I.J. Mathematical Sciences and Computing, **2020, 1, 25-32**
Published Online February 2020 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2020.01.03

Available online at http://www.mecs-press.net/ijmsc

Ferrer diagram based partitioning technique to decision tree using genetic algorithm

Pavan Sai Diwakar Nutheti*, Narayan Hasyagar*, Rajashree Shettar*, Shankru Guggari† and Umadevi V†

* Rashtreeya Vidyalaya College of Engineering, Bengaluru, India
† *B.M.S College of Engineering, Bengaluru, India*

Received: 24 June 2019; Accepted: 11 August 2019; Published: 08 February 2020

Abstract
Decision tree is a known classification technique in machine learning. It is easy to understand and interpret and widely used in known real world applications. Decision tree (DT) faces several challenges such as class imbalance, overfitting and curse of dimensionality. Current study addresses curse of dimensionality problem using partitioning technique. It uses partitioning technique, where features are divided into multiple sets and assigned into each block based on mutual exclusive property. It uses Genetic algorithm to select the features and assign the features into each block based on the ferrer diagram to build multiple CART decision tree. Majority voting technique used to combine the predicted class from the each classifier and produce the major class as output. The novelty of the method is evaluated with 4 datasets from UCI repository and shows approximately 9%, 3% and 5% improvement as compared with CART, Bagging and Adaboost techniques.

Index Terms: Data mining, Decision tree, Ferrer diagram, Vertical Partitioning.

1. Introduction

Data Mining is the process of finding new patterns in large data sets. DT algorithms have attracted and are given more significant interest both in machine learning and data mining. DT are also called as hierarchical classifiers and tool for classification, it's structure is very simple and easy to interpret. The main aim of the DT classifier is to build a model that predicts the target variable based on various input [1]. DT method is a widely applied for supervised classification technique for data analytics [2]. It can handle both continuous and categorical values. Various applications include: location finding by observing certain set of human

* Corresponding author.
E-mail address:

activities [3], prediction of student success rate [4]. etc. As we know attribute selection is a basic task for decision tree generation; DT produces sequences of rules that can be used to recognize the classes for decision making. It is well-known fact that the partitioning based methods in pattern recognition are better as compared to traditional methods in terms of computational efficiency and utilizing local information [5, 6, 7]. Genetic algorithm is a very popular evolutionary algorithm which is widely used in prediction of students academic performance [5], monitoring of aircraft air condition [6] etc.

Partitioning framework helps in managing memory and accelerate learning process [7]. Selecting the right set of attributes for classification is one of the most important problem in designing a good classifier. In the emerging area of data mining applications, users of data mining tools are facing problems with the datasets containing of large number of attributes. Mining process can be made easier to perform by focussing on a subset of relevant attributes while ignoring all others. The proposed study explore the potential of decision trees using partitioning approach, where dataset is partition into set of sub problems based on ferrer digram to reduce the adverse effects of high dimensionality.

The paper is organized as follows: Section 2 describes literature review Section 3 formally defines proposed technique Section 4 demonstrates experimental analysis and Concludes with future directions in Section 5.

2. Literature Review

In this section, various existing techniques and their corresponding challenges in the domain of selecting optimal features is discussed. Data mining techniques are used to analyse huge amount of data and provides useful information for making constructive decisions. Decision tree is non-parametric type of the classifier used to classify the given instance. The decision tree is one of the very popular data mining algorithm widely used for both classification and regression. Each internal node corresponds to testing variable and is split into child nodes based on some splitting criteria. Classes are determined at leaf node after DT is constructed using samples and decides class value based on the majority class at the leaf node [8]. The nodes of the decision tree are chosen based on entropy, Information gain, and Information Gain Ratio as shown in the equations [1,3,4].

Dimensionality reduction is a popular problem in decision tree. Lior Rokach [1] discusses feature selection using genetic algorithm and evaluates the efficiency and performance using Vapenik-Chirvonenkes dimension bound. The novelty of the method evaluated based on different domain dataset from UCI repository. Behzad Rabiee-Ghahfarrokhi use genetic algorithm with combination of C4.5 decision tree to predict microRNA molecules . The superiority of the method is evaluated with TarBase database (version 3.0) with 10-fold cross validation technique and achieves 93.9% classification accuracy [9]. Privacy-Preserving for ID3 decision tree proposed using Vertical partitioning technique [10] similarly Hari seetha et.al discussed a vertical partitioning approach using SVM classifier where features of the dataset are divided based on the mutual exclusive property [11]. Similarly a non-sequential vertical partitioning method proposed to improve both stability and classification rate of the decision tree [12]. More recently partitioning based decision

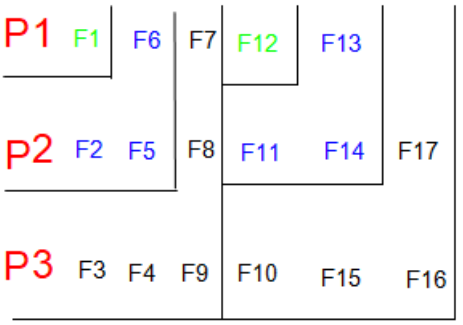

Fig. 1: Creation of blocks (partitions)

tree technique is discussed using bell triangle and ferrer diagram for low dimensional datasets and shown improvement in both classification accuracy and stability in decision tree [12].

Genetic algorithm is an evolutionary technique which is widely used in selection of best features from the dataset. Forecasting aircraft air condition system is mentioned in [6] and selected best features based on genetic algorithm. Decision tree is used to forecast the condition of the air condition system. GAWES algorithm is described to improve energy efficiency and accuracy to solve NP-complete problem [13]. Similarly multi-stage genetic algorithm is proposed by limiting individuals in the same group to avoid the structure of the evolution and showed superiority over traditional genetic algorithm in building decision tree [14]. Similarly student academic performance based on decision tree and fuzzy genetic algorithm [5]. Optimal Machine learning Model is developed to predict Software Defect. It uses feature selection to select minimum number of features and computed classification accuracy, mean square error. Support vector machine shows high classification accuracy and low mean square error [15].

2.1. Partitioning Techniques

Partitioning is a method which divides the problem into set of sub problems, where sub-problem is relatively small in dimension and are more appropriate for user driven visualization techniques and helps to improve the performance of the traditional methods. There are lot of problems in the feature selection of sequential partitioning mining such as domain specific knowledge [16], regular expressions [17], mining specific pattern and counting of their occurrences [18] and dependencies between the various sub-problems [19]. Theme based partitioning addresses the dimensionality reduction problem of decision tree [20]. Partitioning approach enables the researchers to understand the attribute and discover relationship with the other attributes hence enhance the knowledge about the concept. The present study describes homogeneous ensemble method, where number of classifiers are generated by the same classification algorithm but with different training datasets.

Advantages of the Partitioning methods

- Partitioning Methods improve the accuracy level of the classical DT classifiers [21].
- It helps to explore specialized capabilities of the attributes. For instance, accuracy level of clinical diagnosis can be improved using neural network classifier [22].
- In real time scenario; With the change of dimensionality we can easily rebuild the model and also able to maintain sub modules over period of the time as suggested in [23].
- Partitioning methodology suggests the ability to use different classifiers for each sub-problems [19].

2.2 Factors Involved in achieving high classification accuracy

- **Feature Selection:** A evolutionary algorithm called Genetic algorithm (GA) is used for feature selection with 500 generations. It improves both classification accuracy and time complexity [6,9]. Number of feature selection from GA is indicated in Table.2
- **Ferrer diagram:** It is well known optimization technique widely used in machine learning, pattern recognition and data mining technique [12,24].
- **Majority voting:** Present study uses it for combining class prediction to obtain the majority class as the predicted class (Predicted class value).

3. Ferrer Diagram based Partitioning technique to Decision tree using genetic algorithm (FDPGA)

In this paper we describe a simple vertical partitioning method, which partitions the complex problem into several sub-problems, that is, selecting attributes using Ferrer diagram as shown in [12].

Consider Numerical dataset(D) with M dimensions and I instances, which are denoted by (D) $_{I*M}$ = $[D_1, D_2, \ldots, D^n]^T$. Assume that the training data belongs to class C with labels, $c_1, c_2, \ldots, c_s$.

1) Select the feature based on genetic algorithm as indicated in section II-B.
2) Create blocks and assign features into each of them according to ferrer diagram as shown in Fig.1. P1, P2 and P3 are the blocks. Consider a dataset with 17 features. Features with similar color are put in the same block as shown in Fig.1 .
3) Build decision tree for each partition namely $DT_1, DT_2, \ldots, DT^n$.

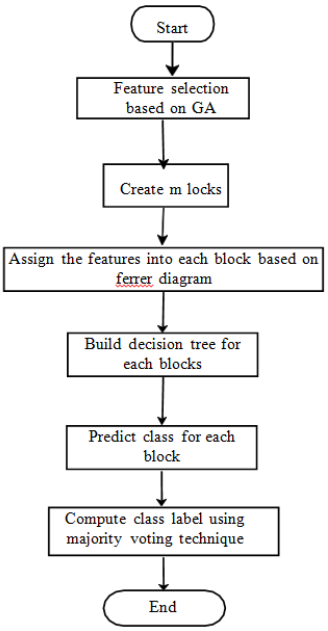

Fig. 2: Flow chart of the proposed method

4) Predict class label of a Testing object, T_i: (i) Divide T_i into sub-objects, $\{T_i^1, T_i^2, ..., T_i^m\}$, as given in Step-1. (ii) For each of the Testing sub-objects, T_i^j, $j = 1,2,\ldots,m$, predict c class memberships, p, using the Decision Tree, DT^j, as given by

$$(p^j)_{c\times 1} = predict(T^j, DT^j) \quad (1)$$

(iii) Compute final class label of Test object , T_i, by combining the class memberships, p^j, of Testing sub-objects which appeared maximum number of times.

5) Repeat step 2 and step 3 to build Decision tree.

Initially, features are selected from the genetic algorithm then create blocks as indicated in Fig.1. Models are build using these blocks using decision tree algorithm (CART). Combine these models using majority voting technique as shown in Fig.2. Finally, performance of the method is evaluated using test instances using classification accuracy .

4. Experimental Analysis

The performance of the proposed method is evaluated with 4 datasets from UCI repository [24] and used 10 fold cross validation(CV) technique. The dataset is partitioned vertically from 2 to 5 partitions. Table 1 shows the characteristics of datasets such as number of instances, number of features and number of classes and Table 2 describes number of feature selected from GA. It also indicates 45% to 50% reduction in the number of features as considering the original features of the dataset as shown in Table 1.

We used a computer system (Windows 7 OS, i5 core processor and 8GB RAM) and python (Version 3.7.3) [25] to obtain our experimental results.

Superiority of the proposed method based on the creation of the partitions. It shows nearly 11% improvement in classification rate for both LSVT Voice rehabilitation and colon tumor datasets. Proposed method improve the classification rate to 8% and 4% for both movement library and CNAE datasets respectively as shown in Fig 3. It shows 2.55% and 9.69% lowest improvement in classification accuracy for LSVT voice rehabilitation dataset as compared to both bagging and adaboost techniques respectively and 7.63% and 22.15% highest improvement in classification accuracy for colon tumor dataset as compared to both stacking and bagging techniques respectively. GA+CART means it uses features of GA to build CART model and shows improvement in classification accuracy as compared to CART indicated in Fig. 3

Wilcox test and t-test statistical tests are performed to evaluate the statistical significance of the proposed method. It shows statistical significant with the p-values of 0.018 and 0.028 as compared to both CART and stacking techniques for $\alpha = 0.05$ and other methods are not statistically significant.

TABLE 1: CHARACTERISTICS OF THE DATASETS

Sl.No	Datasets	No. of features	No. of Instances	No. of Classes
1	Movement Library	91	360	15
2	CNAE	856	1080	9
3	LSVT Voice rehabilitation	310	126	2
4	Colon Tumor	2000	62	2

TABLE 2: FEATURE SELECTION FROM GENETIC ALGORITHM (GA)

Sl.No	Datasets	No. of features	Selection from GA
1	Movement Library	91	48
2	CNAE	856	446
3	LSVT Voice rehabilitation	310	145
4	Colon Tumor	2000	986

5. Conclusions

In this paper we proposed a vertical partitioning approach to decision tree based on Ferrer diagram. We used genetic algorithm for feature section which help to improve the performance of the traditional decision tree methods. Superiority of the proposed method is evaluated based on the classification accuracy as compared to traditional decision tree and other popular ensemble techniques. Current study uses till 5 partitions to build ensemble of decision tree. Since in this study we considered only low dimensional datasets to understand the significance of the proposed method. In future, we would like to use high dimensional datasets and class imbalanced datasets to evaluate the performance of the proposed technique.

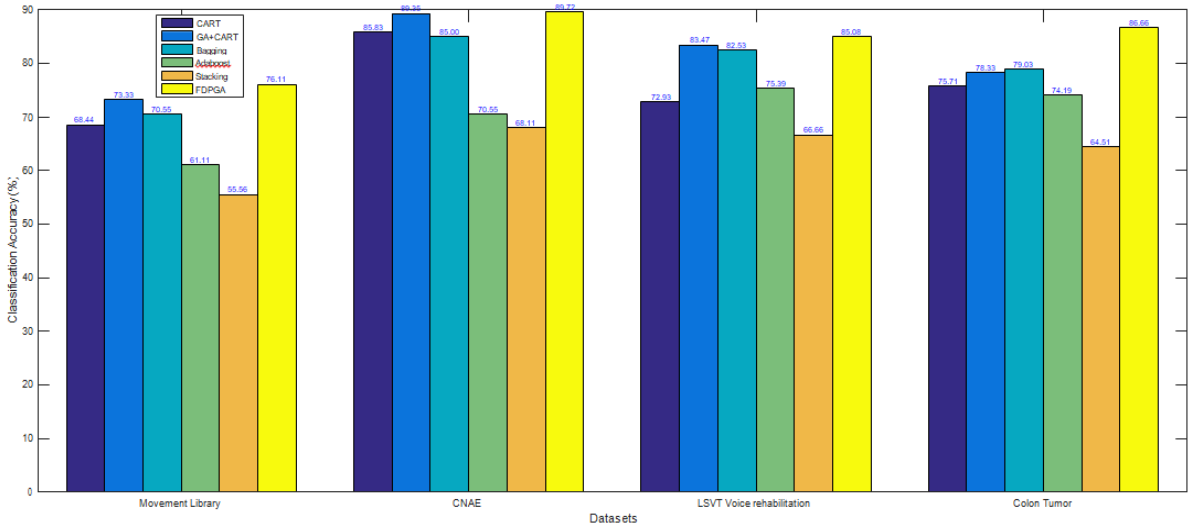

Fig. 3: Classification accuracy of the proposed method with traditional methods

References

[1] Diogo R, Ferreira A, Evgeniy Vasilyev, "Using logical decision trees to discover the cause of process delays from event logs," Computers in Industry, vol. 70, pp. 194–207, 2015.

[2] Gregory-Piatetsky-Shaoiro, "KDnuggets Website," in http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html, 2013.

[3] Jae Sung Lee and Eun Sung Lee, "Exploring the Usefulness of a Decision Tree in Predicting Peoples Location," in Procedia - Social and Behavioural Sciences, 2014, pp. 447–451.

[4] Srecko Natek and Moti Zwilling, "Student data mining solution knowledge management system related to higher education institutions," Computers in Industry, vol. 41, pp. 6400–6407, 2014.

[5] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student academic performance prediction model using decision tree and fuzzy genetic algorithm," Procedia Technology, vol. 25, pp. 326 – 332, 2016.

[6] M. Gerdes, "Decision trees and genetic algorithms for condition monitoring forecasting of aircraft air conditioning," Expert Systems with Applications, vol. 40, no. 12, pp. 5021 – 5026, 2013.

[7] I. B. Yashkov, "Feature selection using decision trees in the problem of jsm classification," Automatic Documentation and Mathematical Linguistics, vol. 48, pp. 6–11, 2014.

[8] Kyoungok Kim, "A hybrid classification algorithm by sub space partitioning through semi-supervised decision tree," Pattern Recognition, vol. 60, pp. 157–163, 2016.

[9] B. Rabiee-Ghahfarrokhi, F. Rafiei, A. A. Niknafs, and B. Zamani, "Prediction of microrna target genes using an efficient genetic algorithm-based decision tree,", FEBS Open Bio, vol. 5, pp. 877 – 884, 2015.

[10] Jaideep Vaidya et. al, "Privacy-preserving decision trees over vertically partitioned data," ACM Transactions on Knowledge Discovery from Data, vol. 2, 2008.

[11] Hari Seetha,M. Narasimha Murty, R. Saravanan, "Classification by majority voting in feature partitions," International Journal of Information and Decision Sciences, vol. 8, no. 2, pp. 109–124, 2016.

[12] S. Guggari, V. Kadappa, and V. Umadevi, "Non-sequential partitioning approaches to decision tree classifier," Future Computing and Informatics Journal, vol. 3, no. 2, pp. 275 – 285, 2018.

[13] H. Akcan, “A genetic algorithm based solution to the minimum-cost bounded-error calibration tree problem,” Applied Soft Computing, vol. 73, pp. 83 – 95, 2018.
[14] L. Yi and K. Wanli, “A new genetic programming algorithm for building decision tree,” Procedia Engineering, vol. 15, pp. 3658 – 3662, 2011.
[15] T. Lamba, Kavita, and A.K.Mishra, “Optimal machine learning model for software defect prediction,” International Journal of Intelligent Systems and Applications(IJISA), vol. 11, no. 02, pp. 36 – 48, 2019.
[16] D.R. Ferreira and M. Zacarias and M. Malheiros and P. Ferreira , “Approaching process mining with sequence clustering: experiments and findings,” in Proceedings of the 5th International Conference on Business Process Management (BPM 2007), vol. 4714, 2007.
[17] M.N. Garofalakis and R. Rastogi and K. Shim , “Spirit: Sequential pattern mining with regular expression constraints,” in Proceedings of the 25th International Conference on Very Large Data Bases (VLDB), 1999, pp. 223–234.
[18] C. Wang and A. Kao and J. Choi and R. Tjoelker, “Discovering Time-Constrained Patterns from Long Sequences,” Advances of computational intelligence in industrial systems, vol. , no. , pp. 99–116, 2008.
[19] Lior Rokach, “Decomposition Methodology for Classification Tasks - A Meta Decomposer Framework,” Pattern Analysis and Applications, vol. 9, pp. 257–271, 2006.
[20] Vijayakumar Kadappa and Shankru Guggari and Atul Negi, “Decision Tree Classifier using Theme based Partitioning,” in IEEE International Conference on computing and network Communications (CoCoNet’15), 2015, , pp. 546–552.
[21] Lior Rokach and Oded Maimon, “Data mining for improving the quality of manufacturing: a feature set decomposition approach,” J Intell Manuf, vol. 17, pp. 285–299, 2006.
[22] Baxt W. G, “Use of an artificial neural network for data analysis in clinical decision making: The diagnosis of acute coronary occlusion,” Neural Computation, vol. 2(9), pp. 480–489, 1990.
[23] Kusiak, A, “Decomposition in Data Mining: An Industrial Case Study,” in IEEE Transactions on Electronics Packaging Manufacturing, 2000, pp. 345–353.
[24] Aha D and Murphy P, “UCI Repository of machine learning databases,” in http://www.ics.uci.edu/mlearn/MLRepository.html . Irvine, CA: University of California, Department of Information and Computer Science, 1994.
[25] R, “The R project for statistical computing,” in http://www.r-project.org/.

Authors’ Profiles

Mr. Pavan Sai Diwakar Nutheti graduated from R. V College of Engineering, Bengaluru in July 2019 with a bachelor of engineering degree in Computer Science and Engineering. He is currently working as a Software Development Engineer.

Mr. Narayan Ganesh Hasyagar graduated from R. V College of Engineering, Bengaluru in July 2019 with a bachelor of engineering degree in Computer Science and Engineering. He is currently working as a Software Development Engineer.

Dr. Rajashree Shettar is currently working as Professor, Dept of Computer Science, RV College of Engineering, Bengaluru. Her research work focuses on Knowledge Discovery in Semi-structured Data. She has around 45 publications in various International Journals and Conferences. She has authored a book on Sequential Pattern Mining from Web Log Data: Concepts, Techniques and Applications of Web Usage Mining, published by LAMBERT Academic Publishingcompany, Germany and co-authored book chapters.She has published Indian Patent titled Developing a Therapeutic Biomarker for Ebola Viral Disease along with Dr. Vidya Niranjan, Sanchit Mittal and Nishka Ranjan

Mr. Shankru Guggari (Ph.D), M.Tech., Research scholar, in the Dept. of Computer science and Engineering, B.M.S. College of Engineering, Bangalore. He is currently working in classification technique area for his Ph.D dissertation. Recently, he has won the best research paper award in the international conference. Pattern recognition, IOT and Machine learning are the interested research area of him. He has published some of his research works in international conferences and a research paper in Elsevier publication journal. He has more than 4 years of industry experience and more than 3 years in academic research experience.

Dr. Umadevi V obtained her Ph.D from IIT Madras and currently working as Associate Professor and Head for Computer science and engineering Department at B.M.S. college of Engineering, Bengaluru. She has published her work in many reputed international conferences and also published many articles in leading journals with well known publishers (Elsevier etc.,). She served as resource person for many Workshops and Faculty development programs. Recently she got international grants from Amoudi Scientific Research Foundation of Majmaah University, Kingdom of Saudi Arabia to conduct research in the area of Medical Thermography.

How to cite this paper: Pavan Sai Diwakar Nutheti, Narayan Hasyagar, Rajashree Shettar, Shankru Guggari, Umadevi V,"Ferrer diagram based partitioning technique to decision tree using genetic algorithm", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.6, No.1, pp.25-32, 2020. DOI: 10.5815/ijmsc.2020.01.03