A new decision tree algorithm IQ Tree for class classification problem in Data Mining

C.V.P.R.Prasad
Research Scholar,
Acharya Nagarjuna University,
Guntur, Andhra Pradesh, India.

Dr. Bhanu Prakash Battula
Associate Professor,
Vignan College,
Andhra Pradesh, India.

ABSTRACT :

Data mining and knowledge discovery is used for discovery of hidden knowledge from large data sources. Decision trees are one of the most famous classification techniques with simple and efficient generalization technique. This paper presents a new decision tree algorithm IQ Tree for class classification problem. The IQ Tree assumes using an inter quartile range conversion of attributes with C4.5 as the base algorithm for performing induction can improve all the measures such as accuracy, tree size.

Keywords:

DataMining, Classification, Decision Tree,inter quartile range.

1. INTRODUCTION:

In Machine Learning community, and in data mining works, classification has its own importance. Classification is an important part and the research application field in the data mining [1].A decision tree gets its name because it is shaped like a tree and can be used to make decisions. —Technically, a tree is a set of nodes and branches and each branch descends from a node to another node. The nodes represent the attributes considered in the decision process and the branches represent the different attribute values.

To reach a decision using the tree for a given case, we take the attribute values of the case and traverse the tree from the root node down to the leaf node that contains the decision.” [2]. A critical issue in artificial intelligence (AI) research is to overcome the so-called —knowledge-acquisition bottleneck” in the construction of knowledge-based systems.

Decision tree can be used to solve this problem. Decision trees can acquire knowledge from concrete examples rather than from experts [3]. In addition, for knowledge-based systems, decision trees have the advantage of being comprehensible by human experts and of being directly convertible into production rules [4].

2. LITERATURE REVIEW :

In Data mining, the problem of decision trees has also become an active area of research. In the literature survey of decision trees we may have many proposals on algorithmic, data-level and hybrid approaches. The recent advances in decision tree learning have been summarized as follows: A parallel decision tree learning algorithm expressed in MapReduce programming model that runs on Apache Hadoop platform is proposed by [5].A new adaptive network intrusion detection learning algorithm using naive Bayesian classifier is proposed by [6]. A new hybrid classification model which isestablished based on a combination of clustering, feature selection, decision trees, and genetic algorithmtechniques is proposed by [7]. A novel roughest based multivariate decision trees (RSMDT) method in which, the positive region degree of condition attributes with respect to decision attributes in rough set theory is used for selecting attributes in multivariate tests is proposed by [8]. A novel splitting criteria which chooses the split with maximum similarity and the decision tree is calledmstreeis proposed by [9].

An improvedID3 algorithm and a novel class attribute selection method based on Maclaurin-Priority Value First method is proposed by [10]. A modified decision tree algorithm for mobile user classification, which introducedgenetic algorithm to optimize the results of the decision tree algorithm, is proposed by [11].

Volume No: 1 (2015), Issue No: 1 (June)
www. IJRACSE.com

June 2015
Page 5

Volume No:1, Issue No:1 (June-2015)

International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

A new parallelized decision tree algorithm on a CUDA (compute unified device architecture), which is a GPG-PU solution provided by NVIDIA is proposed by [12]. A Stochastic Gradient Boosted Decision Trees based method is proposed by [13].

A modified Fuzzy Decision Tree for the fuzzy rules extraction is proposed by [14].Obviously, there are many other algorithms which are not included in this literature. A profound comparison of the above algorithms and many others can be gathered from the references list.

3. THE PROPOSED APPROACH :

In this Section, we investigate to propose a new decision tree induction algorithm known as Inter Quartile (IQ) Range Decision Tree. Our IQ Decision Tree induction method depends on inter quartile ranges which was descried in the above section. We assume that the subset of the training data is small, i.e. it is computationally cheap to act on such a set in a reasonable time. We focus on a set of improved attribute range filters using attribute transformations.

Next, we try to adapt and deploy them as IQ Tree components. Since the IQ Tree scheme is based on a restricted list of candidates, this list could be represented by features that seems to be relevant or those that might provide incremental usefulness to the selected feature subset.

For the IQ Tree construction stage we opt for selection scheme capable of generating attribute ranking. Hence, the weights associated to features will serve as one of the selection criterion in the new heuristic function for inducing decision trees. The next stage of IQ Tree tries to consider both entropy and weights for splitting of attributes.

The quality of solution fine-tuning, mainly, depends on the nature of the filter involved and the parameters of attribute transformation. The following algorithm, detail different design alternatives for both attributes transform and filter procedure search for IQ Tree components.

The algorithm for IQ decision tree is shown below

Algorithm 4: New Decision Tree (D, A, RGR)

Input: D   – Data Partition
      A   – Attribute List
    GR   – Gain Ratio

Output: Decision Tree Measures – Accuracy, Tree Size.

Procedure:

1.   Create a node N
2.   **If** samples in N are of same class, C **then**
3.   return N as a leaf node and mark class C;
4.   **If** A is empty **then**
5.   **return** N as a leaf node and mark with majority class;
6.   **else**
7. (D_w, A_w) = apply Inter Quartile Range (D, A)
8.   apply Gain Ratio(D_w, A_w)
9.   label root node N as $f(A)$
10. **for** each outcome j of $f(A)$**do**
11.   subtree j =New Decision Tree($D_w j, A_w$)
12.   connect the root node N to subtree j
13. **endfor**
14. **endif**
15. **endif**
16. Return N

The algorithm: IQ Tree can be explained as follows:
The inputs to the algorithm are data partition "D", attribute set "A" and splitting criteria gain ratio "GR". The output of the algorithm will be the average measures such as accuracy and tree size produced by the IQ Tree method. The algorithm begins with the create node for same class. In the next stage, attribute rages are applied to the inter quartile for transformation. In the later on stage, the transformed dataset is applied for the splitting criteria gain ratio for decision tree induction. The induced decision tree is applied for the tree pruning process for generalization of the tree. In the final the measures for decision tree validation i.e accuracy ad tree size are generated.

4. RESULTS AND DISCUSSION:

We experimented with 10 standard datasets from the UCI repository, these datasets are standard benchmark imbalanced datasets used in the context of supervised learning. The goal is to examine whether the proposed IQ Tree achieve better predictive performance than a number of existing standard learning algorithms.

Volume No: 1 (2015), Issue No: 1 (June)
www. IJRACSE.com

June 2015
Page 6

We compared the above method with the C4.5, BPN, REP, CART and NB Tree state-of-the-art metric learning algorithms. In all the experiments we estimate accuracy using 10-fold cross-validation and control for the statistical significance of observed differences using t-test (sig. level of 0.05). In Table 3 and 4, we present the results of the comparison between C4.5, BPN, REP, CART, NB Tree and IQ Tree.

From these results we can make several observations. The developed IQ Tree compared with C4.5, REP, CART, NB Tree generally given competitive results; the advantage of our methods is most visible in the balance, diabetes, glass, ionosphereand sonar datasets. Finally, the method that most often win is IQ Tree.

Table 3 Summary of tenfold cross validation performance for Accuracy on all the datasets

S.No Datasets	C4.5	REP	CART	NB Tree	IQ Tree
1. Balance-scale	77.82●	78.54●	78.73●	75.96●	97.24
2. Breast-cancer	74.28● 69.35●	70.22● 70.99●	98.7		
3. Pima_diabetes	74.49●	74.46●	74.56●	74.96● 95.25	
4. Glass	67.63●	65.54●	71.26●	69.84●	76.48
5. Heart-statlog	78.15●	76.15●	78.07●	80.93● 91.51	
6. Ionosphere	89.74●	89.46●	88.87●	90.03●	93.16
7. Iris	94.73●	93.87●	94.20●	93.47●	95.19
8. Sonar	73.61●	72.69●	70.72●	77.11●	82.82
9. Vehicle	72.28●	70.18●	69.91●	70.98●	86.64
10. Waveform	75.25●	76.57●	76.65●	79.84●	89.66
Win/Tie/Loss	(10/0/0)	(10/0/0)	(10/0/0)	(10/0/0)	

● Bold dot indicates the win of proposed method;   ○ Empty dot indicates the loss of proposed method.

Table 4 Summary of tenfold cross validation performance for Tree Size on all the datasets

S.No Datasets	C4.5	REP	CART	NB Tree	IQ Tree
1. Balance-scale	82.20●	42.36○	55.28●	17.38○	30.64
2. Breast-cancer-w	23.46●	13.76●	15.90●	5.68○	9.58
3. Pima_diabetes	43.40○	30.98○	17.36○	5.18○ 27.58	
4. Glass	46.16●	19.70○	21.16○	10.0○	39.98
5. Heart-statlog	34.64●	14.78●	15.36●	9.62○ 19.38	
6. Ionosphere	26.74●	8.76○	8.42○	16.20○	18.42
7. Iris	8.28●	5.84○	7.40○	4.38○	10.90
8. Sonar	27.90●	10.20○	10.50○	13.74○	26.46
9. Vehicle	138.0●	58.52○	92.54○	57.70○	84.62
10. Waveform	591.94●	167.24○	98.32○	94.48○	290.44
Win/Tie/Loss	(9/0/1)	(2/0/8)	(3/0/7)	(0/0/10)	

● Bold dot indicates the win of proposed method;   ○ Empty dot indicates the loss of proposed method.

Table 3and4 presents the comparative results of proposed algorithm IQ Tree against C4.5, REP, CART and NB Tree. The value in the table; example: "11/0/2" specifies that the proposed algorithm has registered 11 wins, 0 ties and 2 losses against compared algorithm for that specified measure. One can observe from the table 3 and 4 that our proposed algorithms have registered good number of wins against the compared algorithms on all the datasets. These results suggest that in the majority of the high dimensional datasets, the feature interactions are not important, and hence the methods that do not account for feature interactions

have in general better performances. Alternatively, it might suggest that stronger regularization is needed. Moreover, it is interesting to note that the cases for which the good performance are difficult classification problems from the UCI datasets. This hints that there might be a bias of method development towards methods that perform well on UCI datasets; however, one can argue that they are really representative of the real world.These results are remarkable since IQ Tree, which is based on a simple idea, performs equally well as the more elaborate standard learning algorithm that has been reported to consistently outperform other

Volume No: 1 (2015), Issue No: 1 (June)
www. IJRACSE.com

June 2015
Page 7

metric learning techniques over a number of non-trivial learning problems. Finally, we mention that the surprisingly poor performance of IQ Tree on sonar, vehicle and vowel datasets in tables 4, might be explained by the fact that its conversion function is not convex and hence it is sensitive to the unique properties of the datasets.

In overall, from all the tables and figures we can conclude that our proposed IQ Tree have given good results when compared to benchmark algorithms. The unique properties of datasets such as size of the dataset and the number of attributes will also effect on the results of our proposed IQ Tree. The above given results are enough to project the validity of our approach and more deep analysis should be done for further analysis.

5. CONCLUSION:

This paper presents a new decision tree algorithm IQ Tree for class classification problem. The IQ Tree assumes using an inter quartile range conversion of attributes with C4.5 as the base algorithm for performing induction can improve all the measures such as accuracy, tree size. The experiments conducted with IQ Tree specify that improved performance can be achieved. We have conducted experiments on 10 datasets from UCI which suggest that IQ Tree can quickly remove redundant, irrelevant and weak attributes as long as the properties of the dataset are normal.

Excellent improvement in measures on some natural domain datasets shows the compatibility of IQ Tree approach on real-time applications. One of the shortcomings seen in IQ Tree is when used for datasets with unique properties; Because IQ Tree will not consider unique properties of datasets for removing instances from data source. Finally, we can conclude that IQ Tree can be a good contribution as a decision tree induction method for efficient learning of the varied datasets.

References:

1.Juanli Hu, Jiabin Deng, Mingxiang Sui, A New Approach for Decision Tree Based on Principal Component Analysis, Proceedings of Conference on Computational Intelligence and Software Engineering, page no:1-4, 2009.

2.Shane Bergsma, Large-Scale Semi-Supervised Learning for Natural Language Processing, PhD Thesis, University of Alberta, 2010.

3.J. Durkin. Expert Systems: Design and Development, Prentice Hall, Englewood Clis, NJ, 1994.

4.J. Quinlan. C4.5 Programs for Machine Learning, San Mateo, CA:Morgan Kaufmann, 1993.

5.VasilePurdila, Ştefan-Gheorghe Pentiuc" MR-Tree - A Scalable MapReduce Algorithm for Building Decision Trees", Journal of Applied Computer Science & Mathematics, no. 16 (8) /2014, Suceava.

6.Dewan Md. Farid, NouriaHarbi, and Mohammad Zahidur Rahman" Combining naive bayes and decision tree for adaptive intrusion detect", International Journal of Network Security & Its Applications (IJNSA), Volume 2, Number 2, April 2010.

7.Mohammad Khanbabaei and Mahmood Alborzi" THE USE OF GENETIC ALGORITHM, CLUSTERING AND FEATURE SELECTION TECHNIQUES IN CONSTRUCTION OF DECISION TREE MODELS FOR CREDIT SCORING", International Journal of Managing Information Technology (IJMIT) Vol.5, No.4, November 2013. DOI : 10.5121/ijmit.2013.5402

8.Dianhong Wang, Xingwen Liu, Liangxiao Jiang, Xiaoting Zhang, Yongguang Zhao" Rough Set Approach to Multivariate Decision Trees Inducing?", JOURNAL OF COMPUTERS, VOL. 7, NO. 4, APRIL 2012.

9.Xinmeng Zhang, Shengyi Jiang "A Splitting Criteria Based on Similarity in Decision Tree Learning", JOURNAL OF SOFTWARE, VOL. 7, NO. 8, AUGUST 2012.

10.Ying Wang, Xinguang Peng, and Jing Bian" Computer Crime Forensics Based on Improved Decision Tree Algorithm", JOURNAL OF NETWORKS, VOL. 9, NO. 4, APRIL 2014.

11.Dong-sheng Liu, Shujiang Fan" A Modified Decision Tree Algorithm Based on Genetic Algorithm for Mobile User Classification Problem", Scientific World Journal, Volume 2014, Article ID 468324, 11 pages, http://dx.doi.org/10.1155/2014/468324, Hindawi Publishing Corporation.

Volume No: 1 (2015), Issue No: 1 (June)
www. IJRACSE.com

June 2015
Page 8

12.Win-Tsung Lo, Yue-Shan Chang, Ruey-Kai Sheu, Chun-Chieh Chiu and Shyan-Ming Yuan,'' CUDT: A CUDA Based Decision Tree Algorithm'', e Scientific World Journal, Volume 2014, Article ID 745640, 12 pages, http://dx.doi.org/10.1155/2014/745640. Hindawi Publishing Corporation.

13.Tarun Chopra, JayashriVajpai'' Fault Diagnosis in Benchmark Process Control System Using Stochastic Gradient Boosted Decision Trees'', International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-1, Issue-3, July 2011.

14.S.V.S. Ganga Devi'' FUZZY RULE EXTRACTION FOR FRUIT DATA CLASSIFICATION'', COMPUSOFT, An international journal of advanced computer technology, 2 (12), December-2013 (Volume-II, Issue-XII).

15.HamiltonA. Asuncion D. Newman. (2007). UCI Repository of Machine Learning Database (School of Information and Computer Science), Irvine, CA: Univ. of California [Online]. Available: http://www.ics.uci.edu/mlearn/MLRepository.html

16.Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.

17.J. Quinlan. Induction of decision trees, Machine Learning, vol. 1, pp. 81C106, 1986.

18.L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.

19.Nitesh V. Chawla et. al. (2002). Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16:321-357.

Volume No: 1 (2015), Issue No: 1 (June)
www. IJRACSE.com

June 2015
Page 9