phase includes symbolic expression generation where terminals are substituted by new terminals or function symbols are substituted by some new function symbols.

They proposed data driven approach along with preprocessing approach. Combination of these two approach produce more efficient attribute for predictive resultant. In this paper genetic programming for attribute construction is used in medical domain to find hidden pattern of medical data to define it with more accuracy and precisely towards early disease prediction. In this research work, heart and liver disease dataset has been taken from UCI machine learning repository. When decision tree classification algorithm is implemented on these datasets using Waikato Environment for Knowledge Analysis (WEKA) to find correctly classified instances, it provide 74.7475.% accuracy in heart disease dataset and 68.6859% accuracy in liver disease dataset respectively however when preprocessing and data driven approach is implemented means new attributed are constructed from original attributes then passed to decision tree it enhances the prediction accuracy of decision tree classifier as 83.8384% on heart disease dataset and 72.4638% in liver disease dataset.

3. Materials and Methods

For experimentation purpose, authors have considered two disease datasets taken from UCI repositories related to heart and liver disorders are considered respectively. The first dataset and very primarily dataset is heart disease dataset. In this dataset total 14 attributed has been defined as age, sex, thal, number major vessel, slope of peak exercise ST segment, resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate achieved, chest pain, resting electrocardiographic result, ST depression, exercise induced angina, and num class for diagnosis of heart disease which whether the patient is suffering from heart disease or not. There are total 08 nominal and 06 numerical values are present with the total instances of 303 patients are present. The liver disorder dataset contains total 7 attribute has been included as alkaline phosphate, gamma-glutamyl transpeptidase, aspartate aminotransferase, alamine aminotransferase, mean corpuscular volume, drinks as alcoholic beverages, selector field is used to split data into two sets. There are 01 nominal and 06 numeric attributes with the total instances of 345. For better accuracy all instances carrying missing values have been removed. In heart dataset total 6 rows have been eliminated, so total instances considered for this experiment is 297 where as in liver disorder dataset complete dataset with a total 345 instances have been considered for the study. Then all attribute are classified by decision tree classifier on WEKA tool for analysis of accuracy, ROC, kappa statistics, TP, FP, precision, recall and f-measures. Next the genetic programming is applied on same data set by choosing WEKA's filter that is GP Attribute Generation, new attributes are generated with the help of set of function (+, -, *, /, ^, sin, cos,

In, exp, sqrt), max Depth = 5, max Time = 600, number Of Generations = 100, operator Porpotion = 0.9,0.1, population Generation Method = Ramped half and half, population size=100, seed=1, selection Method = Single Fitness, target Tree Accuracy = 1.0 and fitness Evalution Method used for classified the instances is J48. Table 1 and Table 2 show the confusion matrix for the J48 classifier on different datasets. Table 3 shows the details performance measures with J48 and J48+GP for decision tree classifier.

Table 1. Confusion matrix (heart disease dataset)

A	B	classified as
131	29	a = level 0
19	118	b = level 1

Table 2. Confusion matrix (liver disease dataset)

A	B	classified as
99	46	a = no
49	151	b = yes

Table 3. Classification Results and Accuracy Measures for Heart and Liver Disease Data

Performance Measures	Heart Disease Dataset		Liver Disease Dataset	
	J48	J48 + GA	J48	J48 + GA
Accuracy	74.747	83.8384	68.6957	72.4638
MAE	0.2922	0.178	0.3673	0.2886
RMSE	0.4677	0.3862	0.5025	0.4954
Kappa Values	0.49	0.6765	0.3401	0.4365
F-Measure	0.747	0.839	0.680	0.725
Precision	0.747	0.841	0.683	0.725
Recall	0.747	0.838	0.687	0.725
ROC s	0.740	0.845	0.665	0.741

4. Result and Discussion

The classification statistics shows the performance measure of the J48 classifier such as correctly classified instances also defined as accuracy, incorrectly classified instances, ROC, kappa statics and different error measures like MAE, RMSE. The correctly classified instances measure shows the ability of a classifier to correctly classify the data. The percentage of incorrectly classified instances shows the incorrectly classified instances of classifier for the given data set. In Figure 1 it has been shown that J48 tree based classifier shows an accuracy of 83.8384% with newly generated attribute by genetic programming. Figure 2 defines J48+ GA gives the MAE value 0.178 for heart and 0.2886 values for liver disease dataset and RMSE value 0.3862 and 0.4954 value for heart and liver disease dataset so the error rate is also reduces by implementation genetic programming on 10 fold classification. The kappa statics value 1 indicates the good chance of agreement between the classification in its classification task and the kappa statics value 0 indicates the disagreement among the group classified. In Figure 3 J48+GA shows the good kappa statics value 0.6765 and 0.4365 respectively for heart and liver disease dataset after classification. Precision and recall performance measures are shown in figure 4. Figure 5 depicted the ROC value 0.845 and 0.741 for heart and liver disease dataset respectively in case of J48+GA.

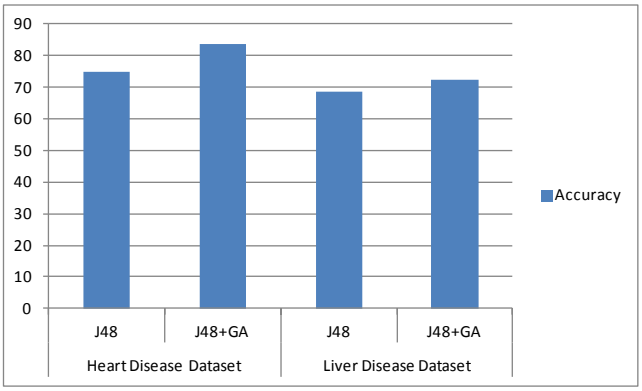

Figure 1. Classification accuracy values.

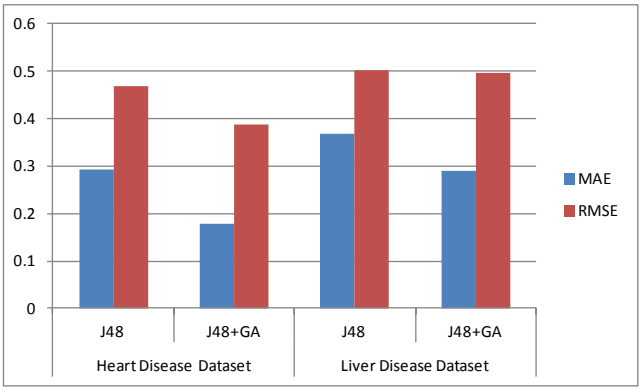

Figure 2. RMSE vs. MAE values.

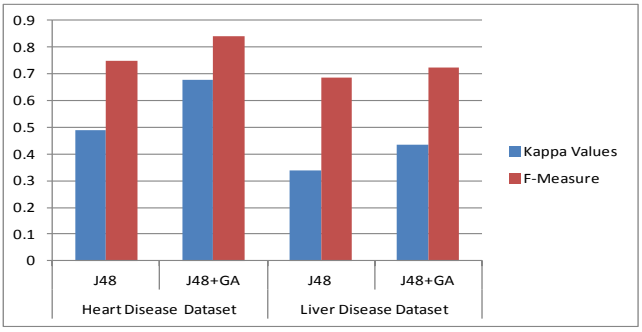

Figure 3. Kappa values vs. f-measure values.

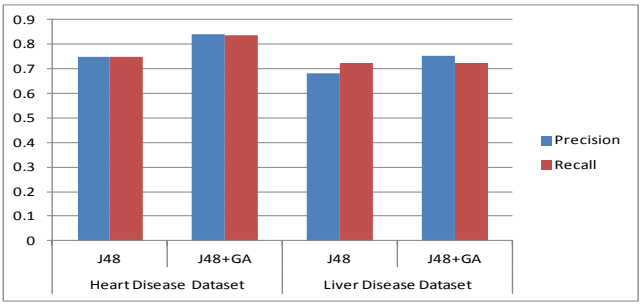

Figure 4. Precision and recall values.

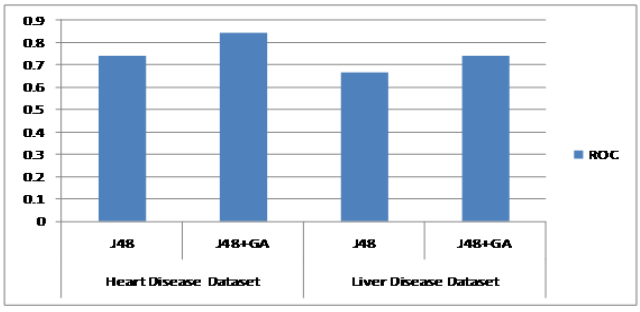

Figure 5. ROC for Heart and Liver Disease Dataset

If we compare the experimental results of this study with earlier results mentioned in the literature, then we find that most of the authors have measured the classification accuracy by considering the disease dataset with only original attribute. The objective of this research work is to enhance the predictive capability of J48 classifier in order to identify the disease with greater accuracy. This machine learning based model could be helpful to medical practitioner by strengthening the early prediction of diseases with a higher degree of accuracy. This leads to the fact that tree base method like J48 with genetic programming approach can be proved as better prediction classifier for medical data. The study has been conducted on two major disease datasets after adequate preprocessing and data driven approach the J48 classifier has been analyzed and implemented in this research work, which justified the prediction approach.

5. Conclusion

We performed on two different datasets with 10 fold cross validation and performance has been evaluated for the J48 classifier, the result shows the significant improvements in terms of accuracy and better predictive capabilities. The constructed attributes with the help of genetic programming not only helpful in improving the data space but also better chances for further predictions , the algorithm like J48 is significantly impacted by the added attribute set because it helps the tree generation in more accurate fashion. The chosen disease datasets heart and liver shows improvements of more than 9% and 6% for heart and liver disease dataset respectively. The value of kappa statics showing better chance of agreement, as well as ROC value is also on the higher side and error measuring parameter MAE and RMSE also reduces by applying genetic programming as well as TP Rate, Precision, Recall, F-Measure values for J48 classifier is better with extended attributes on same datasets. This research work can be further extended towards applying genetics for different classifier like Naïve Bayes, Multilayer perceptron, and KNN in order to enhance the performance on different disease datasets.

6. References

1. Fayyad U, Shapiro GP, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Communications of the Association for Computing Machinery (ACM). 1996 Nov; 39(11):27–34. Crossref.
2. Murthy SK. Automatic construction of decision trees from data: a multi-disciplinary survey. Data Mining Knowledge Discovery. 1998 Dec; 2(4):345–89. Crossref.
3. Estebanez C, Aler R, Valls JM. A method based on genetic programming for improving the quality of datasets in classification problems. International Journal of Computer Science and Applications. 2007; 4(1):69–80.
4. Paliwal P, Malviya M. An efficient method for predicting heart disease problem using fitness value. International Journal of Computer Science and Information Technologies. 2015; 6(2):1290–3.
5. Santhanam T, Ephzibah EP. Heart disease prediction using hybrid genetic fuzzy model. Indian Journal of Science and Technology. 2015 May; 8(9):797–803. Crossref.

6. Sonawane R, Patil S. Diabetes detection using genetic programming. International Journal of Computer Application. 2015 Oct; 127(10):12–6. Crossref.
7. Suganya R, Rajaram S, Abdullah AS, Rajendran V. A novel feature selection method for predicting heart disease with data mining techniques. Asian Journal of Information Technology. 2016; 15(8):1314–21.
8. Koza JR. Genetic programming. The MIT Press, Cambridge, MA. 1992; 4(2):87–112.
9. Sivanandan SN, Deepa SN. Introduction of genetic algorithm: evolutionary computation. Springer, Berlin; 2008. p. 1–13. Crossref.
10. Pappa GL, Kaestner AA. Attribute selection with a multi-objective genetic algorithm. In the Proceedings of the 16th Brazilian Symposium on Artificial Intelligence, Advances in Artificial Intelligence, Lecture Notes in Computer Science, Springer. 2002; 2507:280–90.
11. Otero FEB, Silva MMS, Freitas AA, Nievola JC. Genetic programming for attribute construction in data mining. In the Proceedings of the Association for Computing Machinery (ACM) 6th European conference on Genetic programming (EuroGP), Essex, UK; 2003 Apr 14–16. p. 384–93.
12. Smith MG, Bull L. Using genetic programming for feature creation with a genetic algorithm feature selector. In 8th International Conference on Parallel Problem Solving from Nature (PPSN), Birmingham, UK; 2004 Sep 18–22. p. 1163–71.
13. Muni DP, Nikhil R, Pal, Das J. A novel approach to design classifiers using genetic programming. Institute of Electrical and Electronics Engineers (IEEE) Transactions on Evolutionary Computation. 2004 Apr 19; 8(2):183-96. Crossref.
14. Aslam MW, Nandi AK. Detection of diabetes using genetic programming. 18th European Signal Processing Conference (EUSIPCO), Denmark; 2010 Aug 23–27. p. 1184–8.
15. Pradhan MA, Rahman A, Acharya P, Gawade R, Pateria A. Design of classifier for detection of diabetes using genetic programming. International Conference on Computer Science and Information Technology (ICCSIT); 2011 Dec. p. 125–30.
16. Kumar S, Sahoo G. Classification of heart disease using naive bayes and genetic algorithm. Computational Intelligence in Data Mining. 2014 Dec 11; 2:269–82.
17. Hall M, Frank E, Holmes G, Pfagringer B, Reutemann P, Witten LH. The WEKA data mining software: an update. Association for Computing Machinery (ACM) SIGKDD Explorations Newsletter. 2009 Jun; 11(1):10–8.