IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 April 2022, accepted 13 May 2022, date of publication 29 June 2022, date of current version 7 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3187172

Decision Tree Algorithm Considering Distances Between Classes

SANGYONG LEE[1], CHULHEE LEE[1], KWON GI MUN[2], AND DOHYUN KIM[1]
[1]Department of Industrial and Management Engineering, Myongji University, Yongin 17058, Republic of Korea
[2]Technology and Operations Management, California State Polytechnic University, Pomona, CA 91768, USA

Corresponding author: Dohyun Kim (ftgog@mju.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1F1A1071421).

ABSTRACT Decision tree algorithm (DT) is a commonly used data mining method for classification and regression. DT repeatedly divides a dataset into pure subsets based on impurity measurements such as entropy and Gini. Then relatively ''pure'' partitions consisting of observations with the (almost) same class are obtained. Gini index is one of the representative indices for measuring the impurity of data. However, the Gini index does not take into account distances between classes. If the distances between classes are considered when measuring impurity, the decision tree algorithm can distinguish clearly observations with different classes. To the end, a new decision tree algorithm based on Rao-Stirling index is proposed considering distances between classes. Rao-Stirling index considers distances between classes in such a way that weights more to pairs of references in more distant classes when measuring data impurity. Experimental results indicate that the proposed method is superior in terms of accuracy, implying that considering the distances between classes can help improve accuracy in DT.

INDEX TERMS Decision tree, distance between classes, Rao-Stirling index.

I. INTRODUCTION

Decision trees (DT), which are named after their tree-like structure, are commonly used in data mining. A DT divides the whole data set into several subgroups containing instances with (almost) the same classes. In general, a DT consists of parent nodes and child nodes, and the parent nodes break down the data into smaller and smaller child nodes (subsets) using specific variables selected by split criterion. Partitioning progresses in the direction reducing impurities by measuring the impurity of the child nodes until the stop rule has been reached.

DTs offer several advantages. First, they can create a non-parametric model because there is no assumption regarding the data. Second, DTs can be visualized using a tree structure, so it is easy to interpret the results and to know which variables are important. Finally, the computational cost of a DT is relatively low, and tree-based decision rules for large data sets can be generated relatively quickly.

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

However, DTs also have limitations in that they only consider the impurity measure and do not take advantage of other data features when splitting data. For example, the concept of distance can be used to take advantage of data features, as shown in Figure 1. There are six different classes of data, and these can be partitioned based on the Gini index, the impurity-based splitting criteria for CART algorithm to generate decision rules. For a detailed description of the Gini, the reader is referred to Gini [21]. Figures 1(a)-(c) all have the same Gini index decrease of 1.667 when any one split is performed, so the existing algorithm using the Gini index as the impurity measure cannot distinguish (a)-(c) at all. If the distance between classes is considered, (c) has a larger decrease than (a) and (b), as expected. In other words, the distance between classes should be considered when splitting in DT to obtain the generalized splitting boundary.

Related studies have considered the distance in decision trees. Mantaras [18] used the distance as an impurity measure by applying the distance of two partitions divided by split criterion in the attribute selection process. Another related study was presented by the work of Takahashi and Abe [4].

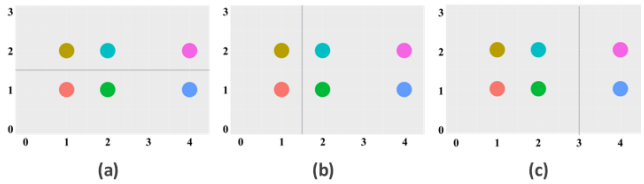

FIGURE 1. Illustration showing the importance of considering class distance when splitting observations.

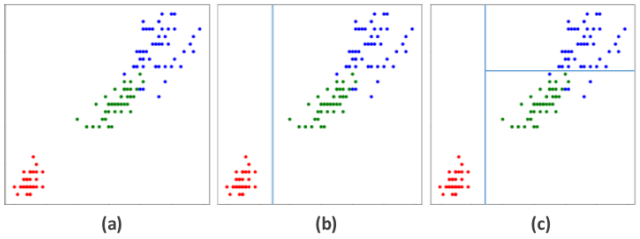

FIGURE 2. Process of splitting for the iris dataset using DT (Each color represents each class).

They proposed decision tree-based multiclass support vector machines that use the DT structure to solve multiple classification problems with SVM, and the distance between classes is used to classify distant classes first.

Related research has shown that the DT can obtain a general partition boundary if the distance between classes is considered, and this can improve the predictive performance with new data. Therefore, we propose a decision tree algorithm that considers the distance between classes as well as the impurity.

This paper is structured as follows. Section II explains the decision tree algorithms and their impurity measure. Section III describe the details of the proposed decision tree algorithm considering class distances. The experimental results are shown in Section IV, and the conclusions are given in Section V.

II. DECISION TREE AND RANDOM FOREST

This section describes how a decision tree works. A DT is a top-down approach to divide a data subset, and a variable and splitting boundary are selected at each stage of the process. Then, the dataset is repeatedly divided into pure subsets based on the impurity measure (See Figure 2 for the iterative splitting process of the DT). The DT defines the goodness of split as the difference between the degree of impurity before and after division. Therefore, a greater purity in the divided data results indicates a higher goodness in the split. As a result, the data set is split through division boundary R with the highest goodness of split defined as:

$$G(T, R) = I(T) - I(T|R)$$

where T is a set of the training example. $G(T, R)$ indicates the goodness of split when the training set T is divided by R and $I(T)$ and $I(T|R)$ indicate the impurities before and after division based on the division boundary.

DT applies its goodness of split criteria to each split point and evaluates the reduction in the impurity. Then, DT selects the best split point of the variable in which the reduction in the impurity is the highest. DT has impurity metrics that can be used to determine the splitting boundary. The impurity metrics are defined according to informatics and statistical approaches, such as the Information gain, Gini Index, gain ratio, distance measure [19]. The information gain [20] is an impurity-based criterion that uses entropy (origin from information theory) as a measure of impurity. The Gini index measures the divergence between the probability distributions of the target attribute's values [6]. The gain ratio [7] is a measure that "normalizes" the information gain divided by the entropy, and the distance measurements [18] differs from other measures in that they use the distance to normalize impurity measurements. However, to the best of my knowledge, there is no method to determine the splitting criteria considering the distance between classes among representative impurity-based splitting criteria.

DT methods have been developed over the years, e.g., ID3 [5], CART [6], C4.5 [7], and CHAID [8]. A detailed review of the structure of DT and the developed methods was provided by Murthy [9]. The included methods differs in the data type of the dependent variable, with impurity measures used to select variables and division boundaries. A summary of the decision tree algorithms is provided in Table 1. Of those, CART adopts a statistical approach using a statistical index called the Gini index. The Gini index is a measure of the degree or probability of samples being incorrectly classified when it has been randomly chosen. ID3 and C4.5 use entropy and CHAID uses chi-square as impurity measures.

Random forest (RF) is an ensemble method that randomly trains many decision trees. DT-based ensemble methods have been used effectively in various domains. DTs can produce results with a large variation and a high variance, which leads to inconsistency and overfitting. Random forest methods alleviate some of these shortcomings by constructing many trees with various properties (See Figure 3 for the "bagging" construction). Such methods reduce the variance because bagging aggregates results of multiple models. Random forest methods adopt random feature selection, and when modeling each subtree, only some features are randomly used for all of the subtrees to then be aggregated by average or vote. This process prevents the use of only specific features and reduces the correlation between subtrees. As a result, the random forest model is robust against noise.

To construct various trees, random forest generates subsets of size D to use a bootstrap technique, which is a sampling method that samples the dataset with a replacement. Then, RF uses subsets to generate a classifier of size D and then combines the generated classifiers into one classifier as:

$$p(c|x) = \frac{1}{D}\sum_{i=1}^{D} p_i(c|x)$$

TABLE 1. Comparison of the four most used decision tree methods [22].

	CART	ID3	C4.5	CHAID
Measurement index	Gini index / Twoing criteria	Entropy	Entropy info-gain	Chi-square
Dependent Variable	Categorical / Continuous	Categorical	Categorical / Continuous	Categorical / Continuous
Input Variables	Categorical / Continuous	Categorical / Continuous	Categorical / Continuous	Categorical / Continuous
Division Space	Binary	Multiple	Multiple	Multiple

CART: Classification and Regression Trees, ID3: Iterative Dichotomiser 3, C4.5: An extension of ID3, CHAID: Chi-square Automatic Interaction Detector

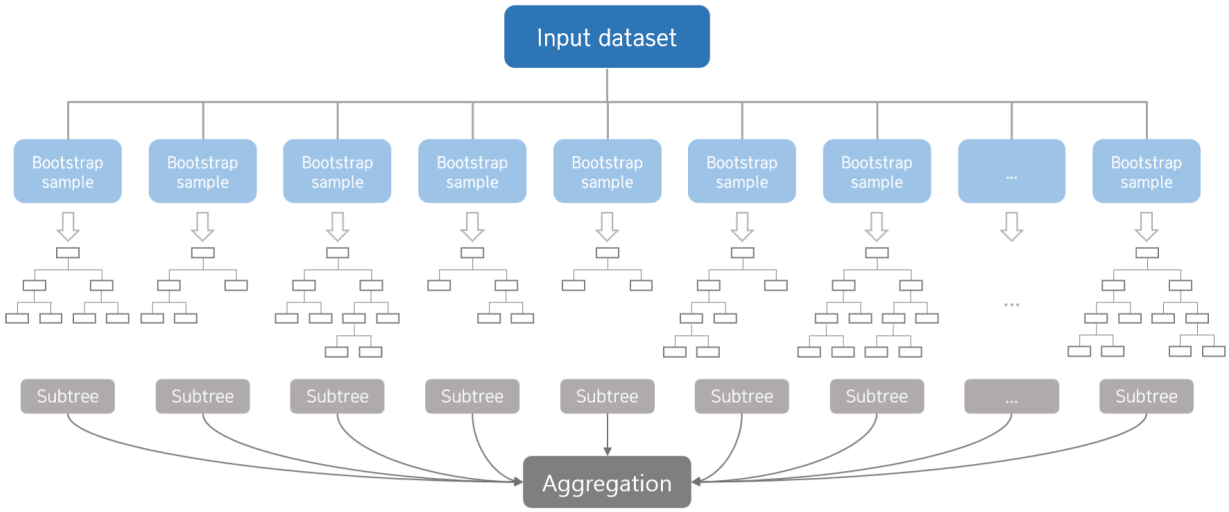

FIGURE 3. Illustration of the bagging algorithm.

where $p(c|x)$ is the probability of class c when x is given. This structure is called bagging (bootstrap aggregating).

III. RAO-STIRLING BASED CLASSIFICATION TREE (RSCT)

A. RAO-STIRLING BASED IMPURITY

In this section, we propose a new decision tree algorithm that considers not only the impurity but also the distance between two classes. First, we present a new impurity measure, the Rao-Stirling measure, that considers not only the impurity but also the distances between classes. The Rao-Stirling measure is one of a family of diversity measures used to consider distances between fields, and it is extensively used in interdisciplinary research [1]–[3].

The proposed Rao-Stirling measure multiplies the Gini index, a representative measure of impurity in decision trees, by the distance between the classes. The Rao-Stirling based impurity is defined as:

$$\sum_{i=1}^{M} \sum_{j=1(i\neq j)}^{M} p_i p_j d_{ij}$$

where, p_i and p_j denote the probabilities for the i-th and j-th classes, respectively. d_{ij} denotes the distance between the i-th class and j-th class.

In the Rao-Stirling measure, the splitting boundary is determined to classify samples with similar classes into the same partition. A simple example shown in Figure 1 can also be split using the Rao-Striling based impurity. The decrease in the Gini index after the partition in Figure 1(a)-(c) has the same value of 0.1667 while the decrease in the Rao-Stirling based impurity isn't the same; (a) 0.257, (b) 0.301, (c) 0.8546. In other words, the Rao-Stirling based impurity approach preferentially determine the splitting boundary so that observations of similar classes can be gathered.

To calculate the distance between the classes, we use the class's center point. When a data set is given $X \in \mathbb{R}^{n\times p}$, n is the number of data and p is the number of features. The center point is defined as the average value of the features for each class as follows:

$$CP_i = [\bar{X}_1^i, \bar{X}_2^i, \ldots, \bar{X}_p^i]$$

where $\bar{X}_p^i$ is average value of each feature of i-th class.

TABLE 2. Experimental data.

Dataset	Variable Type	Number of Classes	Number of Instances	Number of Attributes
CMC	Mixed	3	1,473	9
Car	Categorical	4	1,728	6
Yeast	Real	10	1,484	8

First, the center points (the average value of the variables in each class) are calculated and then, using the obtained center points of each class, the distances between the classes are calculated.

In this study, we use the Euclidean distance and cosine distance as the measure of the distance between classes.

Euclidean distance d_{ij} is calculated as:

$$d(CP_i,\ CP_j) = ||CP_i - CP_j|| = \sqrt{\sum_{p=1}^{P} (\bar{X}_p^i - \bar{X}_p^j)^2}$$

where the Euclidean distance d_{ij} $(for\ i,\ j = 1, \ldots, M)$ has the range, $-\infty \leq d_{ij} \leq \infty$.

The cosine distance d_{ij} is calculated as:

$$d_{ij} = 1 - s_{ij}$$

where s_{ij} is the cosine similarity. The cosine similarity can be applied to any number of dimensions and is used to measure cohesion within clusters. The cosine similarity is defined as:

$$s_{ij} = \frac{CP_i \cdot CP_j}{||CP_i||||CP_j||}$$

The value of the cosine similarity $s_{ij}(for\ i,\ j = 1,\ \ldots\ , M)$ has a range from -1 to 1. -1 means exactly the opposite, 1 is exactly the same, and 0 indicates orthogonally between the two classes. The lower the degree of similarity, the farther the distance between the two classes. So, we use the cosine distance instead of the cosine similarity. The cosine distance has a range from 0 to 2. With a closer distance between the classes, the value approaches 0, and the farther the distance, the closer to 2. The cosine distance has a similar trait with the Euclidean distance, but it does not have the triangle inequality property, that is, $d_{ij} \leq d_{il} + d_{lj}$.

B. PROCEDURE

This section introduces the overall process of the proposed method, with the process map described in Figure 4. First, Steps (2)–(3) in Figure 4 comprise the preprocessing. Steps (2) and (3) find the central value of each class and calculate the distance between the classes based on the center points. Then, Steps (4) calculate the Rao-Stirling measure for all candidates of the splitting boundary to divide the data into sub data. Finally, in Steps (5)–(6), the goodness of split is calculated for each candidate, and the best division boundary is determined with the largest goodness of split. Then, if the goodness of split is positive, the procedure repeats the steps from (4). Finally, the procedure stops when the goodness of split is zero.

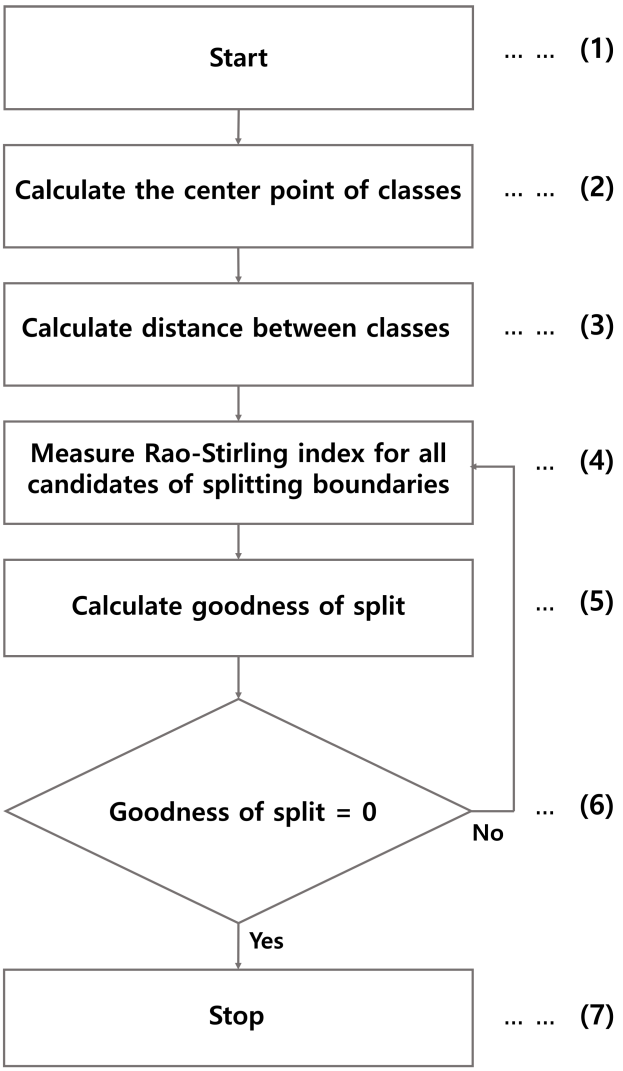

FIGURE 4. Process of the Rao-Stirling measure-based classification tree (RSCT).

IV. EXPERIMENTAL RESULTS

Computational experiments were conducted using three data sets to compare the proposed Rao-Stirling based model with the Gini based model. In addition, two methods were compared to assess their results in measuring the class distance.

A. DATASETS

The datasets used for the computational experiments include the contraceptive method choice, car evaluation, and yeast datasets obtained from the UCI repository for machine learning databases. All data have positive values without missing values. The characteristics of the three datasets

TABLE 3. Classification results in terms of accuracy for each dataset.

	Measure	CMC	Car	Yeast
Decision tree	Gini	55.9	76.5	56.5
	Rao-Stirling (Euclidean)	**57.8**	76.8	**58.2**
	Rao-Stirling (Cosine)	56.9	**77.0**	57.7
Random forest	Gini	56.0	76.1	56.6
	Rao-Stirling (Euclidean)	**57.2**	77.6	**58.2**
	Rao-Stirling (Cosine)	57.0	**77.8**	57.9

* The best and second best results of each dataset are marked in bold.

used for the experiments are summarized in Table 2. The contraceptive method choice (CMC) dataset is part of the 1987 National Indonesia Contraceptive Prevalence Survey data. The CMC is used to classify women's current contraceptive methods based on demographic and socio-economic characteristics. The dataset contains nine attributes with categorical and integer types and 1,473 observations with three classes. The car evaluation dataset was donated by Marco Bohanec in 1997. It contains six categorical attributes and 1,728 instances with four classes. The task of this dataset is to classify whether the car is good or not with variables. The yeast data contains 1,484 instances with ten classes and eight attributes with real values. The dataset is used to predict the localization of cellular components consisting of proteins in a yeast cell.

B. RESULTS

The performance of the proposed method is evaluated by comparing the results to those of the existing method in terms of accuracy. For Gini-based DT and the proposed Rao-Stirling based DT, the max depth d was varied such that $d = 1, 2, 3, \ldots, 12$, and pruning was conducted. For Gini-based RF and the proposed Rao-Stirling based RF, the number of trees n was varied such that $n = 300, 310, 320, \ldots, 500$, and the max depth d was varied in the same way as the DTs. For each combination of parameters, 10-fold cross-validation was performed. The best mean classification accuracy over 10 cross-validations are summarized in Table 3.

For the CMC dataset, the best accuracy was achieved when using the Rao-Stirling with Euclidean distance and 5 max depths, regardless of approach taken (i.e., decision tree and random forest). For the car evaluation dataset, Rao-Stirling with a cosine distance of 8 max depths provides the best accuracy, regardless of approach taken. The best accuracy in the yeast dataset was obtained when using Rao-Stirling (Euclidean) with 6 max depths in both decision tree and random forest. The results show that the proposed method considering the distance between classes offers competitive performance. From Table 3, we observe the proposed methods considering class distances are superior to existing methods, regardless of the class distance measuring methods and approach taken (decision tree vs. random forest).

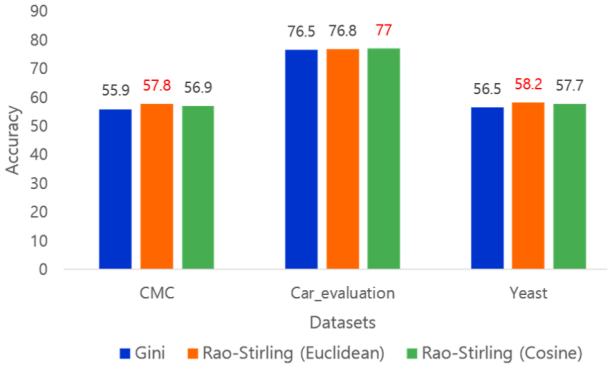

FIGURE 5. Accuracy plots of the proposed Rao-Stirling based DT and the existing DT for CMC, Car Evaluation and Yeast datasets.

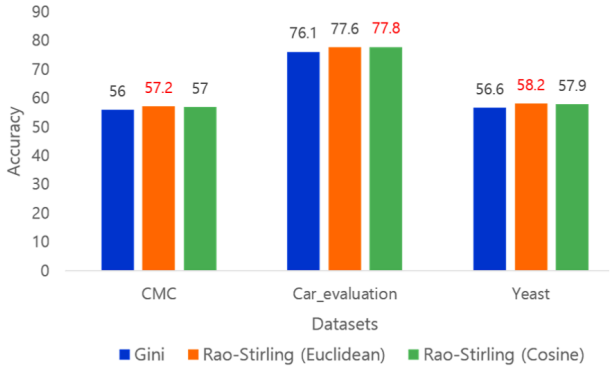

FIGURE 6. Accuracy plots of the proposed Rao-Stirling based RF and the existing RF for CMC, Car Evaluation and Yeast datasets.

In Figure 5, for the CMC dataset, the RSCT with Euclidean class distance provides 1.9 percent higher performance than Gini index based DT. For the car evaluation dataset, the RSCT with cosine class distance shows 0.5 percent higher performance than the existing Gini based DT. For the yeast dataset, the DT using the Rao-Stirling measure and Euclidean class distance yielded a 1.7 percent higher performance than the Gini-based DT.

Figure 6 shows the classification accuracies for the random forest methods. The RF methods that consider class distances are more accurate than RF methods that do not consider class distances. Especially, for the CMC dataset, the Rao-Stirling based RF with Euclidean class distance provides 1.2 percent

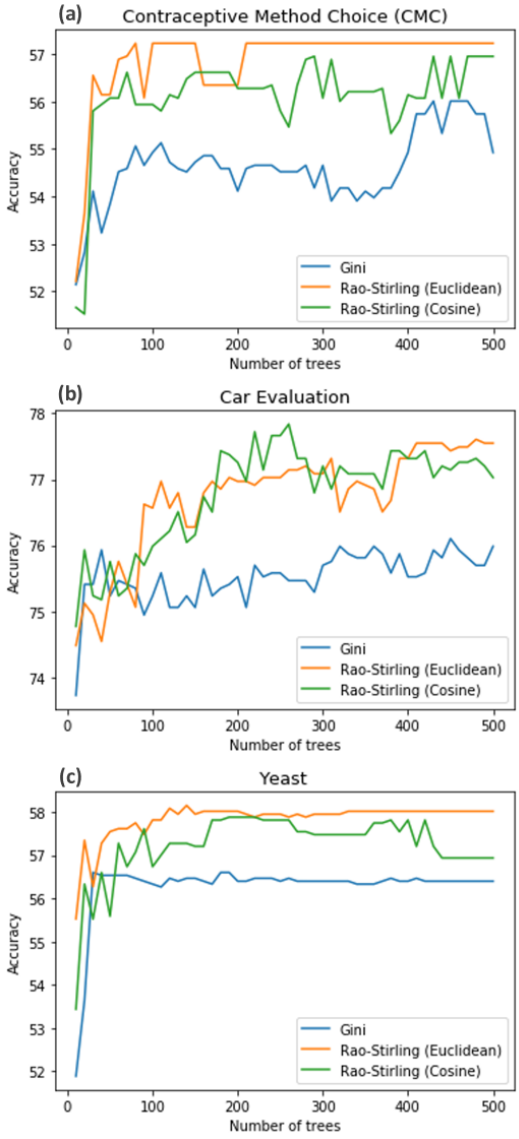

FIGURE 7. Accuracy vs number of trees for three data sets. (a) CMC, (b) Car Evaluation, (c) Yeast.

higher performance than the Gini index-based RF. For the car evaluation dataset, the Rao-Stirling based RF with cosine class distance shows 0.7 percent higher performance than the existing Gini-based RF. For the yeast dataset, the RF using the Rao-Stirling measure and Euclidean class distance yielded a 1.6 percent higher performance than the Gini index-based RF.

Figures 7(a)–(c) show how accuracies of RF change with respect to the number of trees n for the CMC, car evaluation, and yeast data sets, respectively. Note that accuracies start to stabilize when n is approximately equal to 100 for all three data sets, regardless of the datasets. If number of trees is greater than 100, the method that takes the distance of the class into account performs better than the method that does not consider the distance at all.

The Rao-Stirling based impurity approach determines splitting boundaries to classify samples with similar classes into the same partition. Therefore, in the Rao-Stirling based model, similar samples will be classified in a partition, and the distance between samples in a partition will be close.

TABLE 4. Change of sum of distances between samples over all leaf nodes with respect to decision tree depth.

		CMC		CAR	
		GINI	RAO - STIRLING	GINI	RAO - STIRLING
DEPTH	0	4,377,910	4,377,910	8,596,448	8,596,448
	1	3,091,785	3,069,390	3,801,345	3,771,805
	2	1,526,069	1,524,780	2,534,828	2,532,104
	3	810,353	798,518	1,672,261	1,664,332
	4	443,859	434,606	1,272,855	1,268,665

To assess the effects of considering class distances in the Rao-Stirling based model, the change of sum of the distances between the samples over all leaf nodes with respect to decision tree depth for both approaches (i.e., Gini based model and Rao-Stirling based model) is investigated in Table 4. Table 4 shows that for CMC and car evaluation datasets, the sum of distances between samples over all leaf nodes in the Rao-Stirling based model is less than that in the Gini based model regardless of decision tree depth. The results confirm that the Rao-Stirling based impurity approach, taking into account the class distances, allows similar data to belong to the same class.

V. CONCLUSION

A decision tree method based on the Rao-Stirling measure was developed to consider the distance between classes, and it is compared to existing DT methods using the CMC, car evaluation and yeast datasets. While the existing Gini index used in existing DT methods only considers the impurity, the Rao-Stirling measure considers the class distances as well as the impurity. The Experimental results show that the proposed approach performs consistently better than existing approaches, regardless of the data set and class distance measures employed. This is an encouraging result since only considering the distances between classes can improve the performance in both decision tree and random forest analyses. Future work may include verifying the above findings using various datasets and considering other class distance measurements.

ACKNOWLEDGMENT

(Sangyong Lee and Chulhee Lee are co-first authors.)

REFERENCES

[1] L. Leydesdorff and I. Rafols, "Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations," *J. Informetrics*, vol. 5, no. 1, pp. 87–100, Jan. 2011.

[2] L. Leydesdorff, I. Rafols, and C. Chen, "Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal-journal citations," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, no. 12, pp. 2573–2586, Dec. 2013.

[3] L. Leydesdorff, "Can technology life-cycles be indicated by diversity in patent classifications? The crucial role of variety," *Scientometrics*, vol. 105, no. 3, pp. 1441–1451, Dec. 2015.
[4] F. Takahashi and S. Abe, "Decision-tree-based multiclass support vector machines," in *Proc. 9th Int. Conf. Neural Inf. Process. (ICONIP)*, Nov. 2002, pp. 1418–1422.
[5] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[6] L. Breiman, J. H. Friedman, and R. A. C. J. Stone, *Classification and Tree Regression*. Pacific Grove, CA, USA: Wadsworth and Brooks/Cole, Monterey, 1984.
[7] J. R. Quinlan, *C4.5: Programming for Machine Learning*. Burlington, MA, USA: Morgan Kauffmann, 1993.
[8] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Appl. Statist.*, vol. 29, no. 2, pp. 119–127, 1980.
[9] S. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey," *Data Mining Knowl. Discovery*, vol. 2, pp. 345–389, Mar. 2000.
[10] A. Stirling, "A general framework for analysing diversity in science, technology and society," *J. Roy. Soc. Interf.*, vol. 4, no. 5, pp. 707–719, 2007.
[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[12] Z. Botta-Dukát, "Rao's quadratic entropy as a measure of functional diversity based on multiple traits," *J. Vegetation Sci.*, vol. 16, no. 5, pp. 533–540, Oct. 2005.
[13] M. Park, I. Weber, M. Naaman, and S. Vieweg, "Understanding musical diversity via online social media," in *Proc. 9th Int. AAAI Conf. Web Social Media*, 2015, pp. 308–317.
[14] L. Leydesdorff, D. Kushnir, and I. Rafols, "Interactive overlay maps for U.S. patent (USPTO) data based on international patent classification (IPC), *Scientometrics*, vol. 98, no. 3, pp. 1583–1599, 2014.
[15] P. Geurts, A. Irrthum, and L. Wehenkel, "Supervised learning with decision tree-based methods in computational and systems biology, Molecular Biosystems, vol. 5, no. 12, pp. 1593–1605, 2009.
[16] S. Chakrabartty and G. Cauwenberghs, "Gini support vector machine: Quadratic entropy based robust multi-class probability regression," *J. Mach. Learn. Res.*, vol. 8, pp. 813–839, Apr. 2007.
[17] S. A. Mulay, P. R. Devale, and G. V. Garje, "Decision tree based support vector machine for intrusion detection," in *Proc. Int. Conf. Netw. Inf. Technol.*, Jun. 2010, pp. 59–63.
[18] R. L. D. Mantaras, "A distance-based attribute selection measure for decision tree induction," *Mach. Learn.*, vol. 6, no. 1, pp. 81–92, 1991.
[19] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers—A survey," *IEEE Trans. Syst., Man Cybern., C, Appl. Rev.*, vol. 35, no. 4, pp. 476–487, Nov. 2005.
[20] J. R. Quinlan, "Simplifying decision trees," *Int. J. Man-Mach. Stud.*, vol. 27, no. 3, pp. 221–234, Sep. 1987.
[21] L. Ceriani and P. Verme, "The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini," *J. Econ. Inequality*, vol. 10, no. 3, pp. 421–443, 2012.
[22] Y.-Y. Song and L. Ying, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, 2015.

SANGYONG LEE received the B.S. and M.S. degrees from the Department of Industrial and Management Engineering, Myongji University, South Korea, in 2018 and 2020, respectively. His research interests include statistical data mining and deep learning.

CHULHEE LEE received the B.S. and M.S. degrees from the Department of Industrial and Management Engineering, Myongji University, South Korea, in 2018 and 2020, respectively. His research interests include machine learning and deep learning.

KWON GI MUN received the M.A. degree in economics from the University of Missouri, Columbia, and the joint Ph.D. degree in operations research and business and supply chain management from Rutgers University. He was a full-time Staff at Korean Chamber of Commerce and Industry, Seoul, South Korea. He is currently an Assistant Professor with the Department of Technology and Operations Management, California State Polytechnic University, Pomona. He has experience working with industry and government partners, especially in energy policy/system modeling using data analytics. His research interests include operations and energy interface, as well as supply chain strategies.

DOHYUN KIM received the M.S. and Ph.D. degrees from the Department of Industrial Engineering, Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2002 and 2007, respectively. He is currently an Associate Professor with the Department of Industrial and Management Engineering, Myongji University. His research interests include statistical data mining, deep learning, and graph data analysis.