Expert Systems With Applications 105 (2018) 112–128

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Region compatibility based stability assessment for decision trees

Lihong Wang[a], Qiang Li[b], Yanwei Yu[a,c,*], Jinglei Liu[a]

[a] School of Computer and Control Engineering, Yantai University, Yantai, Shandong 264005, China
[b] School of Economics and Management, Yantai University, Yantai, Shandong 264005, China
[c] College of Information Science and Technology, Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 31 May 2017
Revised 19 March 2018
Accepted 20 March 2018
Available online 23 March 2018

Keywords:
Machine learning
Decision tree
Stability measurement
Region compatibility
Evidence theory

ABSTRACT

Decision tree learning algorithms are known to be unstable, because small changes in the training data can result in highly different decision trees. An important issue is how to quantify decision tree stability. Two types of stability are defined in the literature: structural and semantic stability. However, existing structural stability measures are meaningless when applied to apparently different decision trees, and semantic stability only focuses on prediction accuracy without considering structural information. This paper proposes a region compatibility based structural stability measure for decision trees that considers the structural distribution of leaves from the view of basic probability assignments in evidence theory. To the best of our knowledge, we are the first to use basic probability assignments to quantify decision tree stability. We prove convergence for region compatibility, and show that apparently different decision trees have some inherent similarity from the view of region compatibility. We also clarify the meaning of region compatibility for measuring decision tree stability, and derive a method to select a relatively stable learning algorithm for a given dataset. Experimental results validate that region compatibility is effective to quantify the stability of decision tree learning algorithms.

1. Introduction

Knowledge representation and knowledge acquisition are both essential components to design and maintain expert systems. Many knowledge representation methods have been proposed, ranging from production rules, first order logic to semantic networks. Production rules are widely adopted as basic representations in rule based expert systems, automated planning, and decision making, due to their modularity and easy interpretation. Decision tree is a frequently used model for knowledge acquisition from a given dataset, and can be transformed into an equivalent set of rules. A stable decision tree would provide credible rules, hence decision tree learning algorithms have been widely studied.

Decision tree learning is widely used for classification in machine learning, and efficiently infers a tree-like classifier model from a labeled dataset (Quinlan, 1986; Rokach & Maimon, 2005). A decision tree has clear structure and semantic interpretation, which makes it an attractive model for supervised learning. Consequently, decision trees and variations have been employed for many applications, such as credit scoring (Xia, Liu, Li, & Liu, 2017), nominal data clustering (Ghattas, Michel, & Boyer, 2017), and subspace partitioning (Kim, 2016). However, decision tree learners are highly unstable, producing significantly different classifiers from slightly different training sets (Dwyer & Holte, 2007; Turney, 1995) due to the large number of candidate variables with similar discrimination power, from which only a few are selected (Aluja-Banet & Nafria, 2003). In Dwyer and Holte (2007), Dwyer appended a single instance to the training dataset, and showed that C4.5 produced a substantially different decision tree, explicitly demonstrating instability. More specifically, these two trees were produced by C4.5 using data from the lymphography dataset, which was obtained from the UCI repository. T_{106} was induced from a random sample with 106 examples. A single instance, randomly chosen from the unused examples, was appended to this training set, from which C4.5 produced the tree T_{107}. T_{107} contained nearly double the number of decision nodes appearing in T_{106}. This instability undermines the objective of extracting knowledge, and raises suspicion about the validity of the decision tree: is the output model an integration of the information extracted from the training data or is it an artifact reacting to the training instances? (Mirzamomen & Kangavari, 2016). In the context of active learning (Dwyer & Holte, 2007) or incremental

* Corresponding author at: School of Computer and Control Engineering, Yantai University, Yantai, Shandong 264005, China.
E-mail addresses: wanglh@ytu.edu.cn (L. Wang), lq130@163.com (Q. Li), yuyanwei@ytu.edu.cn, yuy174@ist.psu.edu (Y. Yu), jinglei_liu@sina.com (J. Liu).

https://doi.org/10.1016/j.eswa.2018.03.036
0957-4174/

learning (Kalles & Papagelis, 2000), stability problems become more important when an induction algorithm must revise a decision tree.

An ensemble of decision trees, e.g. random forest (Breiman, 2001), usually provides a stable prediction for new instances (Parvin, MirnabiBaboli, & Alinejad-Rokny, 2015; Yang, Roe, & Zhu, 2007). Kuncheva et al. argued that diversity is a key issue for classifier ensembles, and although the general motivation for designing diverse classifiers is correct, the problem of measuring this diversity and using it effectively to build better classifier teams remains open (Kuncheva & Whitaker, 2003). In the ensemble of decision trees, component decision tree diversity can also help to generate smaller ensembles with stronger generalization ability (Banfield, Hall, Bowyer, & Kegelmeyer, 2007; Zhou, Wu, & Tang, 2002). Thus, the proposed decision trees stability metric studied in this paper can also serve as a decision tree diversity metric.

An important issue is how to quantify stability. Two types of stability are defined in the literature: semantic and structural stability (Dwyer & Holte, 2007; Mirzamomen & Kangavari, 2016). Semantic stability measures the degree to which two classifiers make the same predictions, whereas structural stability measures the similarity between particular structural properties of two trees. Structural stability is a sufficient condition for semantic stability, since structurally similar decision trees will produce the same predictions, but the converse is not true.

Structural stability is meaningful when decision trees are identical or partially identical, but it is meaningless otherwise. Semantic stability measures the prediction results of given instances without considering structural information. Therefore, structural and semantic stability are somewhat complimentary metrics, and it would be appropriate to measure decision tree stability considering structural and semantic stability simultaneously.

This paper introduces a structural method, region compatibility, to quantify stability. Region compatibility quantifies the similarity of two trees even if they are apparently different. Although it is defined based on a structural term, called a region, it also considers semantic factors because region compatibility directly compares region instance sets between two trees. Extensive experiments demonstrate the proposed metric effectiveness to quantify decision tree stability. The evaluations also suggest the proposed metric helps to identify stable decision trees that can be transformed to be a set of credible expert system rules. Therefore, the proposed metric constitutes a knowledge evaluation method for expert and intelligent systems, and has significant impact on knowledge acquisition and maintenance.

The main contributions of this paper are as follows:

1. The region compatibility metric is proposed to evaluate decision tree structural stability, even where the trees are apparently different. To the best of our knowledge, this is the first use of basic probability assignments (BPAs) in evidence theory to quantify decision tree stability.
2. Region compatibility advantages over existing metrics are discussed. Since it explicitly considers leaf structural distribution, region compatibility provides more subtle comparison between two decision trees than current structural stability metrics.
3. Region compatibility convergence is proved and validated. Experimental results show an interesting region compatibility convergence property for a special case, providing new insights to understand decision tree stability.
4. Three well-known decision tree learning algorithms are compared, and the algorithm with lowest region compatibility is expected to induce relatively stable decision trees and derive credible rules for a given dataset.

The remainder of the paper is organized as follows. Section 2 reviews previous studies related to decision tree stability. Section 3 presents region compatibility, the proposed decision tree stability metric, and provides a theoretical explanation of its properties. Section 4 presents experiments on UCI datasets to evaluate decision tree region compatibility. Finally, Section 5 summarizes and concludes the paper, and discusses some future research directions.

2. Decision tree stability

2.1. Decision trees

Decision tree induction offers a highly practical method for supervised learning. A decision tree is a directed tree consisting of internal nodes and leaves. The most common approach is to partition labeled examples recursively until a stop criterion is met. Generally, an internal node is created and assigned with a test that has a small set of outcomes, and then a branch for each possible outcome is created, and each example is passed down the corresponding branch. Each partition block is treated as a sub-problem, for which a sub-tree is built recursively. The root of the directed tree is a special internal node, and all non-internal nodes are called leaves (Quinlan, 1986; Rokach & Maimon, 2005).

2.2. Semantic stability

Turney proposed a method to quantify decision tree semantic stability based on agreement between trees built on samples from the same distribution (Turney, 1995). The agreement metric was defined as the probability that a randomly chosen unlabeled example was assigned to the same class by both trees. This approach has subsequently been widely adopted as the decision tree semantic stability metric (Dwyer & Holte, 2007). In practice, the agreement is generally estimated by classifying a randomly selected set of instances, and calculating the ratio of the set assigned to the same class by both trees.

Paul et al. proposed class prediction stability as an indicator of semantic stability for classification algorithms (Paul, Verleysen, & Dupont, 2012). Class prediction stability measures the extent each individual test example is assigned to the same class label across various re-samplings. Stability = 1 when every test example is always assigned the same class label, although this may not necessarily be correct.

2.3. Structural stability

Syntactic similarity measures are not suitable to measuring decision tree similarity, although it seems intuitive. The main syntactic metric drawback is that they are heavily dependent on the chosen representation, and it is difficult to compare stability across different representations. Syntactic similarity metrics can also consider logically equivalent trees as different, because the edit distance between two logically equivalent trees can be significant (Turney, 1995) (refer to Fig. 1 in Section 3.1).

Dwyer appended a single instance to the training dataset, and showed that **C4.5** produced a decision tree containing nearly double the number of nodes compared to before appending that instance, explicitly demonstrating instability (Dwyer & Holte, 2007). Thus smaller difference in tree size and depth have been considered good indicators for structural stability (Mirzamomen & Kangavari, 2016; Zimmermann, 2008). Dwyer also defined a region stability metric of decision tree structural stability (Dwyer & Holte, 2007), where each decision tree leaf was defined as a decision region, and region stability measured the difference between

Table 1
Metric comparison.

Metric	Description	Type	Data usage	Ensemble
Agt	Agreement	Semantic	Training+testing	No
GE	Generalization error	Semantic	Training+testing	Yes
CPS	Class prediction stability	Semantic	Training+testing	Yes
ReSt	Region stability	Structural	Training	No
SD	Tree size and depth	Structural	Training	Yes
VC	Variable selection and Cut-point stability	Structural	Training	No
RC	Region compatibility	Structural	Training+testing	No

Agt: (Turney, 1995), *GE*: (Breiman, 2001), *CPS*: (Paul et al., 2012), *ReSt*: (Dwyer & Holte, 2007), *SD*: (Zimmermann, 2008), *VC*: (Briand et al., 2009), *RC*: this paper.

decision regions in two trees. In particular, Dwyers proposed metric estimates the probability that two trees classify a randomly selected instance into equivalent decision regions.

Briand proposed a similarity metric based on variable selection and cut-point stability associated with internal decision tree nodes to evaluate decision tree stability (Briand, Ducharme, Parache, & Mercat-Rommens, 2009). However, it is very complex to compare two trees node by node.

We compare these stability metrics from three aspects.

1. Type. Almost all semantic stability metrics focus on prediction accuracy of the test set, regardless of the particular prediction error or agreement. On the other hand, structural stability focuses on various decision tree structural aspects, including tree size (i.e., total number of the nodes), depth (i.e., length of the longest branch), leaf intension (i.e., logical formulas for the paths from root node to leaves), to measure similarity for variable selections and cut-point stability with decision tree internal nodes.
2. Data usage. The training set is used to build a decision tree, and the test set to assess semantic stability. Once the tree is constructed, the nodes, size, depth, and leaf intensions are fixed to test new instances. Hence, structural stability assessments are independent of the test set. However, the proposed region compatibility measure applies to both training and test sets, because it also considers semantic factors.
3. Ensemble. Stability indicators are used to test ensemble or single decision tree stability. Table 1 summarizes comparison of seven metrics. Approximately half the approaches assess ensemble stability; i.e., prediction accuracy improvement when a new decision tree is added to the ensemble for the semantic metric, or the accumulated number of nodes of all trees in the ensemble for the structural metric.

3. Proposed region compatibility

3.1. Decision region and two indicators

Dwyer defined region stability (Dwyer & Holte, 2007) with each decision tree leaf corresponding to a decision region. A decision region was then determined by the path from the root to the leaf, i.e., the set of nodes and branches on the path, and decision regions were considered equivalent if they performed the same set of tests and predicted the same class label for the tests.

Given a labeled dataset, D, a randomly selected subset of D is used as training set to produce a decision tree, T, using an induction algorithm, such as C4.5 or CART. This paper does not adopt a specific induction algorithm, i.e., any decision tree learning algorithm is compatible with the proposed stability evaluation metric. We assume that the whole of D is tested by T, i.e., each instance of D is classified exactly by one leaf of T, and the set of instances falling in each leaf forms a decision region.

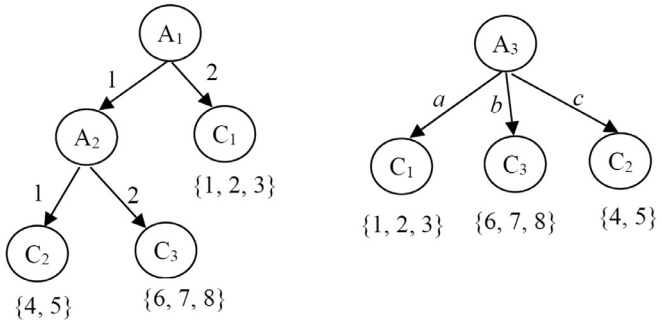


Fig. 1. Apparently different decision trees with identical decision regions.

Let us look closer at Dwyer's decision region definition.

1. A one-to-one mapping exists from the set of leaves to the set of decision regions, i.e., instances falling in a leaf compose its corresponding decision region.
2. Each decision region, R, has a class label, $Label(R)$, determined by majority voting by instance labels in R.
3. Each decision region has an intension and an extension. The intension of a decision region is a logical formula, that is, conjunctions of all tests on the path from the root to the leaf, while the extension of a decision region is the set of instances classified to the leaf. As far as the dataset D is concerned, all instances are partitioned into different decision regions with labels.

Thus, we formally define a decision region as follows.

Definition 1. The decision region, R, of a leaf in decision tree, T, is defined as

$$R(leaf) = \{x \in D | \bigwedge_{t_i \in path(leaf)} t_i(x)\}, \quad (1)$$

where $\wedge$ denotes the conjunction operation, $path(leaf)$ is the set of tests on the path from the root to the leaf, and test $t_i(x)$ is true if x satisfies $t_i(x)$. For example, if $t_i(x)$ is the test "age(x) $<$ 20", then $t_i(Alice)$ is true if she is younger than 20.

The decision region can be used to define two indicators: region stability (Dwyer & Holte, 2007), and agreement (Turney, 1995).

In the following sections, we assume that two decision trees T_1 and T_2 are built from two randomly selected training sets drawn from D, and the whole D is tested by each tree respectively. We obtain the decision region set, F, for each tree,

$$F_1 = \{R_{1,1}, R_{1,2}, ..., R_{1,s}\}, \; F_2 = \{R_{2,1}, R_{2,2}, ..., R_{2,t}\}. \quad (2)$$

Dwyer did not formally define region stability (Dwyer & Holte, 2007), hence we define it as follows.

Definition 2. The region stability (*ReSt*) between T_1 and T_2 is defined as:

$$ReSt(T_1, T_2, D) = \frac{1}{|D|} \sum_{R_{1,i} \in F_1} |R_{1,i}| I(\exists R_{2,j} \in F_2, R_{1,i} = R_{2,j}), \quad (3)$$

where $I(p) = 1$ if p is true, 0 otherwise; and $|D|$ is the cardinality of D.

Intension and extension are both semantic notions, extension is relatively clear, but intension is harder to grasp (Turney, 1995). Thus, we compare the extension (rather than intension) of $R_{1,i}$ with that of $R_{2,j}$, because apparently different path descriptions may cover the same set of instances. Fig. 1 shows that the decision region {1, 2, 3} can be described as $A_1 = 2$ in the left tree, but $A_3 = a$ in the right tree. From the view of extension, the leaf sets of two trees in Fig. 1 are equal to each other, namely, the two decision trees are logically equal. However, the edit distance between two trees are not zero even if we omit the branch values, because converting the left tree into the right one, the edit script needs to delete one vertex (A_2) and substitute one vertex (i.e., replace A_1 with A_3). Therefore, the cost of this edit script turns out to be 1 + 1 = 2 (Pawlik & Augsten, 2016; Tai, 1979). This example also shows that syntactic similarity metrics can consider logically equivalent trees as different, and are not suitable to measuring decision tree similarity.

Agreement is a semantic indicator of decision tree stability, and is defined as the probability that a randomly chosen unlabelled instance is assigned to the same class by both trees (Turney, 1995). This paper estimates this agreement by testing the whole D. Therefore, we represent it from the view of decision regions.

Definition 3. Given two decision trees, T_1 and T_2, with $R_{1,i} \in F_1$ and $R_{2,j} \in F_2$, the agreement (*Agt*) with respect to test data D_1 is defined as

$$Agt(T_1, T_2, D_1) = \frac{1}{|D|} \Big| \Big\{ x \in D_1 | \exists R_{1,i} \exists R_{2,j} \big(x \in (R_{1,i} \cap R_{2,j}) \wedge Label(R_{1,i}) = Label(R_{2,j}) \big) \Big\} \Big|, \quad (4)$$

where $D_1 \subseteq D$, and $Label(R)$ is the majority voting label in region R, as above.

There are some commonalities between the definitions of region stability (*ReSt*) and agreement (*Agt*). On one hand, instances falling in the same region must be predicted with the same label, i.e., when two trees have an identical decision region, they have the consistent agreement in this region. On the other hand, region stability only considers exactly identical decision regions in two decision trees from the view of structural stability, whereas agreement includes all consistent instances with the same labels regardless of the decision region(s) the instances fall into, i.e., agreement is a semantic indicator that only considers the predicted labels of regions.

The following properties are derived from Definitions 2 to 3.

Theorem 1. *If T_1 and T_2 are identical decision trees, then $ReSt(T_1, T_2, D) = 1$ and $Agt(T_1, T_2, D) = 1$.*

Proof. Since T_1 and T_2 are identical decision trees, they have the same set of decision regions. Thus, $F_1 = F_2$ and each decision region $R_{1,i} \in F_1$ is also a decision region of T_2, i.e., $R_{1,i} \in F_2$, and $ReSt(T_1, T_2, D) = 1$ from Definition 2.

Consequently, $Label(R_{1,i})$ is the same in T_1 and T_2, hence $Agt(T_1, T_2, D) = 1$. □

Theorem 2. *Agreement can be calculated recursively if there is exactly one pair of identical decision regions between the trees. Suppose $R_{1,1}$ in T_1 is identical to $R_{2,1}$ in T_2, i.e., $R_{1,1} = R_{2,1}$, $F_1 \cap F_2 = \{R_{1,1}\}$, then*

$$ReSt(T_1, T_2, D) = |R_{1,1}|/|D|$$

and

$$Agt(T_1, T_2, D) = ReSt(T_1, T_2, D) + Agt(T_1, T_2, D - R_{1,1}).$$

Proof. If $F_1 \cap F_2 = \{R_{1,1}\}$, then the other decision regions in T_1 are not equivalent to those in T_2. From Eq. (3), $R_{1,1}$ and $R_{2,1}$ consist of only the pair such that $R_{1,i} = R_{2,j}$, hence $ReSt(T_1, T_2, D) = |R_{1,1}|/|D|$.

Decision region $R_{1,1}$ in T_1 contributes $|R_{1,1}|/|D|$ to *Agt*, and the remaining decision regions in T_1 and T_2 correspond to $Agt(T_1, T_2, D - R_{1,1})$ tested by the remaining dataset $D - R_{1,1}$. □

Corollary 1. *For two decision trees T_1 and T_2 with k pairs of identical decision regions between them, i.e., $R_{1,1} = R_{2,1}, \ldots, R_{1,k} = R_{2,k}$, $F_1 \cap F_2 = \{R_{1,1}, \ldots, R_{1,k}\}$ then*

$$ReSt(T_1, T_2, D) = \frac{1}{|D|} \sum_{i=1}^{k} |R_{1,i}|, \quad \text{and}$$

$$Agt(T_1, T_2, D) = ReSt(T_1, T_2, D) + Agt(T_1, T_2, D - R_{1,1} - R_{1,2} - \ldots - R_{1,k}).$$

Corollary 1 is intuitive, the proof is similar to that for Theorem 2.

Theorem 3. *For two decision trees T_1 and T_2, if no decision region in T_1 is identical to any decision regions in T_2, then $ReSt(T_1, T_2, D) = 0$.*

Proof. From Definition 2, if there no decision region in T_1 is identical to any decision regions in T_2, then for any decision region $R_{1,i} \in F_1$, there exists no $R_{2,j} \in F_2$ such that $R_{1,i} = R_{2,j}$, i.e., $\forall R_{2,j} \in F_2$, $R_{1,i} \neq R_{2,j}$, hence $ReSt(T_1, T_2, D) = 0$. □

Theorems 1–3 show that region stability and agreement are simple to calculate for identical decision trees or partial identical trees, and their values are meaningful. However, region stability is zero, and agreement loses structure comparison clarity and degenerates to simply semantics for apparently different trees.

3.2. Region compatibility based on evidence theory

We propose a *region compatibility* stability indicator based on distance metrics from evidence theory (Dempster–Shafer theory). The concept is to evaluate structural stability between two apparently different trees, because there may be common patterns with the decision regions of decision trees produced from the same D, even if the decision regions are not entirely equivalent.

Table 2 shows the notations used throughout this paper.

3.2.1. Introduction to evidence theory

Evidence theory, also referred to as Dempster–Shafer theory, is a general framework to represent and combine all available evidence from different sources, and is particularly useful in the fields of expert systems and information fusion. The theory was first introduced by Dempster in the context of statistical inference and later developed by Shafer into a general framework for modeling uncertainty. It can be regarded as a generalization of classical probability theory by assigning a basic probability to a subset A of the frame of discernment, D, which is distributed in some unknown manner among the elements of A (Yager, 1987).

Let D be a frame of discernment containing N distinct objects $x_i, i = 1, \ldots, N$. The power set of D (2^D), is the set of the 2^N subsets of D. A basic probability assignment (BPA) m is a mapping from 2^D to [0,1] satisfying $\sum_{A \subseteq D} m(A) = 1$ and $m(\phi) = 0$. A subset A of D is called a focal element if $m(A) > 0$ and we denote the

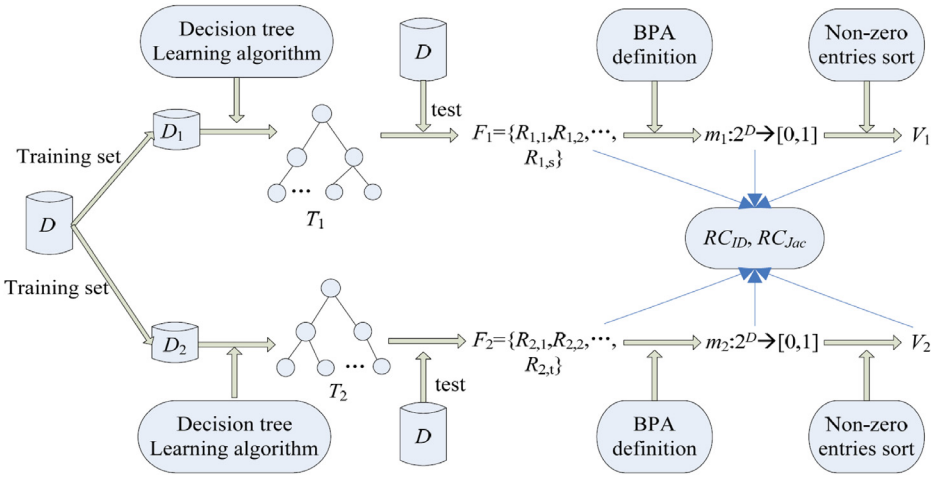


Fig. 2. Proposed region compatibility metric framework.

Table 2
Notations employed throughout this paper.

Notation	Description
$D = \{x_1, x_2, ..., x_N\}$	Data set including N instances.
$\|D\|$	Cardinality of D.
T_i	Decision tree.
R_i	Decision region.
$Label(R_i)$	Label for decision region R_i.
$F = \{R_1, R_2, ..., R_s\}$	Set of decision regions.
$ReSt(T_1, T_2, D)$	Region stability between T_1 and T_2.
$Agt(T_1, T_2, D)$	Agreement between T_1 and T_2.
m	Basic probability assignment, m is a mapping from 2^D to [0,1].
$V = [m(R_1), m(R_2), ..., m(R_s)]^T$	Non-zero entries in m sorted in the form of column vector.
$RC_{ID}(F_1, F_2)$	Region compatibility of F_1 and F_2 based on identity matrix.
$RC_{Jac}(F_1, F_2)$	Region compatibility of F_1 and F_2 based on Jaccard matrix.
W	Jaccard matrix.
W_{12}	Upper right corner block of W.

set of all the focal elements as $F = \{A \subseteq D | m(A) > 0\}$ (Jousselme & Maupin, 2012).

We may regard a decision tree dataset as a frame of discernment, hence each decision region corresponds to a focal element, and F is the set of all decision regions. We may use F to denote the set of all the focal elements as well as the set of all decision regions (Eq. (2)), because they have the same meaning in this paper.

We define BPA m: $2^D \to [0, 1]$ as follows.

$$m(R) = \begin{cases} |R|/|D| & if \quad R \in F \\ 0 & otherwise. \end{cases} \tag{5}$$

Since F is a partition of D, $\sum_{R \in F} m(R) = 1$. Therefore, Eq. (5) satisfies both BPA conditions.

3.2.2. Region compatibility

Fig. 2 shows the framework for the proposed region compatibility metric based on evidence theory. T_1 and T_2 are decision trees trained on randomly selected subsets D_1 and D_2 of D, respectively. Testing the all D by each tree produces decision regions F_1 and F_2, respectively. Then we use Eq. (5) to define m_1 and m_2 as the BPAs of F_1 and F_2, respectively. We can represent sorted non-zero entries in m_1 and m_2 in the form of column vectors. For example, let $V_1 = [m_1(R_{1,1}), m_1(R_{1,2}), \cdots, m_1(R_{1,s})]^T$, s.t. $0 < m_1(R_{1,1}) \le m_1(R_{1,2}) \le \cdots \le m_1(R_{1,s})$, $V_2 = [m_2(R_{2,1}), m_2(R_{2,2}), \cdots, m_2(R_{2,t})]^T$, s.t. $0 < m_2(R_{2,1}) \le m_2(R_{2,2}) \le \cdots \le m_2(R_{2,t})$ where $F_1 = \{R_{1,1}, R_{1,2}, \cdots, R_{1,s}\}$, $F_2 = \{R_{2,1}, R_{2,2}, \cdots, R_{2,t}\}$.

Since m_1: $2^D \to [0, 1]$, m_2: $2^D \to [0, 1]$, then $m_1 - m_2$ is a mapping from 2^D to $[-1, 1]$, i.e., $m_1 - m_2 : 2^D \to [-1, 1]$.

We define two region compatibility indicators for T_1 and T_2 based on distance metrics in the form of inner products (Jousselme & Maupin, 2012).

Definition 4. The region compatibility (RC) of decision trees T_1 and T_2 is defined as

$$\begin{aligned} RC_{ID}(F_1, F_2) &= \sqrt{(m_1 - m_2)^T (m_1 - m_2)} \\ RC_{Jac}(F_1, F_2) &= \sqrt{(m_1 - m_2)^T W (m_1 - m_2)}, \end{aligned} \tag{6}$$

where F_1 and F_2 are decision regions sets corresponding to T_1 and T_2, respectively; T indicates vector transposition; and W denotes the Jaccard matrix, with Jaccard index elements: $W_{A,B} = \frac{|A \cap B|}{|A \cup B|}$ $(A, B \in 2^D, W_{\emptyset,\emptyset} = 0)$.

The region compatibility indicators are both defined on m_1 and m_2, and each BPA is determined by F_1 and F_2, hence we choose F_1 and F_2 as the arguments of RC.

Each indicator in Definition 4 is based on the distance between two BPAs, so $RC = 0$ for two identical trees, and $RC \approx 0$ means the BPAs are only slightly different, i.e., the trees are similar in terms of the decision region extensions.

For example, suppose $D = \{1, 2, ..., 12\}$, with two decision region sets $F_1 = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9, 10, 11, 12\}\}$, and $F_2 = \{\{1, 2, 3\}, \{4, 7, 8, 12\}, \{5, 6, 9, 10, 11\}\}$.

Then, $m_1(\{1, 2, 3\}) - m_2(\{1, 2, 3\}) = 0.25 - 0.25 = 0$, and hence non-zero entries of $m_1 - m_2$ correspond to $\{4, 5, 6\}$, $\{7, 8, 9, 10, 11, 12\}$, $\{4, 7, 8, 12\}$ and $\{5, 6, 9, 10, 11\}$. For simplification, we omit the zero entries and represent $m_1 - m_2 = [0.25, 0.5, -0.33, -0.42]^T$. Hence,

$$\begin{aligned} &RC_{ID}(F_1, F_2) \\ &= \sqrt{(m_1 - m_2)^T (m_1 - m_2)} \\ &= \sqrt{[0.25, 0.5, -0.33, -0.42] \times [0.25, 0.5, -0.33, -0.42]^T} \\ &= \sqrt{0.5978} \\ &= 0.7732. \end{aligned}$$

The Jaccard matrix can be expressed as

$$W = \begin{bmatrix} 1 & 0 & 1/6 & 2/6 \\ 0 & 1 & 3/7 & 3/8 \\ 1/6 & 3/7 & 1 & 0 \\ 2/6 & 3/8 & 0 & 1 \end{bmatrix}.$$

For example, $W_{1,3}$ is the Jaccard index of {4, 5, 6} and {4, 7, 8, 12}, so

$$W_{1,3} = \frac{|\{4,5,6\} \cap \{4,7,8,12\}|}{|\{4,5,6\} \cup \{4,7,8,12\}|} = \frac{1}{6}.$$

Then,

$$\begin{aligned} &RC_{Jac}(F_1, F_2) \\ &= \sqrt{(m_1 - m_2)^T W (m_1 - m_2)} \\ &= \sqrt{[0.25, 0.5, -0.33, -0.42] \times W \times [0.25, 0.5, -0.33, -0.42]^T} \\ &= 0.4487. \end{aligned}$$

Bouchard et al. proved that W with elements that are Jaccard indexes of all pairs of subsets (excluding the empty set) of a reference frame D is positive definite, for any integer $|D| > 1$ (Bouchard, Jousselme, & Doré, 2013). Therefore, the associated distance properties are derived from W, and we propose the following Theorem.

Theorem 4. *If T_1 and T_2 are two decision trees, then $RC_{ID}(F_1, F_2) = 0$ if and only if $F_1 = F_2$ and $RC_{Jac}(F_1, F_2) = 0$ if and only if $F_1 = F_2$.*

Proof. Since the metric is positive definite, $RC_{ID}(F_1, F_2) = 0$ if and only if $m_1 = m_2$. Eq. (5) shows that $m_1 = m_2$ if and only if $F_1 = F_2$.

Similarly, we can prove the second proposition. □

We analyze several useful indicator properties, with proofs given in Appendix A.

Theorem 5. *Region compatibility can be calculated recursively if there is exactly one pair of identical decision regions between two trees. Assume $R_{1,1}$ in F_1 is identical to $R_{2,1}$ in F_2, i.e., $F_1 \cap F_2 = \{R_{1,1}\}$, then*

$$RC_{ID}(F_1, F_2) = RC_{ID}(F_1 - \{R_{1,1}\}, F_2 - \{R_{1,1}\})$$

and

$$RC_{Jac}(F_1, F_2) = RC_{Jac}(F_1 - \{R_{1,1}\}, F_2 - \{R_{1,1}\}).$$

Corollary 2. *For two decision trees, if there are k pairs of identical decision regions, i.e., $R_{1,1} = R_{2,1}, ..., R_{1,k} = R_{2,k}$, $F_1 \cap F_2 = \{R_{1,1}, ..., R_{1,k}\}$, then*

$$RC_{ID}(F_1, F_2) = RC_{ID}(F_1 - \{R_{1,1}, \cdots, R_{1,k}\}, F_2 - \{R_{1,1}, \cdots, R_{1,k}\}),$$

and

$$RC_{Jac}(F_1, F_2) = RC_{Jac}(F_1 - \{R_{1,1}, \cdots, R_{1,k}\}, F_2 - \{R_{1,1}, \cdots, R_{1,k}\}).$$

Theorem 6. *For two distinct decision trees, T_1 and T_2, if there is no decision region in T_1 identical to any decision region in T_2, i.e., $F_1 \cap F_2 = \phi$, then*

$$\begin{aligned} RC_{ID}(F_1, F_2) &= \frac{1}{|D|}\sqrt{\sum_i |R_{1,i}|^2 + \sum_j |R_{2,j}|^2}, \\ RC_{Jac}(F_1, F_2) &= \sqrt{\frac{1}{|D|^2}(\sum_i |R_{1,i}|^2 + \sum_j |R_{2,j}|^2) - 2V_1^T W_{12} V_2}, \end{aligned} \tag{7}$$

where W_{12} denotes a block of Jaccard matrix W; $W_{12}(A, B) = \frac{|A \cap B|}{|A \cup B|}, A \in F_1, B \in F_2$; and V_1 and V_2 are the column vectors of sorted non-zero entries in m_1 and m_2, respectively.

Theorem 7. *If $F_1 \cap F_2 = \phi$ and $V_1 = V_2$, then*

$$RC_{ID}(F_1, F_2) = \sqrt{2V_1^T V_1}.$$

For example, assume $F_1 = \{\{1,2\}, \{3\}, \{4,5,6\}, \{7\}, \{8\}\}, F_2 = \{\{1,3\}, \{2\}, \{5, 7, 8\}, \{4\}, \{6\}\}$, then $V_1 = V_2 = [0.125, 0.125, 0.125, 0.25, 0.375]^T$, and $RC_{ID}(F_1, F_2) = \sqrt{(m_1 - m_2)^T (m_1 - m_2)} = \sqrt{0.5} = \sqrt{2V_1^T V_1}$.

Theorem 6 represents RC_{Jac} as dependent on W_{12}, V_1, and V_2. If V_1 and V_2 are fixed, then the size of each focal element $R_{1,i}$ and $R_{2,j}$ in F_1 and F_2, respectively, is also fixed. However, $R_{1,i}$ and $R_{2,j}$ elements may be different, and RC_{Jac} would still change with W_{12}. Thus, we have the following theorem for the expectation of RC_{Jac}.

Theorem 8. *If $F_1 \cap F_2 = \phi$ and $V_1 = V_2$, then the expectation of RC_{Jac} is*

$$E(RC_{Jac}(F_1, F_2)) = \sqrt{2V_1^T V_1 - 2V_1^T E(W_{12}) V_2},$$

$$E(W_{12}(R_{1,i}, R_{2,j})) = \frac{n_i \times n_j}{N \times (n_i + n_j) - n_i \times n_j},$$

where

$$R_{1,i} \in F_1, R_{2,j} \in F_2, |R_{1,i}| = n_i, |R_{2,j}| = n_j, |D| = N.$$

The above theoretical result can be confirmed by empirical statistics based on the law of large numbers.

Theorem 9 (General weak law of large numbers). *Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed random variables, with $E(|X_1|) < \infty$ and write $E(X_1) = \mu$, and $S_n = X_1 + X_2 + ... + X_n$. Then for any $\varepsilon > 0$, $P(|\frac{S_n}{n} - \mu| > \varepsilon) \to 0$ as $n \to \infty$. (Meester, 2008).*

The general weak law of large numbers is useful here. If we test $RC_{Jac}(F_1, F_2)$ n times for a given dataset, then the average of $RC_{Jac}(F_1, F_2)$ tends to μ (the expectation of $RC_{Jac}(F_1, F_2)$ in Theorem 8) when n tends to infinity. We show how to obtain this expectation value in the following experiments.

3.3. Discussion on region compatibility

The above analysis shows that RC_{Jac} has more interesting properties than RC_{ID}. Therefore we compare RC_{Jac} with other two stability indicators.

Region stability, *ReSt*, is the most closely related concept to RC_{Jac}. Both consider the decision regions, but they have different stability definitions. *ReSt* considers only pairs of identical decision regions, i.e., other decision region pairs are ignored even if a decision region $R_{1,i}$ in T_1 is a subset of a decision region $R_{2,j}$ in T_2. In contrast, RC_{Jac} considers all cases, because $W_{A,B}$ is non-zero for any set A and B if $A \cap B$ is not empty. Therefore, RC_{Jac} provides more subtle observations on the decision region set compared to *ReSt*. *ReSt* and RC_{Jac} are both capable of recognizing two identical sets of decision regions F_1 and F_2, because $ReSt = 1$ if and only if $F_1 = F_2$ (Theorem 1), and $RC_{Jac} = 0$ if and only if $F_1 = F_2$ (Theorem 4).

Agreement is a semantic stability indicator, and significantly different from RC_{Jac}. *Agt* focuses on test instance prediction consistency between the decision trees, whereas RC_{Jac} also considers the decision regions. Two decision trees with equal *Agt* on the same test data may have different decision region sets, because the test instances may fall into different leaves with the same label between the two trees. In short, RC_{Jac} emphasizes decision tree structural stability whereas *Agt* is a semantic stability indicator.

Thus, we can make two conclusions regarding RC_{Jac}.

1. RC_{Jac} is the distance between BPAs, which provides a more subtle comparison than *ReSt* between decision region sets.
2. *Agt* is a pure semantic stability indicator, whereas RC_{Jac} is a structural stability indicator that also includes semantic factors from the Jaccard index. Future research will clarify how to deeply combine structural and semantic stability.

4. Experimental evaluation

4.1. Validation

We provide an example to show how region compatibility works, and validate the proposed approach.

Assume that dataset $D = \{1, 2, ..., 12\}$, with instance labels

$$Label(1) = 0; Label(2) = 0;$$
$$Label(3) = 0; Label(4) = 1;$$

$$Label(5) = 1; Label(6) = 0;$$
$$Label(7) = 1; Label(8) = 1;$$

$$Label(9) = 1; Label(10) = 1;$$
$$Label(11) = 0; Label(12) = 0.$$

4.1.1. Two trees with a single identical decision region

Suppose two trees, T_1 and T_2, are trained by randomly selected subsets of D, producing decision regions $F_1 = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9, 10, 11, 12\}\}$, and

$$F_2 = \{\{1, 2, 3\}, \{4, 7, 8, 12\}, \{5, 6, 9, 10, 11\}\},$$

respectively, with corresponding BPAs m_1 and m_2.

We determinate the decision region labels by majority vote,

$$Label(F_1) = [0, 1, 1],$$

$$Label(F_2) = [0, 1, 1],$$

and

$$m_1 - m_2 = [0.25, 0.5, -0.33, 0.42]^T.$$

Thus, from Section 3.2.2,

$$ReSt(T_1, T_2, D) = 3/12 = 0.25,$$

$$Agt(T_1, T_2, D) = 1,$$

$$RC_{ID}(F_1, F_2) = 0.7732,$$

and

$$RC_{Jac}(F_1, F_2) = 0.4487.$$

4.1.2. Two trees without identical decision regions

Consider a third tree, T_3, with region set

$$F_3 = \{\{1, 2, 4\}, \{3, 11, 12\}, \{5, 6, 7, 8, 9, 10\}\},$$

and corresponding BPA, m_3.

From majority voting, $Label(F_3) = [0, 0, 1]$, hence

$$ReSt(T_1, T_3, D) = 0.0,$$

$$Agt(T_1, T_3, D) = 9/12 = 0.75,$$

and

$$m_1 - m_3 = [0.25, 0.25, 0.5, -0.25, -0.25, -0.5]^T.$$

Thus,

$$RC_{ID}(F_1, F_3) = \sqrt{(m_1 - m_3)^T (m_1 - m_3)} = 0.8660,$$

and

$$RC_{Jac}(F_1, F_3) = \sqrt{(m_1 - m_3)^T W (m_1 - m_3)} = 0.4946,$$

where W is the Jaccard matrix of F_1 and F_3,

$$W = \begin{bmatrix} 1 & 0 & 0 & 2/4 & 1/5 & 0 \\ 0 & 1 & 0 & 1/5 & 0 & 2/7 \\ 0 & 0 & 1 & 0 & 2/7 & 4/8 \\ 2/4 & 1/5 & 0 & 1 & 0 & 0 \\ 1/5 & 0 & 2/7 & 0 & 1 & 0 \\ 0 & 2/7 & 4/8 & 0 & 0 & 1 \end{bmatrix}.$$

Now consider a fourth tree, T_4, with region set

$$F_4 = \{\{1, 2, 7\}, \{3, 5, 6\}, \{4, 8, 9, 10, 11, 12\}\},$$

and corresponding BPA, m_4. Then

$$Label(F_4) = [0, 0, 1],$$

$$ReSt(T_1, T_4, D) = 0.0,$$

$$Agt(T_1, T_4, D) = 8/12 = 0.667,$$

and

$$m_1 - m_4 = [0.25, 0.25, 0.5, -0.25, -0.25, -0.5]^T.$$

Thus,

$$RC_{ID}(F_1, F_4) = \sqrt{(m_1 - m_4)^T (m_1 - m_4)} = 0.8660,$$

and

$$RC_{Jac}(F_1, F_4) = \sqrt{(m_1 - m_4)^T W (m_1 - m_4)} = 0.4247,$$

where W is the Jaccard matrix of F_1 and F_4,

$$W = \begin{bmatrix} 1 & 0 & 0 & 2/4 & 1/5 & 0 \\ 0 & 1 & 0 & 0 & 2/4 & 1/8 \\ 0 & 0 & 1 & 1/8 & 0 & 5/7 \\ 2/4 & 0 & 1/8 & 1 & 0 & 0 \\ 1/5 & 2/4 & 0 & 0 & 1 & 0 \\ 0 & 1/8 & 5/7 & 0 & 0 & 1 \end{bmatrix}.$$

Thus, although $RC_{ID}(F_1, F_3) = RC_{ID}(F_1, F_4)$, $RC_{Jac}(F_1, F_3) \neq RC_{Jac}(F_1, F_4)$, because F_1 and F_3 are not overlapped, hence, $m_1 - m_3 = [m_1, -m_3]^T$. This same situation applies for F_1 and F_4, hence $RC_{ID}(F_1, F_3) = RC_{ID}(F_1, F_4)$. However, the Jaccard matrix of F_1 and F_3 is different from that of F_1 and F_4, thus $RC_{Jac}(F_1, F_3) \neq RC_{Jac}(F_1, F_4)$. Therefore, RC_{Jac} is more sensitive than RC_{ID} to distinguish decision trees.

4.2. Evaluation on UCI datasets

4.2.1. Data sets

We evaluate the proposed region compatibility indicators on 20 real-world datasets from the UCI repository (Lichman, 2017). Table 3 summarizes the datasets.

4.2.2. Region compatibility convergence for a special case

We simulate a special case suitable to apply Theorems 7 and 8. Given a labeled dataset $D = \{x_1, x_2, ..., x_N\}$, let F_1 be the decision region set for a decision tree trained on D, and F_2 be the perturbed version of F_1 obtained as shown in Algorithm 1.

Algorithm 1 (Perturbation algorithm).

Input: F_1, perturbation ratio $\in [0, 0.5]$
Output: F_2

(1) Randomly select a pair i, j, $i \neq j$, from $\{1, 2, \cdots, N\}$, where N is the number of instances of D.
(2) Exchange the positions of x_i and x_j, i.e., if x_i and x_j belong to R_i and R_j, respectively, then they belong to R_j and R_i, respectively, after exchanging.
(3) Repeat (1) and (2) k times, where $k = N \times ratio$.

Table 3
UCI datasets for validation.

Data set	# of attributes	# of classes	# of instances
Balance	4	3	625
Car Evaluation	6	4	1728
Chess	36	2	3196
Ecoli	7	8	336
Glass	9	7	214
Heart	13	2	270
Image Segmentation	19	7	2310
Ionosphere	34	2	351
Iris	4	3	150
Nursery	8	5	12,960
Page-blocks	10	5	5473
Pima Indians	8	2	768
Seed	7	3	210
Soybean	35	4	47
Tic-tac-toc	9	2	958
Waveform	21	3	5000
Wdbc	30	2	569
Wine	13	3	178
Yeast	8	10	1484
Zoo	16	7	101

Thus, Algorithm 1 perturbs F_1, and we generate a new decision region set, F_2. For example, let $F_1 = \{\{1,2\},\{3\},\{4,5,6\},\{7,8\}\}$. If (2, 3) and (4, 7) are selected to exchange positions, then $F_2 = \{\{1,3\},\{2\},\{5,6,7\},\{4,8\}\}$. Some blocks of F_1 are changed in F_2, but F_2 has the same structure as F_1, i.e., $V_1 = V_2$. Therefore, we can achieve $F_1 \cap F_2 = \phi$ if enough instances are exchanged to ensure each F_1 decision region is perturbed.

To calculate the region compatibility for each dataset, we obtain a decision tree T for each dataset using CART with default setting, randomly selecting 70% of the dataset for training. Specifically, we use the function $tree(class \sim ., data = data\_train)$ in R package tree to create a tree. We vary the perturbation ratio from 1% to 49% with step length 1%, and compute region compatibility between T and each perturbed tree. We conducted the experiment on 20 real-world datasets. Due to space limits, we show only graphical results for eight datasets in Fig. 3. Fig. 3 shows the average results for 30 repeated measures. As we see, RC_{Jac} increases with increasing perturbation ratio, because the distance between a decision tree and its perturbed version is positively correlated with the perturbation ratio.

Table 4 compares theoretical and experimental convergence on the 20 datasets, averaged over the 30 repeated cases. In statistics, Mean Absolute Error (MAE) is a measure of difference between two continuous variables (Willmott & Matsuura, 2005). Assume $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots, y_n]$ are variables of theoretical expectation and experimental measurement, MAE is given by following equation:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}.$$

As shown in Table 4, region compatibility shows the interesting convergence properties as expected from Theorems 7 to 8, where the specific convergence value depends on the particular dataset. It is easy to see that the experimental value for each dataset is highly consistent with the theoretical values, which also validate Theorems 7 and 8.

4.2.3. Understanding region compatibility

We need to consider what region compatibility means in terms of decision tree stability. Therefore, let us try to align RC_{Jac} for decision trees T_1 and T_2 to the RC_{Jac} values of T_1 and its perturbed versions.

Using the 30 decision trees with 70% training ratio for each dataset from Section 4.2.2, we computed RC_{Jac} for each pair of trees, i.e., $30 \times 29/2 = 435$ measures, reported the maximum, minimum, and average metrics with the RC_{Jac} convergence for each dataset, and only presented the details of eight datasets in Fig. 4 due to space limits.

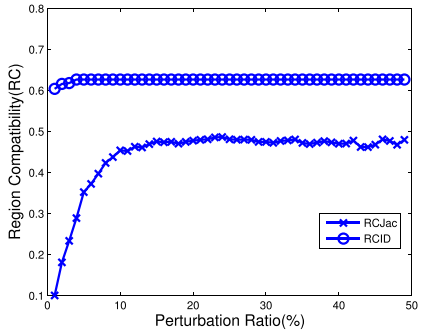
Since the RC_{Jac} curve is monotonic increasing, it is trivial to align RC_{Jac} between two trees T_1 and T_2 to the curve to identify its corresponding perturbation ratio. In particular, Fig. 4(f) shows that $\text{Max}(RC_{Jac})$ corresponds to perturbation ratio $< 4\%$. In other words, the maximum difference between these 30 decision trees on Iris data is less unstable than perturbing one tree within 4% pairs of total instances by exchanging their positions in decision regions.

We also computed RC_{Jac} for other two decision tree induction algorithms: CTREE (R package party with default setting, i.e., $ctree(class \sim ., data = data\_train)$) and C4.5 (R package RWeka with default setting, i.e., $J48(class \sim ., data = data\_train)$). Similar to Fig. 4, Figs. B.6 and B.7 in Appendix B show the performance of

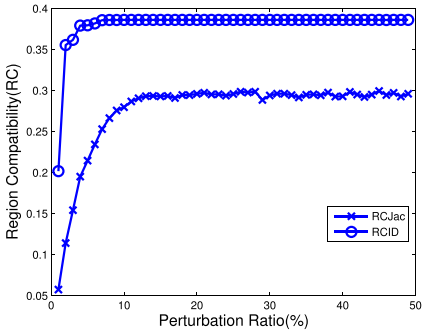
Table 4
Theoretical and experimental convergence.

Dataset	RC_{ID}_T	RC_{ID}_E	MAE_{ID}	RC_{Jac}_T	RC_{Jac}_E	MAE_{Jac}
Balance	0.396 ± 0.023	0.396 ± 0.023	5.83E−17	0.301 ± 0.021	0.301 ± 0.021	0.001
Car Eval.	0.598 ± 0.002	0.598 ± 0.002	1.11E−17	0.460 ± 0.000	0.460 ± 0.002	0.002
Chess	0.550 ± 0.015	0.550 ± 0.015	2.22E−17	0.402 ± 0.007	0.401 ± 0.007	0.000
Ecoli	0.617 ± 0.034	0.617 ± 0.034	2.78E−17	0.464 ± 0.024	0.463 ± 0.024	0.004
Glass	0.398 ± 0.019	0.398 ± 0.019	1.11E−05	0.309 ± 0.015	0.307 ± 0.015	0.003
Heart	0.487 ± 0.043	0.487 ± 0.043	7.11E−06	0.383 ± 0.033	0.382 ± 0.035	0.003
Image Seg.	0.483 ± 0.015	0.483 ± 0.015	2.22E−17	0.341 ± 0.008	0.338 ± 0.008	0.003
Ionosphere	0.725 ± 0.025	0.725 ± 0.025	7.77E−17	0.539 ± 0.014	0.539 ± 0.014	0.004
Iris	0.767 ± 0.014	0.767 ± 0.014	2.22E−17	0.509 ± 0.003	0.506 ± 0.004	0.003
Nursery	0.588 ± 0.008	0.588 ± 0.008	1.05E−16	0.439 ± 0.002	0.438 ± 0.002	0.001
Page-blo.	1.119 ± 0.036	1.119 ± 0.036	5.55E−17	0.650 ± 0.005	0.650 ± 0.006	0.003
Pima Ind.	0.536 ± 0.059	0.536 ± 0.059	5.55E−17	0.403 ± 0.049	0.402 ± 0.049	0.002
Seed	0.674 ± 0.029	0.674 ± 0.029	2.22E−17	0.473 ± 0.014	0.470 ± 0.015	0.004
Soybean	0.734 ± 0.008	0.734 ± 0.008	6.11E−17	0.489 ± 0.007	0.481 ± 0.010	0.008
Tic-tac-toc	0.483 ± 0.081	0.483 ± 0.081	4.44E−17	0.370 ± 0.057	0.370 ± 0.057	0.001
Waveform	0.558 ± 0.021	0.558 ± 0.021	2.22E−17	0.408 ± 0.015	0.408 ± 0.015	0.000
Wdbc	0.834 ± 0.043	0.834 ± 0.043	2.01E−07	0.579 ± 0.019	0.580 ± 0.019	0.003
Wine	0.733 ± 0.015	0.733 ± 0.015	2.19E−06	0.496 ± 0.007	0.495 ± 0.006	0.002
Yeast	0.652 ± 0.037	0.652 ± 0.037	3.33E−17	0.478 ± 0.029	0.478 ± 0.030	0.001
Zoo	0.706 ± 0.009	0.706 ± 0.009	5.55E−17	0.502 ± 0.002	0.498 ± 0.005	0.006

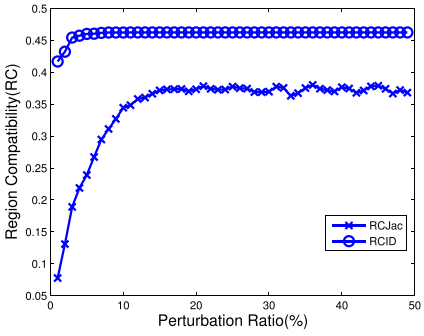
Notes: RC_{ID}_T and RC_{Jac}_T denote the theoretical values of RC_{ID} and RC_{Jac}, while RC_{ID}_E and RC_{Jac}_E denote the experimental values of RC_{ID} and RC_{Jac}, respectively.



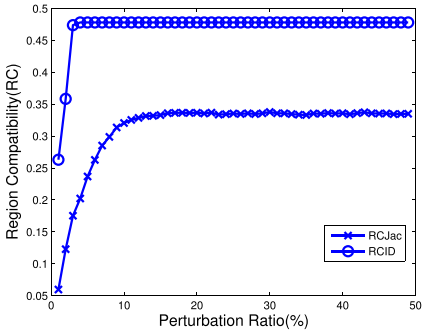
(a) Ecoli



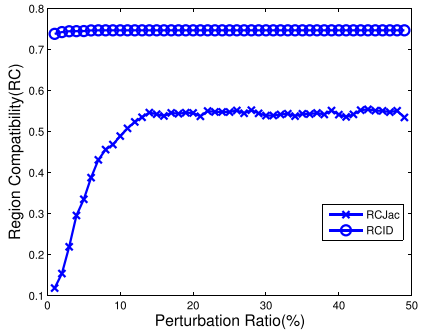
(b) Glass



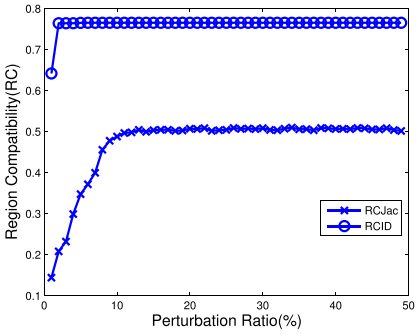
(c) Heart



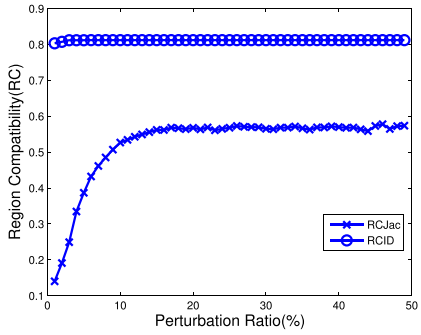
(d) Image Segmentation



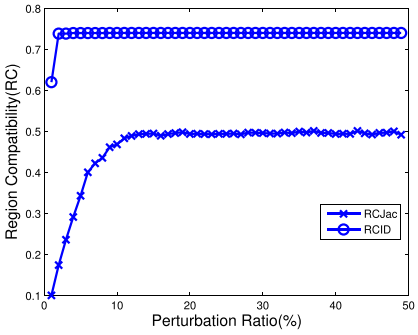
(e) Ionosphere



(f) Iris

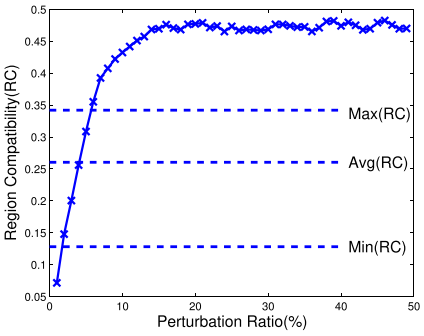


(g) Wdbc

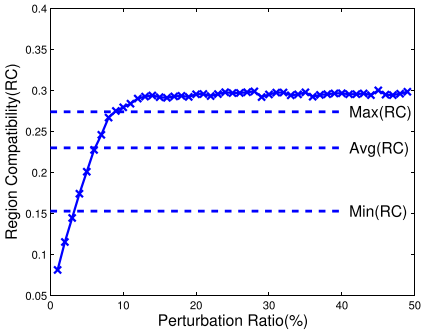


(h) Wine

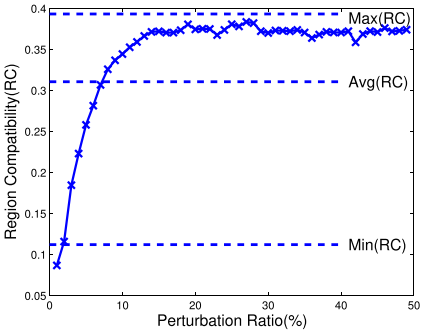
Fig. 3. Region compatibility for eight datasets.



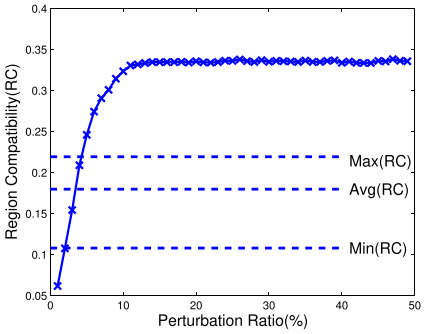
(a) Ecoli



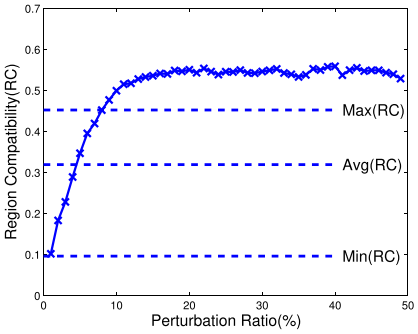
(b) Glass



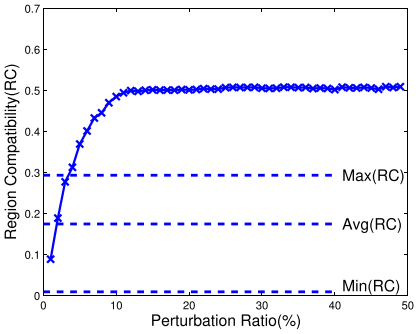
(c) Heart



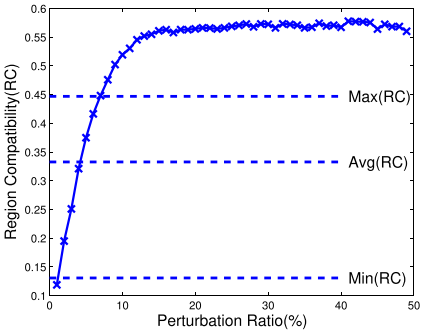
(d) Image Segmentation



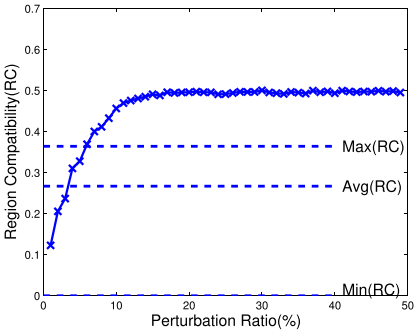
(e) Ionosphere



(f) Iris



(g) Wdbc



(h) Wine

Fig. 4. RC_{Jac} for CART.

Table 5
Rank and *p*-value of Friedman test on eight datasets.

dataset		50%	60%	70%	80%	90%
Ecoli	CART	2.04	2.10	2.02	2.34	2.28
	CTREE	2.36	2.44	2.37	2.10	2.07
	C4.5	1.60	1.46	1.60	1.56	1.65
	p-value	.000	.000	.000	.000	.000
Glass	CART	1.43	1.29	1.62	1.65	1.93
	CTREE	2.78	2.67	2.24	1.93	1.36
	C4.5	1.79	2.04	2.14	2.42	2.73
	p-value	.000	.000	.000	.000	.000
Heart	CART	1.36	1.47	1.54	1.60	1.79
	CTREE	2.37	2.18	2.00	1.80	1.84
	C4.5	2.27	2.35	2.47	2.59	2.37
	p-value	.000	.000	.000	.000	.000
ImageSeg	CART	1.64	1.54	1.71	1.78	1.90
	CTREE	2.79	2.75	2.82	2.83	2.47
	C4.5	1.57	1.71	1.47	1.39	1.63
	p-value	.000	.000	.000	.000	.000
Ionosphere	CART	1.96	1.82	1.64	1.89	1.67
	CTREE	1.74	1.89	1.89	1.73	2.03
	C4.5	2.30	2.29	2.48	2.38	2.30
	p-value	.000	.000	.000	.000	.000
Iris	CART	1.95	1.77	2.06	2.18	2.37
	CTREE	1.99	2.14	1.88	1.87	1.70
	C4.5	2.06	2.09	2.06	1.95	1.93
	p-value	**.215**	.000	.009	.000	.000
Wdbc	CART	2.11	2.40	2.30	2.37	2.27
	CTREE	1.85	1.66	1.63	1.60	1.59
	C4.5	2.04	1.94	2.06	2.03	2.15
	p-value	.000	.000	.000	.000	.000
Wine	CART	1.80	1.78	1.84	1.78	2.04
	CTREE	2.30	2.26	2.23	2.31	1.95
	C4.5	1.90	1.96	1.93	1.94	2.01
	p-value	.000	.000	.000	.000	**.445**

CTREE and C4.5, respectively. We can understand the meaning of RC_{Jac} similarly.

4.2.4. Comparing three decision tree algorithms

Furthermore, we evaluated the impact of training ratio on region compatibility. We repeated each experiment 30 times, with mean values shown in Fig. 5. The different decision tree algorithms produce various mean values for each dataset under the same training ratio. Generally, larger training data implies more stability between the decision trees, and the experimental results confirm that mean RC_{Jac} between decision trees decreased with increased training data.

To compare the three algorithms (i.e., CART, CTREE and C4.5), we evaluated the following hypothesis H_0 using Friedman test (Friedman, 1937).

Null hypothesis H_0: Decision tree induction algorithms (CART, CTREE and C4.5) do not show any significant difference when evaluated using RC_{Jac}.

The statistics of Friedman test is based on chi-square distribution with $n-1$ degrees of freedom, where n corresponds to the number of compared algorithms in this paper. As the study compares 3 algorithms, the degrees of freedom is 2. The Friedman test was applied by using all RC_{Jac} values of 30 rounds under each training ratio. We calculated the rank and p-value for each test, and the hypothesis was checked at $\alpha = 0.05$ significance level, as shown in Table 5.

Only 2 cases in Table 5 have p-value $> .05$, which means the Friedman test results are not significant at $\alpha = 0.05$. Thus, we accept Null hypothesis H_0 and the three algorithms have similar stability on Iris dataset with training ratio 50% and Wine dataset with training ratio 90%.

The other 38 cases show that the Friedman test results are significant at $\alpha = 0.05$. Thus, we reject Null hypothesis H_0. Namely, the three algorithms perform significantly different from each other when evaluated using RC_{Jac}. Fig. 5 gives us a way to select a relatively stable decision tree learning algorithm. We applied the interpretation "the lower RC_{Jac}, the better algorithm", since low RC_{Jac} indicates stable decision trees and credible rules derived from the decision trees. For a given dataset, we can recommend the algorithm with lowest mean RC_{Jac} in Fig. 5. In particular, CTREE is recommended to Wdbc dataset, and C4.5 is recommended to Ecoli dataset according to Fig. 5.

This section evaluated region compatibility on 20 real world UCI datasets and obtained the following important observations.

1. Region compatibility suggests that some similarity exists for any two decision trees, even if they are apparently different.
2. Region compatibility convergence was proved in Section 3.2.2, and validated in the current section using the perturbation algorithm.
3. Region compatibility shows more data involved in decision tree training implies more stable between the resultant trees, which is consistent with intuition.
4. We clarified the meaning of region compatibility in terms of decision tree stability, by aligning RC_{Jac} for two trees to the RC_{Jac} curve of perturbation ratio (Fig. 4) to identify the corresponding perturbation ratio, r. Region compatibility shows difference between two decision trees is less unstable than perturbing one tree within r pairs of total instances by exchanging their positions in decision regions.
5. Three well-known decision tree learning algorithms perform significantly different on tested real-world datasets, and the algorithm with lowest RC_{Jac} is expected to induce relatively stable decision trees and derive credible rules for a given dataset.

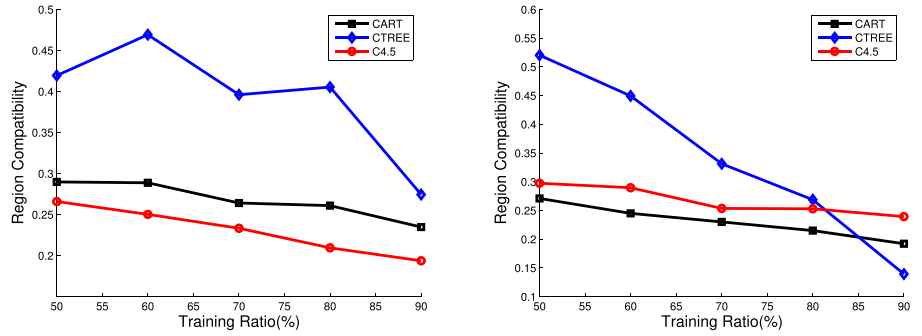
In summary, the experiments show that region compatibility has a solid theoretical foundation, and provides a reasonable and meaningful assessment for decision tree stability.

5. Conclusions

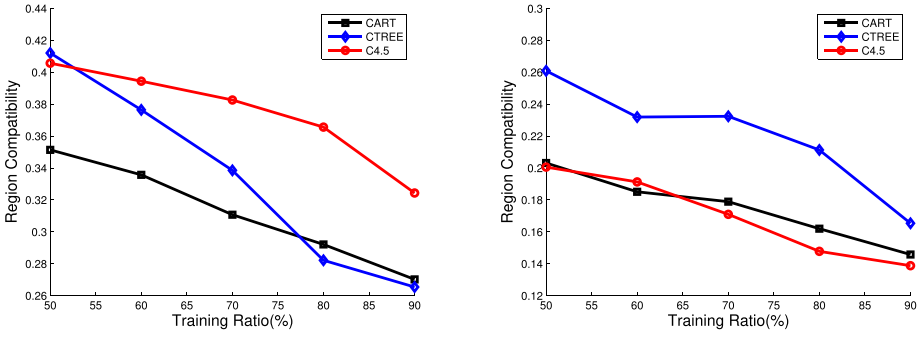
Decision tree learners are highly unstable, producing significantly different classifiers from slightly different training sets, and various metrics have been proposed to quantify stability. However, current metrics fail to consider potential similarities of apparently different decision trees. This paper proposed a region compatibility metric based on Dempster–Shafer theory and explained RC_{Jac} experimentally by its corresponding perturbation ratio. The proposed region compatibility metric was effective and powerful to quantify decision tree learning algorithm stability. Therefore, region compatibility has a potentially broad range of applications, including image understanding (Gardner, Kanno, Duncan, & Selmic, 2014) and preference learning (Liu & Liao, 2015).

In future research, we plan to explore the region compatibility approach for a range of potential applications.

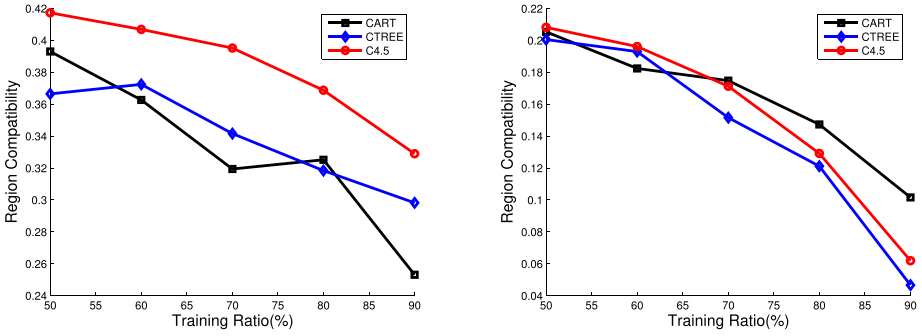
- Compare decision trees labeled regions, matching leaves with the same label to provide more reasonable comparisons.
- Assess preference decision trees stability, to predict the preference relationships between each pair of objects by combining decision trees with conditional preference networks (CP-nets).
- Investigate clustering evaluation metrics using region compatibility, particularly for the k-means algorithm, which suffers from similar stability problem. K-means algorithm is unstable because initial cluster centers are randomly selected, and the random initial centers always lead k-means to local optimum easily.



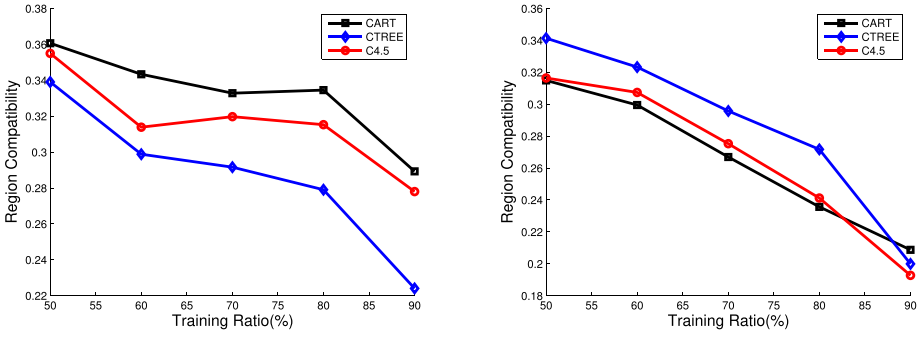
(a) Ecoli (b) Glass



(c) Heart (d) Image Segmentation



(e) Ionosphere (f) Iris



(g) Wdbc (h) Wine

Fig. 5. Mean values of RC_{Jac}.

- Investigate complex object matching in image understanding regarding region compatibility as a distance metric between complex objects that include multiple parts.

Acknowledgments

The authors are very grateful to the anonymous reviewers for their insightful comments and suggestions that have led to an improved version of this paper. This work was supported by National Natural Science Foundation of China (Nos. 61773331, 61572419, 61403328 and 71672166) and a Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA091). We thank International Science Editing (http://www.internationalscienceediting.com) for editing this manuscript.

Appendix A. Proofs of Theorems

A.1. Proof of Theorem 5

Proof. In Eq. (5), the entry that corresponds to $R_{1,1}$ is updated to $m_1(R_{1,1}) - m_2(R_{2,1}) = 0$, hence it has no effect on the inner product $\langle m_1 - m_2,\ m_1 - m_2\rangle$, and region compatibility between two partially identical trees is equivalent to comparing the difference of the trees. □

A.2. Proof of Theorem 6

Proof. For two completely distinct decision trees, $F_1 \cap F_2 = \phi$, non-zero entries of $m_1 - m_2$ can be regarded as a column vector formed by appending $-V_2$ to V_1, i.e.,

$$m_1 - m_2 = \begin{bmatrix} V_1 \\ -V_2 \end{bmatrix}.$$

Hence,

$$\begin{aligned} RC_{ID}(F_1, F_2) &= \sqrt{(m_1 - m_2)^T (m_1 - m_2)} \\ &= \sqrt{[V_1^T, -V_2^T]\begin{bmatrix} V_1 \\ -V_2 \end{bmatrix}} \\ &= \sqrt{V_1^T V_1 + V_2^T V_2}. \end{aligned}$$

Since (see Eq. (5))

$$V_1 = [m_1(R_{1,1}), m_1(R_{1,2}), \ldots, m_1(R_{1,s})]^T,$$

$$V_2 = [m_2(R_{2,1}), m_2(R_{2,2}), \ldots, m_2(R_{2,t})]^T,$$

and

$$m_1(R_{1,i}) = |R_{1,i}|/|D|,\ m_2(R_{2,j}) = |R_{2,j}|/|D|;$$

then

$$V_1^T V_1 = \frac{\sum_i |R_{1,i}|^2}{|D|^2}, \quad V_2^T V_2 = \frac{\sum_j |R_{2,j}|^2}{|D|^2}.$$

Therefore,

$$RC_{ID}(F_1, F_2) = \frac{1}{|D|}\sqrt{\sum_i |R_{1,i}|^2 + \sum_j |R_{2,j}|^2}.$$

The Jaccard matrix, W, can be regarded as a block matrix,

$$W = \begin{bmatrix} I_1 & W_{12} \\ W_{12}^T & I_2 \end{bmatrix},$$

where I_1 is the $|F_1| \times |F_1|$ identity matrix, and I_2 is the $|F_2| \times |F_2|$ identity matrix. Therefore,

$$\begin{aligned} RC_{Jac}(F_1, F_2) &= \sqrt{(m_1 - m_2)^T W (m_1 - m_2)} \\ &= \sqrt{[V_1^T, -V_2^T]\begin{bmatrix} I_1 & W_{12} \\ W_{12}^T & I_2 \end{bmatrix}\begin{bmatrix} V_1 \\ -V_2 \end{bmatrix}} \\ &= \sqrt{V_1^T V_1 + V_2^T V_2 - 2V_1^T W_{12} V_2}. \end{aligned}$$

Thus,

$$RC_{Jac}(F_1, F_2) = \sqrt{\frac{1}{|D|^2}\left(\sum_i |R_{1,i}|^2 + \sum_j |R_{2,j}|^2\right) - 2V_1^T W_{12} V_2}.$$

□

A.3. Proof of Theorem 7

Proof. Since $V_1 = V_2$,

$$\begin{aligned} RC_{ID}(F_1, F_2) &= \sqrt{(m_1 - m_2)^T (m_1 - m_2)} \\ &= \sqrt{V_1^T V_1 + V_2^T V_2} \\ &= \sqrt{2V_1^T V_1}. \end{aligned}$$

□

A.4. Proof of Theorem 8

Proof. Let $D = \{x_1, x_2, \ldots, x_N\}$, $F_1 = \{R_{1,1}, R_{1,2}, \ldots, R_{1,s}\}$, $F_2 = \{R_{2,1}, R_{2,2}, \ldots, R_{2,t}\}$,

$$\begin{aligned} V_1 = {} & [m_1(R_{1,1}), m_1(R_{1,2}), \ldots, m_1(R_{1,s})]^T, s.t. \\ & 0 < m_1(R_{1,1}) \le m_1(R_{1,2}) \le \ldots \le m_1(R_{1,s}), \end{aligned}$$

$$\begin{aligned} V_2 = {} & [m_2(R_{2,1}), m_2(R_{2,2}), \ldots, m_2(R_{2,t})]^T, s.t. \\ & 0 < m_2(R_{2,1}) \le m_2(R_{2,2}) \le \ldots \le m_2(R_{2,t}). \end{aligned}$$

To simplify, we assume that

$$R_{1,1} = \underbrace{\{x_1, x_2, \ldots, x_{i_1}\}}_{n_1},$$

$$R_{1,2} = \underbrace{\{x_{i_1+1}, x_{i_1+2}, \ldots, x_{i_2}\}}_{n_2}, \ldots,$$

and

$$R_{1,s} = \underbrace{\{x_{i_{s-1}+1}, x_{i_{s-1}+2}, \ldots, x_N\}}_{n_s}.$$

Since $V_1 = V_2$ implies cardinality of F_1 equals that of F_2, i.e., $s = t$, and the cardinalities of focal elements in F_1 correspond to those of focal elements in F_2,

$$R_{2,1} = \underbrace{\{x'_1, x'_2, \ldots, x'_{i_1}\}}_{n_1},$$

$$R_{2,2} = \underbrace{\{x'_{i_1+1}, x'_{i_1+2}, \ldots, x'_{i_2}\}}_{n_2}, \ldots,$$

and

$$R_{2,s} = \underbrace{\{x'_{i_{s-1}+1}, x'_{i_{s-1}+2}, \ldots, x'_N\}}_{n_s}.$$

From the definition of W_{12} in Theorem 6,

$$W_{12}(R_{1,i}, R_{2,j}) = \frac{|R_{1,i} \cap R_{2,j}|}{|R_{1,i} \cup R_{2,j}|}, R_{1,i} \in F_1, R_{2,j} \in F_2.$$

We need the expectation of $|R_{1,i} \cap R_{2,j}|$ due to the random nature of F_1 and F_2. If we regard $|R_{1,i} \cap R_{2,j}|$ as a random variable, then it has a hypergeometric distribution (N, n_i, n_j), where $n_i = |R_{1,i}|, n_j = |R_{2,j}|$.

$|R_{1,i} \cap R_{2,j}| = 1$ means there is exactly one element shared by $R_{1,i}$ and $R_{2,j}$, i.e., only one element in $R_{1,i}$ comes from $R_{2,j}$, and the other elements come from focal elements other than $R_{2,j}$. The corresponding probability is

$$P(|R_{1,i} \cap R_{2,j}| = 1) = \frac{\begin{pmatrix} n_j \\ 1 \end{pmatrix}\begin{pmatrix} N - n_j \\ n_i - 1 \end{pmatrix}}{\begin{pmatrix} N \\ n_i \end{pmatrix}}.$$

If $|R_{1,i} \cap R_{2,j}| = 2$, then

$$P(|R_{1,i} \cap R_{2,j}| = 2) = \frac{\begin{pmatrix} n_j \\ 2 \end{pmatrix}\begin{pmatrix} N - n_j \\ n_i - 2 \end{pmatrix}}{\begin{pmatrix} N \\ n_i \end{pmatrix}}.$$

Since $n_i \le N - n_j$, and $|R_{1,i} \cap R_{2,j}| \le min(|R_{1,i}|, |R_{2,j}|)$,

$$P(|R_{1,i} \cap R_{2,j}| = m) = \frac{\begin{pmatrix} n_j \\ m \end{pmatrix}\begin{pmatrix} N - n_j \\ n_i - m \end{pmatrix}}{\begin{pmatrix} N \\ n_i \end{pmatrix}}$$

for $0 \le m \le min(|R_{1,i}|, |R_{2,j}|) = min(n_i, n_j)$, and $P(|R_{1,i} \cap R_{2,j}| = m) = 0$ otherwise.

Thus the expectation of $|R_{1,i} \cap R_{2,j}|$ is

$$E(|R_{1,i} \cap R_{2,j}|) = n_i \times \frac{n_j}{N} = \frac{n_i \times n_j}{N}.$$

Hence,

$$\begin{aligned} W_{12}(R_{1,i}, R_{2,j}) &= \frac{|R_{1,i} \cap R_{2,j}|}{|R_{1,i} \cup R_{2,j}|} \\ &= \frac{|R_{1,i} \cap R_{2,j}|}{|R_{1,i}| + |R_{2,j}| - |R_{1,i} \cap R_{2,j}|}, \\ &= \frac{|R_{1,i} \cap R_{2,j}|}{n_i + n_j - |R_{1,i} \cap R_{2,j}|}. \end{aligned}$$

and

$$E(W_{12}(R_{1,i}, R_{2,j})) = \frac{n_i \times n_j / N}{n_i + n_j - n_i \times n_j / N} = \frac{n_i \times n_j}{N \times (n_i + n_j) - n_i \times n_j}.$$

From Theorem 6, if $F_1 \cap F_2 = \phi$,

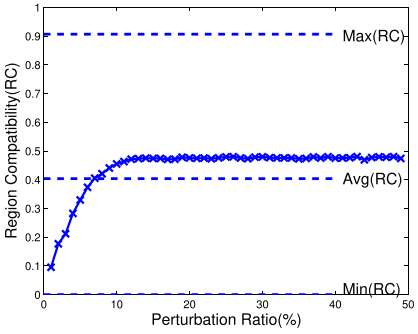
$$\begin{aligned} RC_{Jac}(F_1, F_2) &= \sqrt{(m_1 - m_2)^T W (m_1 - m_2)} \\ &= \sqrt{[V_1^T, -V_2^T]\begin{bmatrix} I_1 & W_{12} \\ W_{12}^T & I_2 \end{bmatrix}\begin{bmatrix} V_1 \\ -V_2 \end{bmatrix}} \\ &= \sqrt{V_1^T V_1 + V_2^T V_2 - 2V_1^T W_{12} V_2}. \end{aligned}$$

Fixing $V_1 = V_2$, the expectation of RC_{Jac} is

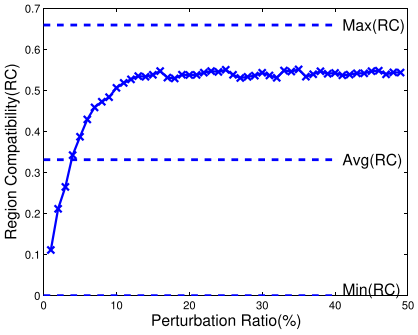
$$\begin{aligned} E(RC_{Jac}(F_1, F_2)) &= \sqrt{V_1^T V_1 + V_2^T V_2 - 2V_1^T E(W_{12}) V_2} \\ &= \sqrt{2V_1^T V_1 - 2V_1^T E(W_{12}) V_2}. \end{aligned}$$

□

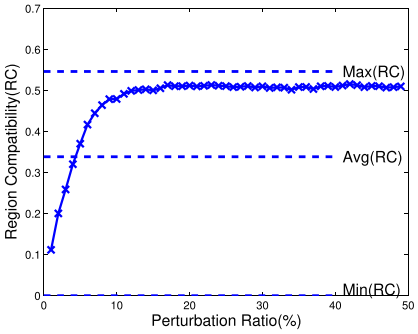
Appendix B. RC_{Jac} for CTREE and C4.5



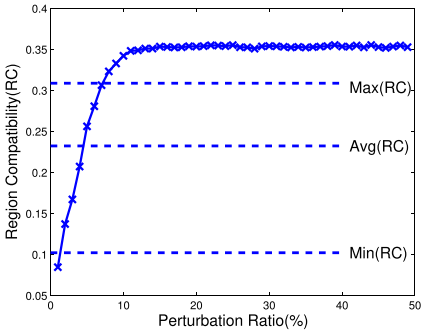
(a) Ecoli



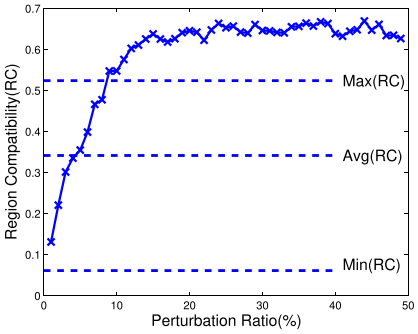
(b) Glass



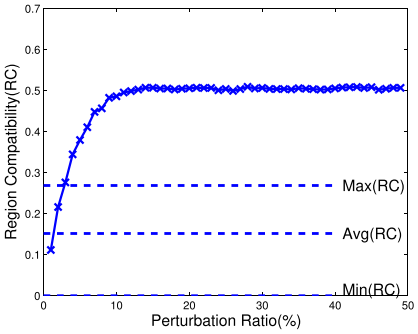
(c) Heart



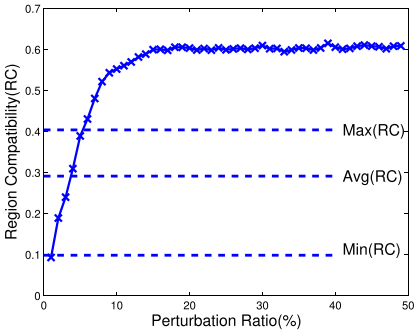
(d) Image Segmentation



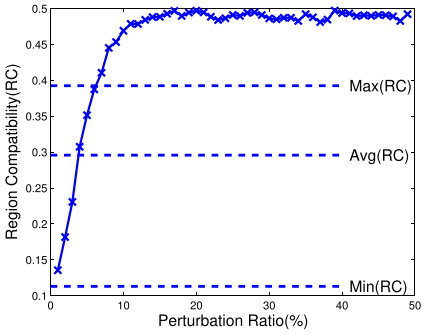
(e) Ionosphere



(f) Iris

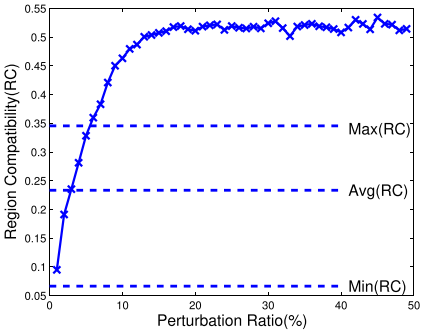


(g) Wdbc

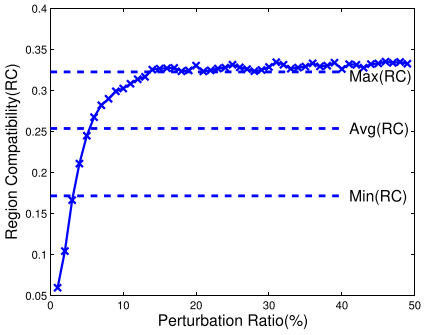


(h) Wine

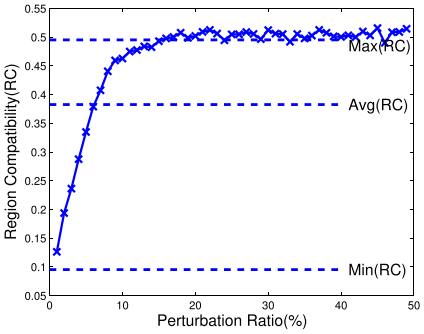
Fig. B.6. RC_{Jac} for CTREE.



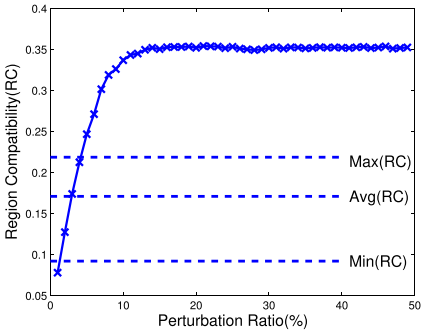
(a) Ecoli



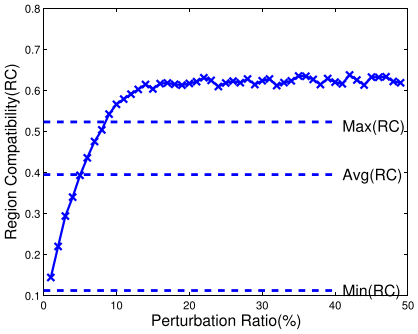
(b) Glass



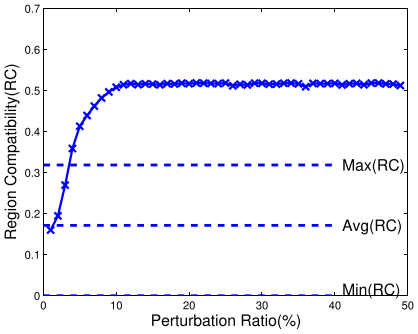
(c) Heart



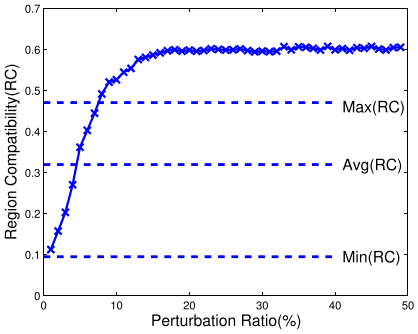
(d) Image Segmentation



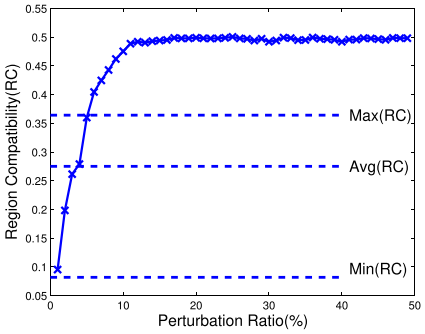
(e) Ionosphere



(f) Iris



(g) Wdbc



(h) Wine

Fig. B.7. RC_{Jac} for C4.5.

References

Aluja-Banet, T., & Nafria, E. (2003). Stability and scalability in decision trees. *Computational Statistics, 18*(3), 505–520.

Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2007). A comparison of decision tree ensemble creation techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(1), 173–180.

Bouchard, M., Jousselme, A.-L., & Doré, P.-E. (2013). A proof for the positive definiteness of the Jaccard index matrix. *International Journal of Approximate Reasoning, 54*(5), 615–626.

Breiman, L. (2001). Random forest. *Machine Learning, 45*, 5–32.

Briand, B., Ducharme, G. R., Parache, V., & Mercat-Rommens, C. (2009). A similarity measure to assess the stability of classification trees. *Computational Statistics & Data Analysis, 53*(4), 1208–1217.

Dwyer, K., & Holte, R. (2007). Decision tree instability and active learning. In *European conference on machine learning* (pp. 128–139). Springer.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701.

Gardner, A., Kanno, J., Duncan, C. A., & Selmic, R. (2014). Measuring distance between unordered sets of different sizes. In *2014 IEEE conference on computer vision and pattern recognition (CVPR): 00* (pp. 137–143). doi:10.1109/CVPR.2014.25.

Ghattas, B., Michel, P., & Boyer, L. (2017). Clustering nominal data using unsupervised binary decision trees: Comparisons with the state of the art methods. *Pattern Recognition, 67*, 177–185.

Jousselme, A.-L., & Maupin, P. (2012). Distances in evidence theory: Comprehensive survey and generalizations. *International Journal of Approximate Reasoning, 53*(2), 118–145.

Kalles, D., & Papagelis, A. (2000). Stable decision trees: Using local anarchy for efficient incremental learning. *International Journal on Artificial Intelligence Tools, 9*(01), 79–95.

Kim, K. (2016). A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. *Pattern Recognition, 60*, 157–163.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning, 51*(2), 181–207.

Lichman, M. (2017). *UCI machine learning repository*. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml.

Liu, J.-L., & Liao, S.-Z. (2015). Expressive efficiency of two kinds of specific CP-nets. *Information Sciences, 295*(2), 379–394.

Meester, R. (2008). *A natural introduction to probability theory*. Springer Science & Business Media.

Mirzamomen, Z., & Kangavari, M. R. (2016). A framework to induce more stable decision trees for pattern classification. *Pattern Analysis and Applications*, 1–14.

Parvin, H., MirnabiBaboli, M., & Alinejad-Rokny, H. (2015). Proposing a classifier ensemble framework based on classifier selection and decision tree. *Engineering Applications of Artificial Intelligence, 37*, 34–42.

Paul, J., Verleysen, M., & Dupont, P. (2012). The stability of feature selection and class prediction from ensemble tree classifiers.. *Esann2012 special session on machine ensembles, European symposium on artificial neural networks, computational intelligence and machine learning*.

Pawlik, M., & Augsten, N. (2016). Tree edit distance: Robust and memory-efficient. *Information Systems, 56*, 157–173.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Rokach, L., & Maimon, O. (2005). Decision trees. In O. Maimon, & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 165–192)). Boston, MA: Springer US.

Tai, K. C. (1979). The tree-to-tree correction problem. *Journal of ACM, 26*(3), 422–433.

Turney, P. (1995). Technical note: Bias and the quantification of stability. *Machine Learning, 20*(1), 23–33.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79.

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications, 78*, 225–241.

Yager, R. R. (1987). On the Dempster-Shafer framework and new combination rules. *Information Sciences, 41*(2), 93–137.

Yang, H.-J., Roe, B. P., & Zhu, J. (2007). Studies of stability and robustness for artificial neural networks and boosted decision trees. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 574*(2), 342–349.

Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence, 137*(1–2), 239–263.

Zimmermann, A. (2008). Ensemble-trees: Leveraging ensemble power inside decision trees. In *International conference on discovery science* (pp. 76–87). Springer.