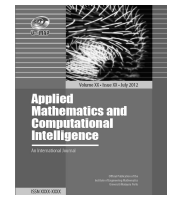
App. Math. and Comp. Intel., Vol. 1 (2012) 33–41
http://amci.unimap.edu.my



Similarity measure exercise for classification trees based on the classification path

N. Hasan[a,b*], M. B. Adam[b], N. Mustapha[b], M. R. Abu Bakar[b]

[a] *Faculty of Science and Mathematics*
Universiti Pendidikan Sultan Idris
35900 Tanjong Malim, Perak, Malaysia.

[b] *Institute for Mathematical Research (INSPEM)*
Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia.

Received: 29 February 2012; Revised: 17 July 2012; Accepted: 18 July 2012

Abstract: *Classification tree models are known for their simplicity and efficiency when dealing with domains contain large number of variables and cases. However, a small perturbation in the data, can lead to a very different tree. We introduce a method for measuring similarity between binary classification trees based on the similarity between the classification paths. The trees to be compared are represented in the form of matrices whose entries are in the interval [0,1]. Overlap similarity measure is used to measure the similarity between each pair of path in two trees, and the best matching paths between trees are used to calculate the similarity measure. This method has advantage to measure trees that possess the same structure and leaf nodes but different internal node.*

Keywords: *Binary classification tree; Classification paths; Overlap similarity measure.*

PACS: *02.70.Rr*

1 Introduction

Classification trees are widely used in various fields such as medicine (diagnosis), computer science (data structures), education (classification and prediction), and psychology (decision theory). Classification tree models are known for their simplicity and efficiency when dealing with domains with large number of variables and cases. Classification trees readily lend themselves to being displayed graphically, helping to make them easier to interpret than they would be if only a strict numerical interpretation were possible. However, classification trees are known for their instability [11]. A small perturbation in the data, or a new sample, can lead to a very different tree particularly if the change occurs in top level nodes and thus

*Corresponding Author: norsida@fsmt.upsi.edu.my (N. Hasan)

give a different misclassification rate.

Classical methods focus on structural and geometrical characteristics of trees, mainly considering the branching structure of classification trees. Zhong, et. al.[10] have develop a webbing matrix method to calculate the overall similarity of two leaf-labelled trees. They use similarity measure for ordinary sets to compare all paires of subtrees which are simply reduced to their respective leaf node sets. Ganesan et.al [7] consider a hierarchical domain structure to produce more intuitive similarity score. The application of fuzzy for measuring an overall degree of similarity between different leaf-labelled trees has been introduced in [12]. Yu, et.al. [5] develop a simple probabilistic approach known as total ancestry method for computing relatedness quantity. This method is based on counting the number of leaf nodes that share exactly the same set of ancestor nodes in comparison to the total number of classified pairs. Briand, et.al [2] measure the similarity of splitting rule at each internal node to construct classification tree that is more stable than the classical classification tree.

Tree $T1$ and Tree $T2$ in figure 1 and figure 2 have similar structure but different splitting rule in some nodes. Existing similarity measure such as overall similarity algorithm is unable to detect the difference and will consider both trees to have 100 percent similarity. In this paper, we develop a new similarity measure method that employ the classification path which lead to a better similarity measure.

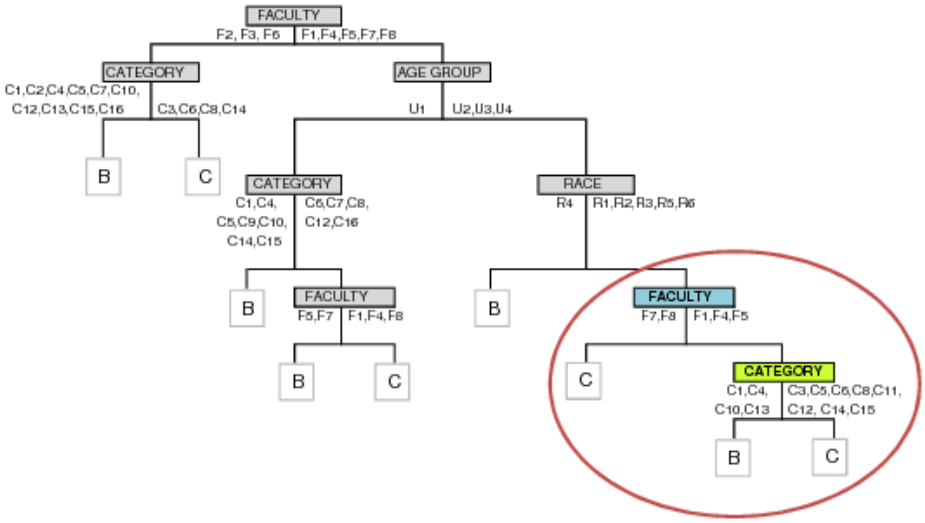


Figure 1: Classification tree $T1$

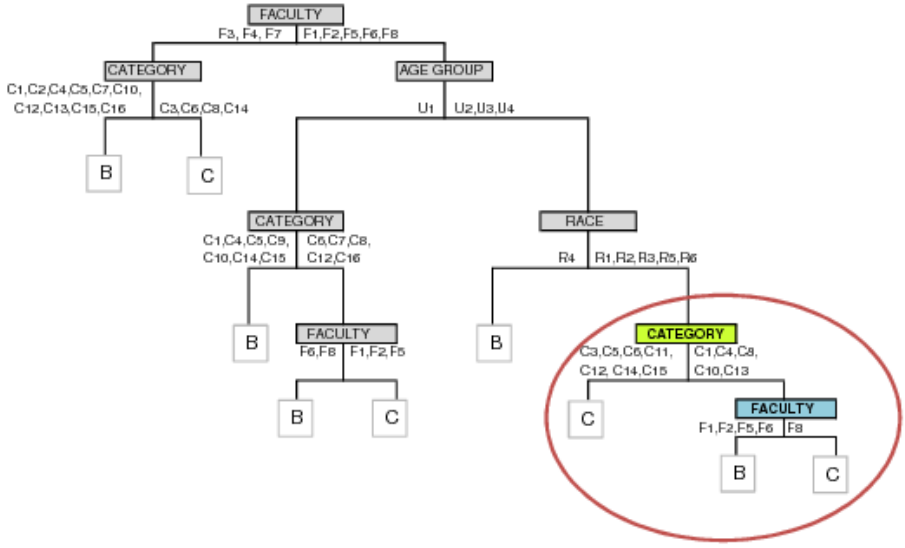


Figure 2: Classification tree $T2$

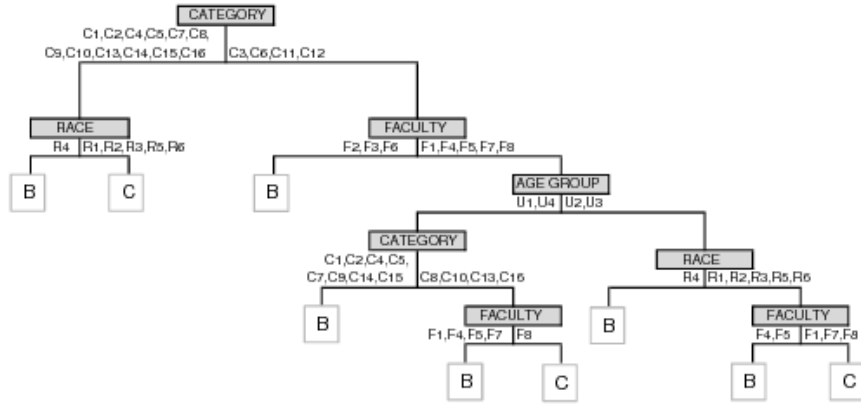


Figure 3: Classification tree $T3$

2 Related Works

2.1 Similarity Measures for Categorical Data

The similarity or distance measure for categorical data is not as straightforward as for continuous data. It is not possible to directly compare two different categorical values they are not inherently ordered. The simplest way to measure similarity between two categorical data is overlap measure [3]. Such measure assign a similarity of 1 if the values are identical and a similarity of 0 if the values are not identical. Similarity measure for two multivariate categorical data points is directly proportional to the number of attributes in which they match. Such measure treats all matches or mismatches as equally importance. Other measures are known as data-driven measures for categorical attributes which take into account the frequency distribution of each attribute values. Boriah, et.al [9] have evaluated the performance of a variety of similarity measure for categorical data. They have shown that no one measure dominates others for all types of problems.

The existing association coefficients for binary characters, i.e., present and absent, can be used to compute the subtree similarity. For two objects, let a be the number of features that are common to both objects, b and c be the number of features that are present in only one object and d be the number of features that both objects lack. Common choices of similarity measure S_{ij} formula are as follows:

$$S_{ij} = \frac{a}{a+b+c} \tag{1}$$

$$S_{ij} = \frac{2a}{2a+b+c} \tag{2}$$

$$S_{ij} = \frac{a+d}{a+b+c+d} \tag{3}$$

$$S_{ij} = \frac{a+d-b-c}{a+b+c+d} \tag{4}$$

Eq.(1) is known as Jaccard coefficient. Jaccard coefficient is the most popular similarity measure. It is a measure that omits occurrences of negative matches and reads as the number of common features in both objects over the number of all features present in either one of them. Eq.(2) is known as Dice coefficient [6]. It is not very different in form from the Jaccard index but has some different properties. Eq.(3) is known as simple matching coefficient [8] while eq.(4) is known as Hamann association coefficient [1]. According to Gower and Legendre [4], one criterion to choose an appropriate similarity measure for a certain problem is whether to include conjoint absences, d or not often leads to a discussion. In some situations, it would seem ridiculous to compare two objects on the basis of the features they both lack, but in other situations it would seem improper to neglect these conjoint absences.

2.2 Overall Similarity Measure

There were quite a number of research done to measure similarity between trees based on terminal node. The most popular method is based on overall similarity [10]. Suppose that two tree $T1$ and $T2$ can be partitioned into M and N complete set of subtrees respectively. A subtree of a tree is defined as part of a tree in which the root of the subtree is a non-terminal node. A complete set of subtrees consist of all non-terminal nodes in a classification tree. The overall similarity $Sim(T1, T2)$ is defined as follows:

$$Sim(T1, T2) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} S_{ij} \tag{5}$$

where
$i = 1, 2, ...M$ is the subtrees in $T1$
$j = 1, 2, ...N$ is the subtrees in $T2$
S_{ij} is the similarity measure between subtrees i and j

In this method, two complete subtree sets in the two trees are written out as the column and row headings of a matrix. All elements in the matrix which are not equal to 1 will be given an actual value of S_{ij}. However, when value of any cell is equal to one, asterisks will be placed in all other elements of its column and row in the matrix. It means that if two subtrees are exactly matched, other subtrees do not need to be compared with either one of the two subtrees. Meanwhile, those asterisks might be assigned a constant value. The span of the output is originally [0, 1] where 0 indicates that there is no overall resemblance and 1 indicates that the trees are identical.

3 Methodology

3.1 Similarity Measures Based on the Classification Path

We proposed a method to compute the similarity of classification trees in term of the agreement of the classification path of each prediction classes. Similarity measures based on classification path for trees $T1$ and $T2$ is obtained by first grouping all paths based on their leaf nodes class. Path is defined as branches in a tree from the root to a leaf. For each leaf nodes class, we compare the splitting variable used at each internal node for each path. The webbing matrix for measures similarity based on classification path has different column and row headings from classical webbing matrix. In this webbing matrix, two complete paths sets in the two trees are written out as the column and row headings of a matrix. All elements in the matrix which are not equal to 1 will be given an actual value of S_{ij}. Then, we select the best matching splitting variables among paths in $T1$ and $T2$.

The procedure to calculate the internal node similarity of trees T1 and T2 can be described in four steps.

- STEP 1: List all paths for all leaf nodes class.
- STEP 2: Calculate distance matrix S_{ij}.
- STEP 3: Find the best matching paths between $T1$ and $T2$.
- STEP 4: Calculate the similarity score based on best matching paths between $T1$ and $T2$.

This dataset produced two outcomes namely class B and class C. Tree $T1$ and $T2$ have nine paths while $T3$ has 10 paths. The classification path for class B and C in Tree $T1$ are given in Table 1.

Table 1: Classification Paths for Tree $T1$

Class	Path
B	Faculty - Category.
	Faculty - Age Group - Category
	Faculty - Category - Age Group - Faculty
	Faculty - Age Group - Race
	Faculty - Age group - Race - Faculty - Category
C	Category, Faculty
	Faculty, Category, Age Group, Faculty
	Category, Faculty, Race, Age Group, Faculty
	Faculty, Race, Age Group, Faculty

T2-B	T1 - B				
	FC	FAC	FACF	FAF	FARFC
FC	1	4/5	2/3	2/5	4/7
FAC	4/5	1	6/7	2/3	3/4
FACF	2/3	6/7	1	4/7	8/9
FAF	2/5	2/3	4/7	1	3/4
FARCF	4/7	3/4	8/9	3/4	1

T2 - C	T1 - C			
	FC	FACF	FARF	FARFC
FC	1	2/3	1/3	4/7
FACF	2/3	1	3/4	8/9
FARC	2/3	3/4	3/4	8/9
FARCF	4/7	8/9	8/9	1

Figure 4: The webbing matrix to compute the similarity between trees $T1$ and $T2$

T3-B	T1 - B				
	FC	FAC	FACF	FAF	FARFC
CFi	1/2	2/5	1/3	2/5	4/7
CF	1	4/5	2/3	2/5	4/7
CFAC	2/3	6/7	3/4	4/7	2/3
CFACF	4/7	3/4	8/9	1/2	4/5
CFARC	4/7	3/4	8/9	1/2	4/5
CFARF	4/7	3/4	8/9	3/4	1

T3 - C	T1 - C			
	FC	FACF	FARF	FARFC
CF	1/2	1/3	1/3	4/7
CFACF	4/7	8/9	2/3	4/5
CFARC	4/7	8/9	2/3	4/5
CFARFR	4/7	4/9	8/9	1

Figure 5: The webbing matrix to compute the similarity between trees $T1$ and $T3$

T3-B	T2 - B				
	FC	FAC	FACF	FAF	FARCF
CFi	1/2	2/5	1/3	2/5	4/7
CF	1	4/5	2/3	2/5	4/7
CFAC	2/3	6/7	3/4	4/7	2/3
CFACF	4/7	3/4	8/9	1/2	4/5
CFAFC	4/7	3/4	8/9	1/2	4/5
CFAFF	4/7	3/4	8/9	3/4	1

T3 - C	T2 - C			
	FC	FACF	FARC	FARFC
CF	1/2	1/3	2/3	4/7
CFACF	4/7	8/9	2/3	4/5
CFAFC	4/7	8/9	1/3	4/5
CFAFR	4/7	4/9	8/9	1

Figure 6: The webbing matrix to compute the similarity between trees $T2$ and $T3$

Variables that appear twice in any path are treated as two different variables. For each of path of $T1$, we calculate how similar it is to each path of $T2$ using the webbing matrix. If $T1$ contains m paths and T2 contains n paths, there will be at most $m \times n$ similarity calculations. The distance matrix, S_{ij} for comparing $T1$, $T2$ and $T3$ are calculated using Dice similarity measure. The webbing matrix for comparing $T1$, $T2$ and $T3$ are illustrated in Figure 4, Figure 5 and Figure 6. The span of the output originally falls into the real interval [0,1] where 0 and 1 indicates totally different and identical matching, respectively. We have converted this scale to a linear percentage where 0 percent represents 0 and 100 percent represents 1.

As expected, the similarity between tree $T1$ and tree $T2$ is 97.2 percent which is closed to 100 percent. Similarly, we obtain the similarity between tree $T1$ and tree $T3$ is 72.0 percent and the similarity between tree $T2$ and tree $T3$ is 72.7 percent.

4 Disscussion

Reconsider our webbing matrix in figure 4, particularly the similarity for class C. There are perfect match for FC (Faculty-Category), FACF (Faculty-Age Group-Category-Faculty) and FARCF (Faculty-Age Group-Race-Category-Faculty) with similarity equal to 1. The best match for FARC (Faculty-Age Group-Race-Category) is FARFC with $S_{ij} = \frac{8}{9}$, but FARFC has been chosen to pair with FARCF as they match perfectly. In this paper we avoid choosing a path in any tree to pair with more than one path in another tree. Therefore, we select the second best match for FARC which is FARF (Faculty-Age group-race-Faculty) with $S_{ij} = \frac{3}{4}$.

As we can see in figure 5 and figure 6, there are total six paths in $T3$ compared with five paths in $T1$ for class B. For this case, we are unable to assign CFAFC (Category-Faculty-Age Group-Faculty-Category) to any pair and therefore $S_{ij} = 0$.

The similarity between $T1$, $T2$ and $T3$ measured using overall similarity and similarity based on the classification path are shown in Table 2. Similarity based on classification path method is able to provide a better similarity measure for our tree models since it can distinguish the small different in $T1$ and $T2$.

Table 2: Comparison of similarity measure using overall similarity and similarity based on classification path

Tree	Overall Similarity ([10])	Similarity based on Path
$T1, T2$	100%	97.2%
$T1, T3$	58.7%	72.0%
$T2, T3$	58.7%	72.7%

5 Conclusion

A new method for measuring the similarity between tree based on the classification path has been introduced as an alternative to the overall similarity measure. Any similarity measure coefficients can be used to measure the similarity between a pair of path in two trees, and the webbing matrix is used to find the best matching paths between trees. Other simmilarity measure such as data-driven similarity measure can be adapted to calculate the similarity between each pair of paths in two trees. This method has advantage to measure the similarity between trees that possess the same structure and terminal nodes but different internal node.

References

[1] A. H. Cheetham and J. E. Hazel. Binary(Present-Absent) Similarity Coefficients. *Journal of Paleontology,* 43(5):1130–1136, 1969.

[2] B. Briand, G. R. Ducharme, V. Parache and C. Mercat-Rommens. A Similarity Measure to Access the Stability of Classification Trees. *Computational Statistics and Data Analysis,* 53:1208–1217, 2009.

[3] C. Stanfill and D. Waltz. Toward Memort-based Reasoning. *Communications of the ACM,* 29(12):1213–1228, 1986.

[4] G. C. Gower and P. Legendre. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification,* 3:5–48, 1986.

[5] H. Yu, R. Jansen, G. Stolovitzky and M. Gerstein. Total Ancestry Measure: Quantifying the Similarity in Tree-like Classification With Genomic Applications *Bioinformatics,* 23(16):2163–2173, 2007.

[6] L. R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology,* 26(3):297–302, 1945.

[7] P. Ganesan, H. Garcia-Molina and J. Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Trans. Inf. Syst.,* 21(1):64–93, 2003.

[8] R. R. Sokal and C. D. Michener. A Statistical Method for Evaluating Systematic Relationships *Univ. Kans. Sci. Bull.,* 38:14093–1438, 1958.

[9] S. Boriah, V. Chandola and V. Kumar. Similarity measures for categorical data: A comparative evaluation. *Computer and Information Science,* 30(2):243–254, 2008.

[10] Y. Zhong, A. C. Meacham and S. Pramanik. A General Method for Tree-comparison Based on Subtree Similarity and its Use in a Taxonomic Databased. *BioSystem,* 42:1–8, 1997.

[11] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees*, Chapman and Hall, New York, 1984.

[12] H. Demeyer, B. De Baets and S. Janssens. Similarity measurement on leaf-labelled trees. *In EUSFLAT Conf.'01*, pages 23–256, 2001.