ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

An Improved Tree Using Discrete Haar Wavelet Transform

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 is a popular knowledge discovery technique. In data mining decision trees are of the simple and powerful decision making models. One of the limitations in decision trees is towards the data source which they tackle. If data sources which are given as input to decision tree are of imbalance nature then the efficiency of decision tree drops drastically, we propose a decision tree structure which uses discrete haar wavelet transformation technique along with a filter. In this paper, we propose a novel method WT Tree based on above strategy. Extensive experiments, using C4.5 decision tree as base classifier, show that the performance measures of our method is comparable to state-of-the-art methods.

Keywords:

Knowledge Discovery, Data Mining, Classification, Decision Trees, Discrete haar.

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]. A decision tree not only provides the solution for a given case, but also provides the reasons behind its decision. So the real benefit of decision tree technology is that it avoids the need for human expert. Because of the above advantages, there are many successes in applying decision tree learning to solve real-world problems.

2 RECENT ADVANCES IN DECISION TREES :

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 is established based on a combination of clustering, feature selection, decision trees, and genetic algorithm techniques 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 called mstree is proposed by [9]. An improved ID3 algorithm and a novel class attribute selection method based on Maclaurin-Priority Value First method is proposed by [10].

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A modified decision tree algorithm for mobile user classification, which introduced genetic algorithm to optimize the results of the decision tree algorithm, is proposed by [11]. A new parallelized decision tree algorithm on a CUDA comparison of the above algorithms and many others can be gathered from the references list.

3.THE PROPOSED METHOD :

In this section, the proposed approach is presented. In the framework of WT Tree a base algorithm is used in the implementation and the efficiency of the WT Tree will also depend on the fine tuning of the parameters and the base algorithms etc. As to find the efficiency of WT Tree for different parameters, we have designed different variations of WT Tree by varying the type of base algorithm and fine tuning parameters in implementation.

The WT Tree follows a feature conversion and instance misclassification detection approach for continuous improvement. The above said strategy is implemented in the WT system. In the initial stage the decision tree learning process will initiate with the identification of influential features or attributes from the data source.

The selected or identified prominent features from the data source are grouped together and again used for conversion by using wavelet transform. In this case, we have used discrete wavelet transform. In the discrete wavelet transform, one of the best applicable approaches for the real world data is Haar wavelet transform.

The so called process of features conversion is done. After the conversion of the influential features, the learning process is initiated for identification of misclassified instances by using a base algorithm. The converted "n" features are used to build the model for that particular data source and the efficiency of the model is evaluated. In the next, phase the improved data source is applied to the base algorithm (C4.5) and the learning process is continued for the evaluation of efficiency of the model.

The algorithm for WT Tree approach is given below,

Algorithm: Wavelet Transform Decision Tree

Input: Tr: a training sample set
La: Label for Tr
M: The number of selected features
Output Measures: Accuracy, Tree size.
Procedure:
Phase I:
Finding important features using filter
Features = Apply Filter (Data set, CFS)
Phase II:
Conversion of the features
for feature i = 1 to M//M represents # features selected
for Tr (j) = 1 to N//N represents # Transform the features
Transform (WT, j-1)
j=j-1
end for
end for
Phase III:
Training on Converted features
For Tr (i,j) = 1 to O//O represents # converted features
Learn Model = Build (Tr (i.j), C4.5)
End for
Predict (Learning Mode, Measure)

The algorithm : Wavelet Transform Decision Tree can be explained as follows.

The inputs to the algorithm are training sample set "Tr" and label of training sample set "La." The output of the algorithm will be the average measures such as accuracy and tree size produced by the WT Tree method. The algorithm begins with the initial stage of identifying important features M, where M is the number of features extracted by applying correlation based feature subset filter on the data set. The "M" value will change from one data set to other, and depending upon the unique properties of the data set, the number of features can be more or less. In the next stage, features are converted and consecutively trained and decremented on the base algorithm.

4.RESULTS :

We experimented with 10 standard datasets from the UCI repository (Breast_w, Diabetes, Hepatitis, Sonar, Ionosphere, Vote, Colic, labor, Breast and Sick); these datasets are standard benchmark imbalanced datasets used in the context of supervised learning.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The goal is to examine whether the WT Tree achieve better predictive performance than a number of existing standard learning algorithms. We compared the above methods with the C4.5, 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 2-3, we present the results of the comparison between C4.5, REP, CART, NB Tree and WT Tree. From these results we can make several observations. The developed WT Tree based on C4.5, CART and REP generally given competitive results for C4.5, REP and traditional benchmarks; the advantage of our methods is most visible in the breast_w, diabetes, vehicle, ionosphere and sonar datasets. Finally, the method that most often win is WT Tree.

Table 2 Summary of tenfold cross validation performance for Accuracy on all the datasets

Datasets	C4.5	REP	CART	NB Tree	WT 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●	98.74
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 3 Summary of tenfold cross validation performance for Tree Size on all the datasets

Datasets	C4.5	REP	CART	NB Tree	WT 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)**	**(3/0/7)**	**(3/0/7)**	**(0/0/10)**	

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

These results are remarkable since WT Tree, which are based on a simple idea, performs equally well as the more elaborate standard learning algorithm that has been reported to consistently outperform other metric learning techniques over a number of non-trivial learning problems. Finally, we mention that the surprisingly poor performance of WT Tree on iris, waveform and glass datasets in tables 3, might be explained by the fact that its learning function is not convex and hence it is sensitive to the tree size.Table 2 and 3 presents the summary of experimental results conducted using non-parametric statistical tests with Wilcoxon test for all the datasets.

The last row in tables2 and 3 describes number of wins or tie of WT Tree results of accuracy and tree size for every dataset. The values M/N/O specify the number of wins, number of ties and number of losses with compared WT Tree algorithm. Tables 2 and 3 presents the comparative results of proposed algorithm WT Tree against C4.5, REP, CART and NB Tree. The value in the table; example: "(9/0/1)" specifies that the proposed algorithm has registered 9 wins, 0 ties and 1 loss against compared algorithm on that dataset for that specified measure. One can observe from the tables 2 and 3 that our proposed algorithm has registered good number of wins against the compared algorithms

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

on all the datasets. In overall, from all the tables we can conclude that our proposed WT Tree has 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 WT 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:

In this paper, we proposed a new decision tree algorithm dubbed as WT Tree for classification problem. The WT Tree assumes using a discrete haar wavelet transform for attribute transformation with eliminating mostly misclassified instances can improve all the classification measures such as accuracy and tree size.

The experiments conducted with WT Tree specify that improved classification measures can be achieved. We have conducted experiments on 10 datasets from UCI which suggest that WT Tree can quickly remove redundant, irrelevant and weak instances and attributes as long as the properties of the dataset are normal. Excellent improvement in classification measures on some natural domain datasets shows the compatibility of WT Tree approach on real-time applications. Finally, we can conclude that WT Tree can be a good contribution as a decision tree induction method for efficient learning of the 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.Vasile Purdila, 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, Nouria Harbi, 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.

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.

3.Tarun Chopra, Jayashri Vajpai" Fault Diagnosis in Benchmark Process Control System Using Stochastic Gradient Boosted Decision Trees", International

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

1Journal 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.