International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 4 (2015) pp. 11297-11306
© Research India Publications
http://www.ripublication.com

An Efficient Approach For Knowledge Discovery In Decision Trees Using Attribute Transform and Outlier Detection

C.V.P.R.Prasad*, Dr. Bhanu Prakash Battula**
**Research Scholar, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.Email:prasadcvpr@gmail.com*
***Associate Professor, Vignan College, Andhra Pradesh, India. Email: battulaphd@gmail.com*

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 proposes a novel algorithm know as Attribute transform and Outlier Detection (ATOD) for classification of varied and noisy datasets. We consideredthe problem of classification as a two-step process, firststep of which dealt with the attribute transformation and outlier detection whereas the second step was involved withthe classification of these data sources by using C4.5 as the base algorithm. The performance of the proposed algorithm is impressive.

Keywords: DataMining, Classification, Decision Tree,Attributes Transform and Outlier Detection.

Introduction

Classification is a well know knowledge discovery technique used in data mining. In classification, decision tree are of the most widely used techniques for decision making. A decision tree is a flow-chart-like tree structure, where each internal node denotes atest on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions [1]. A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision.One of the main problems in the effective operation of decision trees is its complexity. The complexity of the decision trees should be minimized to have better generalization.

In this paper, the statistical procedure concerned with elucidating the covariance structure of a set of variables and outlier detection techniques are introduced to

provide improved performance. The rest of this paper was organized as follows: The related work is given in section 2. The proposed algorithm is discussed in section 3. In section 4, the details of experimental framework are presented. Simulation results are listed in section 5 and conclusion is presented in final section.

Literature Review

There is a rich literature related to the topic of decision trees. In this section, we review the most closely related works. Our review is by no means comprehensive. We refer [2] interested readers to for a more thorough survey.

N. Sivaramet al. [3] have investigated empirical characterization and evaluation of pruned and unpruned trees construction using ID3, C4.5 and CART decision trees algorithms for the recruitment problem domain. P. Karthigayaniet al. [4] haveproposed a Decision Tree Based Occlusion Detection (DTOD) classifier for Occlusion detection in face verification.Hiroshi Imamuraet al. [5] haveproposed a decision tree for deciding the safe limit in preoperative assessment of liver function and prediction of postoperativeliver function to minimize surgicalrisk, especially in patients with hepatocellular carcinoma.Masud Karimet al. [6] haveinvestigated on two decision tree algorithms the Naïve Bayes and the C4.5 to predict whether a client will subscribe a term deposit.

S. KluskaNawareckaet al. [7] havedescribes the methodology and the process of developing fuzzy logic-based models of decision making based on preprocessed data with classification trees, where the needs of the diverse characteristics of copper alloys processing are the scope. Ihsan A. Kareem et al. [8] havepresented a problem of finding the parameter settings of decision tree algorithmin order to build an accurate tree.The applied technique is an unsupervised filter and the suggested discretization applies on C4.5 algorithm to construct adecision tree. Mohammad Nazaripouret al. [9] have conducted a study which uses a two-step procedure for the evaluation of B2C controls, first, using a Data EnvelopmentAnalysis (DEA) model, second using decision trees. The results of the DEA model indicate that retail firms andinformation service providers implement B2C controls more effectively than financial firms do. Thedecision tree model issued to suggest the level of controls and argued rules for controls guidance. After analyzing the existing recent literature, we found that new classification algorithm for varied data source is the need of the hour.

The Proposed Approach

AttributeTransform with Outlier Detection (ATOD) algorithm is a linear phase, attribute-based optimizationalgorithm. ATOD works with the main principle of optimizing the complexity in terms of tree size and maximizing the accuracy.The optimization process is conducted by means attribute transformation and outlier detection.

In attribute transformation, astatistical approach concerned with elucidating the covariance structure of a set of variables is introduced. The main goal is to identify the principal direction in which the data transformation is acceptable. To understand

the covariance initially we need to understand variance. Variance canbe defied as follows,

Variance is another measure of the spread of data in a data set. In fact it is almost identical to the standard deviation. The formula is given in eq (1),

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} \text{ ----------} \quad (1)$$

The formula for covariance is very similar to the formula of variance. The formula for variance could also be written like given in eq (2),

$$\text{var}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)} \text{ ----------} \quad (2)$$

The formula for covariance of variables (x,y) is given in eq (3),

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)} \text{ ----------} \quad (3)$$

Let us consider, a triangle represent a two variable data set which have measuredin the X-Y coordinate system. The acceptable direction in which the data varies is shown by the U axis and the second most important direction is the V axis orthogonal to it. If we place the (U-V)axis system at the mean of the data it gives us a compact representation. If we transform each (X; Y) coordinate into its corresponding (U; V) value, the data is de-correlated, meaning that the co-variance between the U and V variables is zero. For a given set of data, principal component analysis finds the axis system defined by the principal directions of variance (ie the U-V axis system). The directions U and V are known as attribute transformable directions.In our implementation of the transformation, the covariance is set as 0.95. In the next stage of our frame work, the outlier are detected by using some of the specific properties such as the density, deviation from the average values, most misclassification of instances etc.. In the final stage, we use a base algorithm i'e C4.5 to build and evaluate model for classification results. The algorithm for ATOD can be given as follows.

Attribute Transform Outlier Detection (ATOD)

Algorithm: New Decision Tree (D, A)

Input: D – Data Partition
A – Attribute List

Output: A Decision Tree

Procedure:
1. **Attribute Transformation (D, A)**

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

Experimental Setup and Algorithms Compared

In the study, we have considered 24 data-sets which have been collected from the UCI [11] machine learning repository web sites. The complete details regarding all the datasets can be obtained from UCI Machine Learning Repository.

We have obtained the accuracy ad tree size metric estimates by means of a 10-fold cross-validation. That is, the data-set was split into ten folds, each one containing 10% of the patterns of the dataset. For each fold, the algorithm is trained with the examples contained in the remaining folds and then tested with the current fold.Table 1 summarizes the properties of the selected datasets.

Table 1 The 24 UCI datasets and their properties

S.no.	Dataset	Instances	Missing	Numeric Values	Nominal attributes	Classes attributes
1.	Anneal.ORIG	898	Yes	5	28	6
2.	Balance-scale	625	No	4	0	3
3.	Breast-cancer	286	Yes	0	9	2
4.	Breast-w	699	Yes	9	0	2
5.	Horse-colic	368	Yes	7	15	2
6.	Credit-a	690	Yes	6	9	2
7.	Credit-g	1,000	No	7	13	2
8.	Pima diabetes	768	No	8	0	2
9.	Glass	214	No	9	0	6
10.	Heart-c	303	Yes	6	7	2

11.	Heart-h	294	Yes	6	7	2
12.	Heart-statlog	270	No	13	0	2
13.	Hepatitis	155	Yes	6	13	12
14.	Ionosphere	351	No	34	0	2
15.	Iris	150	No	4	0	3
16.	Labor	57	Yes	8	8	2
17.	Lympho	148	No	3	15	4
18.	Mushroom	8,124	Yes	0	22	2
19.	Primarytumor	339	Yes	0	17	21
20.	Sonar	208	No	60	0	2
21.	Vehicle	846	No	18	0	4
22.	Vowel	990	No	10	3	11
23.	Waveform	5,000	No	41	0	3
24.	Zoo	101	No	1	16	7

The algorithms used in the experimental study and their parameter settings, which are obtained from the WEKA [10] software tools. Several decision tree methods have been selected and compared to determine whether the proposal is competitive in different domains with the other approaches. Algorithms are compared on equal terms and without specific settings for each data problem. The parameters used for the experimental study in all decision tree methods are the optimal values from the tenfold cross-validation, and they are now detailed in Table 2.

Table 2: Experimental Settings for standard decision tree algorithms

Algorithm	**Parameter**	**Value**
C4.5	confidence factor	0.25
	min number of objects	2.0
REP	maximum depth	no restriction
	min number of objects	2.0
	min variance proportion	0.001
CART	number of folds pruning	5
	min number of objects	2.0
NB Tree	technique used at leaves	naive bayes

Results And Discussion

In this section, we investigate the results of thecompared and proposed algorithms from boththe macro and micro perspectives. From the macro level, we analyze the general properties and classifications of the instances. In addition, we test thecomplexity of the formed tree. From the micro level, we study the detailed data sources which are affectedby the proposed technique. Through the analysis, we aim at gaining more insights of the mechanism of decision tree induction with the new proposed approach.

The comparative experimental results of the proposed algorithm with benchmarks are presented in tables 3, 4 and Figure 1, 2. In this study evaluation metrics such as accuracy ad tree size are used for the comparison.The win and loss of the proposed algorithm is indicated by the '●' and '○' symbols respectively beside every value. The significance for the win and loss is considered according to the T-test at 95% confidence level. The experimental comparison of the proposed algorithm is done with each and every algorithm independently. This methodology will expose strengths and limitations of our proposed algorithm with respect to compared algorithms.

We perform classification and report the evaluation results in Table 3 and 4. Evaluationresults show the overall accuracy and tree size values are improved for our proposed ATOD algorithm. It shows, in general, high accuracy for all the datasets and relatively lower tree size values are responsible for such an overall accuracy. The performance of the event classification phase directly depends on the simplicity of the tree generated. Thus, the classification performance can be improved if the errors of detection can be reduced. Thereafter, we perform summary of experiential results and report in Table 5. Comparisons between Table 3, Table 4 and Table 5 clearly show that two-phase implementation (i.e., when attribute transformation and outlier detection performed in a series) is better than one-pass implementation (i.e., when detection and classification performed together).

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

S.No	Datasets	C4.5	REP	CART	NB Tree	ATOD
1.	Anneal.ORIG	92.35●	91.89●	93.36●	97.13	97.13
2.	Balance-scale	77.82●	78.54●	78.73●	75.96●	94.29
3.	Breast-cancer	74.28●	69.35●	70.22●	70.99●	90.46
4.	Breast-cancer-w	95.01●	94.77●	94.74●	96.37●	99.82
5.	Horse-colic	85.16●	84.94●	85.37●	81.11●	87.62
6.	Credit-rating	85.57●	84.75●	84.99●	85.42●	95.71
7.	German_credit	71.25●	72.02●	73.43●	74.64●	88.92
8.	Pima_diabetes	74.49●	74.46●	74.56●	74.96●	90.34
9.	Glass	67.63●	65.54●	71.26●	69.84●	78.91
10.	Heart _c	76.94●	77.02●	78.68●	80.03●	91.34
11.	Heart-h	80.22●	78.56●	79.02●	81.50●	90.43
12.	Heart-statlog	78.15●	76.15●	78.07●	80.93●	96.94
13.	Hepatitis	79.22●	78.62●	77.10●	81.30	81.64
14.	Ionosphere	89.74●	89.46●	88.87●	90.03	90.96
15.	Iris	94.73●	93.87●	94.20●	93.47●	97.31
16.	Labor	78.60●	78.27●	80.03●	91.63●	97.80
17.	Lymphography	75.84●	75.33●	77.21●	81.90●	85.32
18.	Mushroom	100.00	99.98	99.95	100.00	99.83
19.	Primary-tumor	41.39●	38.71●	41.42●	47.50●	74.87
20.	Sonar	73.61●	72.69●	70.72●	77.11○	76.76
21.	Vehicle	72.28●	70.18●	69.91●	70.98●	85.43

22.	Vowel	80.20●	66.67●	79.61●	92.35○	83.19
23.	Waveform	75.25●	76.57●	76.65●	79.84●	94.59
24.	Zoo	92.61●	40.61●	40.61●	94.73	94.60
Win/Tie/Loss		**(23/1/0)**	**(23/1/0)**	**(23/1/0)**	**(17/5/2)**	

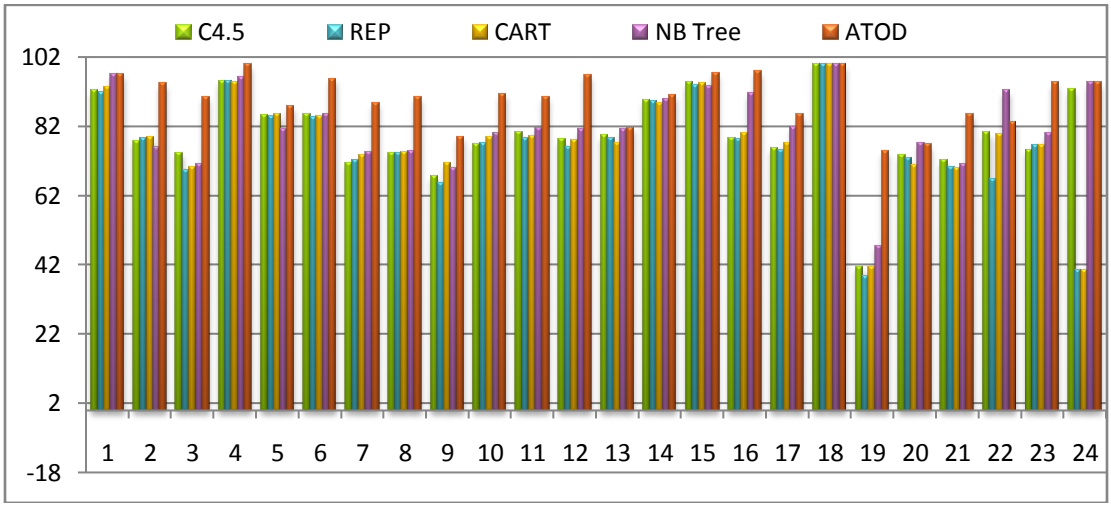
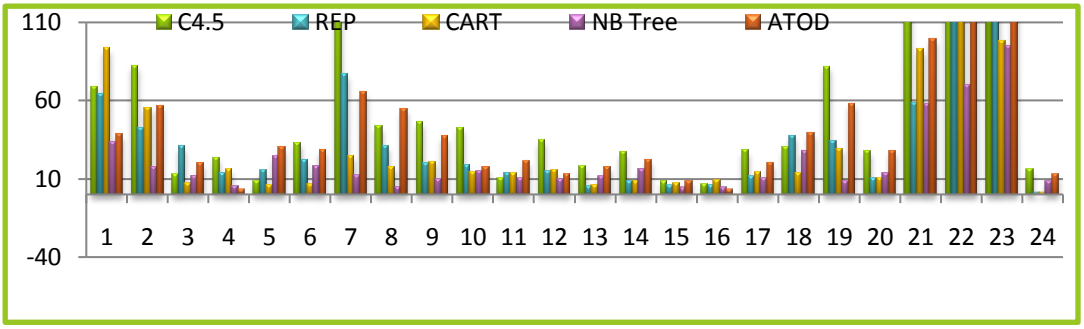


Table 4: Summary of tenfold cross validation performance for Tree Size on all the datasets

Datasets	**C4.5**	**REP**	**CART**	**NB Tree**	**ATOD**
Anneal.ORIG	68.64●	63.53●	93.22●	32.93○	38.46
Balance-scale	82.20●	42.36○	55.28○	17.38○	56.72
Breast-cancer	12.78○	30.70●	7.16○	11.90○	19.72
Breast-cancer-w	23.46●	13.76●	15.90●	5.68●	3.00
Horse-colic	8.80●	15.19○	6.42○	24.27○	29.84
Credit-rating	32.82●	22.03○	6.54○	17.90○	28.44
German_credit	126.85●	76.81●	24.46○	12.07○	65.00
Pima_diabetes	43.40○	30.98○	17.36○	5.18○	54.34
Glass	46.16●	19.70○	21.16○	10.0○	37.12
Heart-c	42.52●	18.39●	13.82○	14.58○	17.40
Heart-h	10.53●	13.63○	13.42○	10.61○	21.34
Heart-statlog	34.64●	14.78●	15.36●	9.62○	12.88
Hepatitis	17.66	5.64○	6.04○	11.56○	17.46
Ionosphere	26.74●	8.76○	8.42○	16.20○	22.14
Iris	8.28	5.84○	7.40○	4.38○	8.48
Labor	6.92●	6.15●	9.32●	4.46●	3.00
Lymphography	28.00●	11.46○	13.92○	10.24○	20.42
Mushroom	29.94○	37.54○	13.24○	27.55○	39.12
Primary-tumor	81.51●	33.50○	29.04○	8.79○	57.96
Sonar	27.90	10.20○	10.50○	13.74○	27.70
Vehicle	138.0●	58.52○	92.54○	57.70○	99.14

Vowel	209.81●	254.36	171.74○	70.10○	203.28
Waveform	591.94●	167.24○	98.32○	94.48○	171.44
Zoo	15.70●	1.00○	1.00○	8.34○	13.00
Win/Tie/Loss	**(18/3/3)**	**(17/1/16)**	**(4/0/20)**	**(2/0/22)**	



Results also indicate that anneal, balance scale, breast cancer-w, heart-c, heart-h, heart-statlog, labor, primary tumor, vehicle, vowel, waveform and zoodatasets are relatively easier for both improved classification and decrease in tree size. In contrast, complex datasets, i.e. breast cancer, pima diabetes and sonarare difficult to identify and/or classify. This led the importance of proper feature selection for event identification and classification both.

Following conclusions can be drawn

1. ATOD decision tree algorithm classify UCI datasets with animproved accuracy, there are only 2loses against NB Tree and 0 loses against C4.5, REP and CART decision tree classifiers.

Table 5: Summary of experimental results of ATOD vs compared algorithms

System	**Wins**	**Ties**	**Loss**
Accuracy			
ATOD vs C4.5	23	1	0
ATOD vs REP	23	1	0
ATOD vs CART	23	1	0
ATOD vs NB Tree	17	5	2
Tree Size			
ATOD vs C4.5	18	3	3
ATOD vs REP	7	1	16
ATOD vs CART	4	0	20
ATOD vs NB Tree	2	0	22

2. ATOD algorithm is best classification technique with highest accuracy and minimal tree size to generate decision tree for classification.
3. ATOD as a data mining technique is very useful in the process of knowledge discovery. In addition, using this technique is very convenient since the

decision tree are simple to understand, works with mixed data types, models non-linear functions, handles classification, and most of the readily available tools use it.

4. Using the same data sets with different mining techniques and comparing results of each technique in order to construct a full view of the resulted patterns and levels ofaccuracy of each technique may be very useful for exploring strengths and exposing weakness of the proposed model.

Conclusion

In this paper we have proposed a supervised machine learning approach for improved classification that involves identification of attribute transformationandclassification of them into the predefined classes. We have used statistical attributetransformation technique and outlier detection technique for improved classification performance. Firstly, we consideredthe problem of classification as a two-step process, firststep of which dealt with the attribute transformation and outlier detection whereas the second step was involved withthe classification of these data sources by using C4.5 as the base algorithm. The performance of the proposed algorithm is impressive.

References

[1]. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2nd edn. (2001)

[2]. C. V. P. R. Prasad, Dr. Bhanu Prakash Battula"A Survey On decision Tree Learning Algorithms for Knowledge Discovery", Int. Journal of Engineering Research and Application, Vol. 5, Issue 2, (Part -2) February 2015.

[3]. N. Sivaram and K. Ramar "KNOWLEDGE ENGINEERING TO AID THE RECRUITMENT PROCESS OF AN INDUSTRY BY IDENTIFYING SUPERIOR SELECTION CRITERIA", ICTACT JOURNAL ON SOFT COMPUTING, JANUARY 2011, ISSUE: 03, pp; 138-144.

[4]. P. Karthigayani, S. Sridhar" DECISION TREE BASED OCCLUSION DETECTION IN FACE RECOGNITION AND ESTIMATION OF HUMAN AGE USING BACK PROPAGATION NEURAL NETWORK" Journal of Computer Science 10 (1): 115-127, 2014, Science Publications, doi:10.3844/jcssp.2014.115.127.

[5]. Hiroshi Imamura, Keiji Sano, Yasuhiko Sugawara, Norihiko Kokudo, and Masatoshi Makuuchi" Assessment of hepatic reserve for indication of hepatic resection:decision tree incorporating indocyanine green test", J Hepatobiliary PancreatSurg (2005) 12:16–22,DOI 10.1007/s00534-004-0965-9.

[6]. Masud Karim, Rashedur M. Rahman"Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for

Direct Marketing", *Journal of Software Engineering and Applications*, 2013, 6, 196-206 http://dx.doi.org/10.4236/jsea.2013.64025 Published Online April 2013 (http://www.scirp.org/journal/jsea).

[7]. S. Kluska-Nawarecka, Z. Górny, B. Mrzygłód, D.Wilk-Kołodziejczyk, K.Regulski"Methods of development fuzzy logic driven decision-support models in copper alloys processing", ARCHIVES of FOUNDRY ENGINEERING, Academy of Sciences

[8]. Ihsan A. Kareem 1, Mehdi G. Duaimi 2" Improved Accuracy for Decision Tree Algorithm Based on Unsupervised Discretization", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.6, June-2014, pg. 176-183.

[9]. Mohammad Nazaripour" Data Analysis and Decision Trees for Analysis and B2C Controls", Asian Journal of Business Management 4(4): 376-385, 2012.

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

[11]. Blake C, Merz CJ (2000) UCI repository of machine learning databases. Machine-readable data repository. Department of Information and Computer Science, University of California at Irvine, Irvine. http://www.ics.uci.edu/mlearn/MLRepository.html