# Multi-path Decision Tree

Huaping Guo, Ming Fan
School of Information Engineering, ZhengZhou University, P. R. China
hpguo.gm@gmail.com, mfan@zzu.edu.cn

***Abstract*—Decision trees are well-known and established models for classification and regression. In this paper, we propose multi-path decision tree algorithm (MPDT). Different from traditional decision tree where the path of each record is deterministic and exclusive, a record can trace several paths simultaneously in multi-path decision tree so that it has the effect of ensemble classifiers with only one classifier. Local class information gain is the value of class information (entropy or Gini, etc) given the value of an attribute relative to class information unsupervised. We examine the MPDT on a random selection of 26 benchmark data sets from the UCI repository and compared it with Bagging, AdaBoost and C4.5. The results note that MPDT has better performance.**

***Keywords-component; formatting; style; styling; insert***

## I. INTRODUCTION

Decision trees, which encompass many diverse applications including diagnostics for linear regression [1], hierarchical multi-label classification [2] etc, have been among the most wide-spread models in machine learning. Their advantages include the interpretability of the model and the capability to handle both numerical attributes and categorical attributes even with missing values. Some decision tree classifier's algorithms, such as ID3 [3], C4.5 [4], CART [5] etc, are often considered as the benchmark of evaluating the performance of other classifiers. Traditional decision tree classifications are all single path—all training and test records trace only one path from root to some or other leaf. When building the decision tree, records are partitioned into successively purer subsets in a recursive fashion, based on an attribute test condition. It can work very well in general circumstance; however, there is no remedy when error occurred.

Combining classifiers is now an established research area known under different names in the literature: committees of learners, mixtures of experts, classifier ensembles, multiple classifier systems, etc. Ensemble Methods, such as Bagging [6], Adaboost [7,8], Random Forests [9], etc, based on decision tree can improve decision tree classification accuracy effectively. For the theories, you can refer to [10, 11]. The key to the success of these algorithms is that, intuitively at least, they build a set of diverse classifiers. We can penetrate the issue with another perspective that each test record traces several paths, in which each path belongs to one decision tree in the combining classifier, offsetting the deficiencies with each other. However, we argue that, for some records, there may be some redundancies in the paths, even all the paths to be the same in the extreme. For these cases, the combining classifier doesn't have the effect of enhancing the performance for some test records, and on the contrary, slows down the speed of classifying them.

To solve problems mentioned above and, at the same time, utilize the merit of decision tree and combining classifiers, we present Multi-path decision tree. In the tree, there are three kinds of node called test node, leaf node and hub node which distributes the record arriving at the node to all of its children nodes immediately. Experimental results denote that MPDT has better accurate.

The paper is organized as follows. Section 2 gives the definition of some concepts used in this paper. Section 3 introduces how a multi-path decision tree works and gives the algorithm of building the tree. Section 4 presents the method of controlling the tree's size. Section 5 proves prove the equipollence in effect between Multi-path decision tree and combining classifiers based on decision tree. Section 6 describes the experiments conducted, and discusses the results. Conclusions are summarized in Section 7.

## II. DEFINITION

**Definition 1**: A directed acyclic graph $G = (V, E)$ consists of a finite, non-empty set of *nodes (or vertices)* $V$ and a set of *edges* $E$. A directed tree is a directed acyclic graph satisfying the following properties

- There is exactly one node, called the *root*, which no edges enter. The root node contains all the class labels.
- Every node except the root has exactly one entering edge.
- There is a unique path from the root to each node.

If $(v, w)$ is an edge in a tree, then $v$ is called the father of $w$, and $w$ is a *son* of $v$. If there is a path from $v$ to $w$ ($v \neq w$), then $v$ is a proper ancestor of $w$ and $w$ is a proper descendant of $v$. A node with no proper descendant is called a leaf (or a terminal). All other nodes (except the root) are called internal nodes. The depth *of a node* $v$ in a tree is the length of the path from the root to $v$. The height of node v in a tree is the length of a largest path from $v$ to a leaf. The height of a tree is the height of its root. The level of a node $v$ in a tree is the height of the tree minus the depth of $v$.

**Definition 2**: A decision tree is a directed tree satisfying the following properties

- Each leaf node is assigned a class label.
- The non-terminal nodes, which include the root and other internal nodes, contain attribute test conditions to separate records that have different characteristics.

Published by Atlantis Press, Paris, France.

**Definition 3**: Based on the definition 1, $V$ contains a kind of node called combination node $v_i$. The tree is called the Combination tree if the node $v_i$ satisfying the following properties

- There is entering edge $(v_j, v_i)$ at most and one outgoing edge at least.
- Combination node $v_i$ disassembles the tree into $\{V_1, V_2, \ldots, V_n, V-V_1-V_2 - \ldots-V_n, \{ v_i \}; E_1, E_2, \ldots, E_n, \{(v_i, v_c)| c = 1,2,3,\ldots,n\}, \{(v_j, v_i)\}, E - E_1 - E_2 - \ldots-E_n\}$. $G' = \{ V_c , E_c\}$ is combination tree whose root is $v_c$,. $v_c$ is end node of the edge $(v_i, v_c)$.
- the combination tree $G$ denotes: $G$ is the combination model of the combination trees--$\{V_1 \cup (V-V_1-V_2 - \cdots-V_n), \{(v_j, v_1)\} \cup (E - E_1 - E_2 - \cdots-E_n) \cup E_1\}, \{V_2 \cup (V-V_1-V_2 - \cdots-V_n), \{(v_j, v_2)\} \cup (E - E_1 - E_2 - \cdots - E_n) \cup E_2\}; \cdots; \{V_n \cup (V-V_1-V_2 - \cdots - V_n), \{(v_j, v_n)\} \cup (E - E_1 - E_2 - \cdots-E_n) \cup E_n\}$

**Algorithm 1:** Create a multi-path decision tree

**Input:** a set of training samples, and a set of attributes
**Output:** a multi-path decision tree
**begin:**
1. Create a hub node as the root of MPD-tree, denoted as *root*;
2. $D \leftarrow$ all training samples
3. *AttributeList* $\leftarrow$ all attributes
4. MPDT-Growth(*D*, *AttributeList*, *root*) //where，MPDT-Growth is defined as following:

**MPDT-Growth**(*D*, *AttributeList*, *T*)
1. **if** stopping criterion is met **then**
2. change *T* into leaf node
3. *T.label* $\leftarrow$ class distribution of samples in *D*;
4. **else**
5 *selsetedAttributes* $\leftarrow$ select one or more "best" attributes from *AttributeList*;
6. **for** each *A* in *selsetedAttributes* **do**
// deal with each selected attribute
7. Create a test node *N* as a child of *T*;
8 *N.label* $\leftarrow$ *test-condition* of attribute *A*;
9. **for** each outcome $v_j$ of *test-condition* **do**
// split samples in *D*, and grow sub-trees for each partition
10. Let $D_j$ be the samples in *D* which meet outcome $v_j$ of *test-condition*;
11. Create a hub node $T_j$ as a child of *N*;
12. **if** *A* is discrete attribute **then**
13. MPDT-Growth($D_j$, *AttributeList*−{*A*}, $T_j$);
14. **else**
15. MPDT-Growth($D_j$, *AttributeList*, $T_j$);
16. **end if**
17. **end for**
18. **end for**
19. **end if**

**Definition 4**: A multi-path decision tree is a combination tree satisfying the following properties.

- The combination nodes, called hub nodes here, do not contain attribute test. The main function of it is distribute all the records arrived to all its children's node.
- Each leaf node is assigned a class label.
- All the nodes, except hub nodes and leaf nodes, contain attribute test conditions to separate records that have different characteristics.

Comparing definition 2 with definition 4, we know that the only difference between Decision tree and Multi-path decision tree is multi-path decision tree has one more kind of node called hub node.

## III. ALGORITHM

The algorithm of MPDT (multi-path decision tree) is shown in algorithm 1. The idea of MPDT is to build hub-node which immediately distributes the record arriving the node to all its sub-nodes (refer to 9-19 in algorithm MPDT-Growth).

## IV. EXPERIMENT

The experiment aims to evaluate the MDPT performance, and compares MDPT with traditional ensemble learning algorithms: C4.5, Random Forest, Bagging and Boosting, where the base classifier of ensembles is C4.5. 26 data sets are randomly selected to test the performances of the corresponding algorithms, and the details of the data sets are shown in table 1. For each data set, 10-fold cross-validations are conducted.

Table 1. Data sets used in experiments

| DataSet | #Instance | #Attribute | #Class |
|---|---|---|---|
| anneal | 998 | 38 | 5 |
| australian | 690 | 14 | 2 |
| breast | 699 | 10 | 2 |
| cleve | 303 | 13 | 2 |
| crx | 690 | 15 | 2 |
| debate | 768 | 8 | 2 |
| german | 1000 | 20 | 2 |
| heart | 270 | 13 | 2 |
| hepatitis | 155 | 19 | 2 |
| hypothyroid | 3163 | 25 | 2 |
| ionosphere | 351 | 342 | 2 |
| labor-neg | 57 | 16 | 2 |
| letter | 20000 | 16 | 26 |
| liver | 345 | 6 | 2 |
| mushroom | 8124 | 22 | 2 |
| pima | 768 | 8 | 2 |
| shuttle | 58000 | 9 | 7 |
| sick | 4744 | 29 | 2 |
| sonar | 208 | 60 | 2 |
| tic-tac-toe | 958 | 9 | 2 |
| vehicle | 846 | 18 | 4 |
| vote | 433 | 16 | 2 |
| waveform-21 | 5000 | 21 | 3 |
| waveform-40 | 5000 | 40 | 3 |
| wine | 178 | 13 | 3 |
| yeast | 1484 | 8 | 10 |

The corresponding results are shown in table 2 and table 3, where table 2 is the results of ensemble (Random Forest, Bagging, Boosting) with ten members and table 2 is the result of ensemble with 20 members.

From table 2 and table 3, we observe that comparing with other advanced methods, MDPT has better generalization ability, which indicates that MDPT is an applicable method for ensemble learning.

Table 2 The error rate of algorithms where each ensemble has 10 base classifiers.

| Data Set | CZRDT | C4.5 | Random Forests | Bagging | Boosting |
|---|---|---|---|---|---|
| anneal | 4.94 | 7.99 | 5.63 | 5.51 | 4.83 |
| australian | 13.63 | 14.64 | 16.05 | 13.77 | 16.09 |
| breast | 3.62 | 5.51 | 4.76 | 4.64 | 3.62 |
| cleve | 19.00 | 22.00 | 20.00 | 23.33 | 19.00 |
| crx | 14.49 | 15.36 | 15.51 | 14.35 | 16.23 |
| debate | 23.85 | 25.79 | 27.63 | 24.34 | 28.42 |
| german | 25.30 | 29.40 | 26.60 | 27.70 | 30.20 |
| heart | 16.67 | 21.48 | 18.15 | 20.00 | 21.48 |
| hepatitis | 19.30 | 23.33 | 16.00 | 20.67 | 16.00 |
| hypothyroid | 0.73 | 0.70 | 0.95 | 0.66 | 0.95 |
| ionosphere | 7.14 | 10.83 | 7.14 | 7.14 | 7.14 |
| labor-neg | 8.00 | 18.00 | 16.00 | 14.00 | 10.00 |
| letter | 5.02 | 13.54 | 7.55 | 12.05 | 7.95 |
| liver | 30.29 | 36.47 | 32.35 | 30.18 | 30.88 |
| mushroom | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| pima | 24.42 | 26.32 | 23.95 | 24.87 | 29.21 |
| shuttle | 0.00 | 0.03 | 0.02 | 0.02 | 0.02 |
| sick | 2.03 | 2.03 | 2.44 | 2.03 | 2.88 |
| sonar | 17.50 | 23.50 | 22.50 | 19.50 | 22.00 |
| tic-tac-toe | 8.00 | 15.58 | 21.79 | 17.14 | 15.58 |
| vehicle | 24.89 | 27.62 | 23.57 | 25.95 | 21.69 |
| vote | 8.37 | 11.63 | 9.53 | 10.23 | 12.09 |
| waveform-21 | 17.06 | 24.96 | 17.50 | 18.80 | 17.00 |
| waveform-40 | 16.70 | 23.12 | 17.94 | 18.04 | 17.58 |
| wine | 3.97 | 6.41 | 1.60 | 4.81 | 4.28 |
| yeast | 42.84 | 44.39 | 44.46 | 40.95 | 44.32 |
| **Average** | **13.76** | **17.33** | **15.37** | **15.41** | **15.36** |

Table 3 The error rate of algorithms where each ensemble has 20 base classifiers.

| Data Set | MPDT | C4.5 | Random Forests | Bagging | Boosting |
|---|---|---|---|---|---|
| anneal | 4.94 | 7.99 | 4.94 | 5.73 | 4.94 |
| australian | 13.63 | 14.64 | 14.06 | 13.19 | 15.18 |
| breast | 3.62 | 5.51 | 4.06 | 4.35 | 3.33 |
| cleve | 19.00 | 22.00 | 20.00 | 19.00 | 16.00 |
| crx | 14.49 | 15.36 | 14.20 | 14.06 | 15.07 |
| debate | 23.85 | 25.79 | 26.05 | 23.82 | 27.50 |
| german | 25.30 | 29.40 | 25.40 | 27.70 | 28.10 |
| heart | 16.67 | 21.48 | 19.63 | 18.89 | 21.85 |
| hepatitis | 19.30 | 23.33 | 16.67 | 19.33 | 16.00 |
| hypothyroid | 0.73 | 0.70 | 1.01 | 0.70 | 0.95 |
| ionosphere | 7.14 | 10.83 | 6.29 | 7.14 | 5.71 |
| labor-neg | 8.00 | 18.00 | 14.00 | 14.00 | 8.00 |
| letter | 5.02 | 13.54 | 7.55 | 12.05 | 7.95 |
| liver | 30.29 | 36.47 | 28.82 | 29.12 | 27.94 |
| mushroom | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| pima | 24.42 | 26.32 | 23.98 | 23.55 | 26.76 |
| shuttle | 0.00 | 0.03 | 0.02 | 0.02 | 0.36 |
| sick | 2.03 | 2.03 | 2.21 | 2.12 | 2.38 |
| sonar | 17.50 | 23.50 | 20.00 | 19.00 | 17.00 |
| tic-tac-toe | 8.00 | 15.58 | 5.68 | 7.16 | 1.79 |
| vehicle | 24.89 | 27.62 | 24.17 | 26.90 | 21.90 |
| vote | 8.37 | 11.63 | 10.00 | 9.30 | 12.09 |
| waveform-21 | 17.06 | 24.96 | 16.32 | 17.82 | 17.00 |
| waveform-40 | 16.70 | 23.12 | 16.80 | 17.70 | 17.58 |
| wine | 3.97 | 6.41 | 1.67 | 5.23 | 4.62 |
| yeast | 42.84 | 44.39 | 40.81 | 40.14 | 42.03 |
| **Average** | **13.76** | **17.33** | **14.01** | **14.54** | **13.92** |

## REFERENCES


[1] S. Xiaogang, T. Chih-Ling and C. W. Morgan, “Tree-structured model diagnostics for linear regression,” Mach Learn, 74: 111–131, 2009.

[2] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, H. Blockeel, “Decision trees for hierarchical multi-label classification” Mach Learn, 73: 185–214, 2008.

[3] J. R. Quinlan. “Discovering rules by induction from large collection of examples”, In D. Michie, editor, Expert Systems in the Micro Electronic Age, Edinburgh University Press, Edinburgh, UK, 1987.

[4] J. R. Quinlan, “C4.5: programs for machine learning”, San Mateo: Morgan Kaufmann, 1993.

[5] L. Breslow, D. W. Friedman, R. Olshen and C. J. Stone, “Classification and Regression Trees”, Chap & Hall, New York, 1984.

[6] L. Breiman, “Bagging Predictors”, Machine Learning, 24(2):123 – 140, 1996.

[7] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning of on-line learning and application to boosting”, Journal of Computer and System Sciences, 55(1):119 – 139, 1997.

[8] L. Breiman, “Random Forests”, Machine Learning, 45(1): 5 – 32. 2001

[9] I. Kuncheva and L. J. Whitaker, “Measures of Diversity in classifier Ensembles and Their Relationship with the Ensemble Accuracy”, Machine Learning, 51, 181-207, 2003.