Sensitivity Analysis of Bipartition Dissimilarity under Tree Rearrangement Operations

Xin Xiao
College of Foreign Studies
Shandong Institute of Business and Technology
Yantai, China
xinxiaoyt@hotmail.com

Li Xinbo
Yantai No. 1 Middle School
Yantai, China
xinboyt@hotmail.com

Abstract—Trees are a powerful structure for representing hierarchical relations in a natural way. Comparison of trees is a recurrent task in various computer science related fields. The widely used Robinson-Foulds distance for comparing leaf labeled trees is overly sensitive to very small changes in the tree. The measure of bipartition dissimilarity refines Robinson-Foulds metric by comparing the quality of the tree bipartitions instead of their quantity. Sensitivity analysis is used in this paper which shows that bipartition dissimilarity has smaller sensitivity to small modifications in the tree.

Keywords- leaf labeled trees; Robinson-Foulds distance; bipartition dissimilarity; sensitivity analysis; tree rearrangement operations

I. INTRODUCTION

Representing data for which hierarchical relations can be defined in a tree-like structure is ubiquitous in many areas, such as text document analysis [1], natural language processing [2, 3], image representation and analysis [4], protein structure prediction [5], to name but a few.

In all such areas, it is important to be able to compare trees. Different methods have been posed in order to perform this comparison. Some of them are proposed to work with fully labeled trees. The method presented in this paper is to work with partially labeled trees, i.e., the trees labeled only at the leaves. Leaf labeled trees arise in the areas such as classification, biology, etc.

One way for tree comparison is to define a dissimilarity measure to determine how distant two trees are from each other. A number of dissimilarity measures for leaf labeled trees have been defined in the literature [6-12]. The Robinson-Foulds distance [6] is by far the most widely used dissimilarity measure which enumerates all edges in the trees and counts how many of the induced bipartitions differ between the two input trees. The quartet distance [7] is based on comparing all quartets (subsets of leaves of size four) in the trees and counts how many of the induced quartets differ between the two input trees. The path difference metric [8] is based on the comparison of the vectors of lengths of paths connecting pairs of taxa and quantifies the rate at which pairs of taxa that are close together in one tree lie at opposite ends in another tree.

The Robinson-Foulds distance is overly sensitive to some small changes in the tree. For example, just moving one leaf at the end of a caterpillar tree to the other end will result in a tree that has maximum distance to the original one; but the two trees are identical if the single leaf is removed. A caterpillar tree is a binary tree for which the induced subtree on the internal vertices forms a path graph.

II. PRELIMINARIES

Let \( G = (V, E) \) be an undirected graph with set of vertices \( V \) and set of edges \( E \). A tree is a connected acyclic graph. A leaf labeled tree is a tree whose leaves correspond to the taxa about which data was collected, while each nonleaf vertex is unlabeled and have degree at least 3. If every nonleaf vertex has degree equal to 3, the tree is said to be binary. Let \( T^n \) denote the set of binary leaf labeled trees on \( n \) taxa.

Cutting an edge \((a,b)\) from the tree \( T \) disconnects the tree, creates two smaller trees, and induces a
bipartition \( A, B \) of the set \( L \) of \( n \) taxa. We denote this
bipartition by an unordered pair \( A \mid B \). If \( \min\{|A|,|B|\} = 1 \), then \( A \mid B \) is trivial, otherwise it is
nontrivial. It is well known that the tree \( T \) can be
reconstructed from the set of the bipartitions it induces \([14, \text{Section 3.1}].\)

In each \( T \in T_n \), there are \( n \) pendant and \( n - 3 \)
internal edges. Let \( \beta(T) \) denote the set of bipartitions of \( T \), so \( |\beta(T)| = 2n - 3 \) and \( T \) has \( n \) trivial bipartitions.

The symmetric difference of sets \( X \) and \( Y \), denoted \( X \Delta Y \), is the set \((X - Y) \cup (Y - X)\).

**Definition 1.** The Robinson-Foulds distance \([6]\) between two trees \( T_1, T_2 \in T_n \) is defined as

\[
d_{RF}(T_1, T_2) = \frac{1}{2} |\beta(T_1) \oplus \beta(T_2)|
\]

(1)

Each bipartition \( A \mid B \) of the tree \( T_1 \) associates with a
binary vector \( V_e \) of length \( n \) : For any leaf \( i \), set \( V_e[i] = 1 \) if \( i \in A \), otherwise set \( V_e[i] = 0 \). Denote by \( BT_1 \) and \( BT_2 \) the sets of binary vectors associated with
the internal bipartitions of the trees \( T_1 \) and \( T_2 \), respectively.
The bipartition dissimilarity measure \( bd \) between
\( T_1 \) and \( T_2 \) \([13]\) is computed as follows:

\[
bd = \left( \sum_{a \in BT_1} \sum_{b \in BT_2} \min\{d_H(a,b), d_H(a,\bar{b})\} \right) + \left( \sum_{a \in BT_1} \sum_{b \in BT_2} \min\{d_H(b,a), d_H(b,\bar{a})\} \right) / 2,
\]

(2)

where \( d_H \) is the Hamming distance between the two
vectors \( a \) and \( b \), and \( \bar{a} \) and \( \bar{b} \) are the complements of
\( a \) and \( b \), respectively.

We now introduce the five types of commonly used
rearrangement operations on leaf labeled trees.

Each internal edge of a tree \( T \) associates four subtrees
which are attached to it. Nearest Neighbour Interchange (NNI)
means swapping two subtrees that are incident to
the same internal edge, as illustrated in Fig. 1.

![Figure 1. Trees \( T_2 \) and \( T_3 \) are obtained from \( T_1 \) by a single NNI operation. Circles are subtrees over sets of leaves \( A, B, C \) and \( D \).](image)

A Subtree Prune and Regraft (SPR) operation is
defined as follows. Delete an edge \( e = (u, v) \) of the tree
\( T \), get a new vertex \( w \) by subdividing an edge in the component
of \( T \setminus e \) that does not contain \( u \), add a new edge between
\( u \) and \( w \), and finally suppress all resulting
degree-two vertices. The operation is illustrated in Fig. 2.

A Leaf Prune and Regraft (LPR) operation is a special case of SPR in which the edge \( e = (u, v) \) is a pedant edge
(i.e., one of the vertices \( u \) and \( v \) is a labeled leaf.)

A Tree Bisection and Reconnection (TBR) operation is
similar to SPR and defined as follows. Delete an edge
\( e = (u, v) \) from \( T \), subdivide an edge in each component
of \( T \setminus e \), connect the two new vertices with an edge, and
finally suppress all resulting degree-two vertices. If a
component of \( T \setminus e \) consists of a single vertex, then the
added edge is attached to this vertex. The operation is
illustrated in Fig. 3.

A Leaf Label Interchange (LLI) operation just
exchanges the labels of two leaves and does not change the
topology of the tree \( T \).

For more details of the tree rearrangement operations
defined above, please see \([10, 14, 15]\).

**III. SENSITIVITY ANALYSIS**

We now investigate the sensitivity of bipartition
dissimilarity measure introduced in \([13]\) under the five tree
rearrangement operations defined in the last section.

For each binary vector \( a \in BT_1 \) associated with a
bipartition of the tree \( T_1 \), the best match \( M(a) \in BT_2 \)
is the binary vector that minimizes the dissimilarity value
between \( a \) and any binary vector in \( BT_2 \), i.e.,
Let \( N(T, \phi) \) be the neighborhood of \( T \) with respect to the operation \( \phi \), i.e., the set of trees that can be obtained by applying \( \phi \) once to \( T \). The gradient of \( \phi \) with respect to a measure \( d \) on \( T_n \), denoted \( G(\phi, d, T_n) \), is
\[
G(\phi, d, T_n) = \max \{ d(T_1, T_2) \mid T_1, T_2 \in T_n, T_2 \in N(T_1, \phi) \}.
\]

**Theorem 1.** [10] The gradients of the five tree rearrangement operations with respect to the Robinson-Foulds distance on \( T_n \) are as follows:

1. \( G(\text{NNI}, d_{RF}, T_n) = 1 \);
2. \( G(\text{SPR}, d_{RF}, T_n) = n - 3 \);
3. \( G(\text{LPR}, d_{RF}, T_n) = n - 3 \);
4. \( G(\text{TBR}, d_{RF}, T_n) = n - 3 \);
5. \( G(\text{LLI}, d_{RF}, T_n) = n - 3 \).

Sensitivity of a measure under an operation is defined to be the ratio of the gradient of the operation with respect to this measure to the diameter of it. Hence we concentrate our attention to the gradients of the five tree rearrangement operations with respect to bipartition dissimilarity measure on \( T_n \).

**Theorem 2.** \( G(\text{NNI}, bd, T_n) = \Theta(n^2) \).

**Proof.** Let \( T_2 \) be in the neighborhood of \( T_1 \) with respect to NNI operation. Clearly, \( BT_1 \) and \( BT_2 \) share \( n - 4 \) binary vectors, and only one binary vector is different in \( BT_1 \) and \( BT_2 \). Since each binary vector in \( BT_1 \) (\( BT_2 \)) has a dissimilarity value at most \( n/2 \) to \( T_2 \) (\( T_1 \)), it follows that \( bd(T_1, T_2) \leq n/2 \). On the other hand, it is easy to construct an example with \( bd(T_1, T_2) = n/2 \). Simply set \( |A| = |B| = |C| = |D| = n/4 \) in Fig. 1.

**Theorem 3.** \( G(\text{SPR}, bd, T_n) = \Theta(n^2) \).

**Proof.** Fig. 4 shows an example where one SPR operation leads to \( bd(T_1, T_2) = \Theta(n^2) \).

**Theorem 4.** \( G(\text{TBR}, bd, T_n) = \Theta(n^2) \).

**Proof.** The theorem follows from Theorem 3 since SPR is a special case of TBR.

**Theorem 5.** \( G(\text{LPR}, bd, T_n) = \Theta(n) \).

**Proof.** The lower bound is obtained by applying one LPR operation to a caterpillar tree, where one leaf at one end of the tree is moved to the other end. For the upper bound, let \( T_2 \) be in the neighborhood of \( T_1 \) with respect to LPR operation. Clearly, each LPR affects only two internal edges of \( T_1 \). Therefore, there is an internal edge \( e_1 \) that is in \( T_1 \) but not in \( T_2 \), and an internal edge \( e_2 \) that is in \( T_2 \) but not in \( T_1 \). Each bipartition of \( T_1 \) (\( T_2 \)) induced by an internal edge other than \( e_1 \) (\( e_2 \)) has a dissimilarity value at most 1, and the bipartition induced by \( e_1 \) or \( e_2 \) has a dissimilarity value at most \( n/2 \). Hence the upper bound is obtained.

**Theorem 6.** \( G(\text{LLI}, bd, T_n) = \Theta(n) \).

**Proof.** The lower bound is obtained by applying one LLI operation to a caterpillar tree, where the two leaf labels at the opposite ends of the tree are interchanged. For the upper bound, let \( T_2 \) be in the neighborhood of \( T_1 \) with respect to LLI operation. Clearly, one LLI affects only two leaves of \( T_1 \). Each bipartition of \( T_1 \) or \( T_2 \) has a dissimilarity value at most 2. Hence the upper bound is obtained.
small modifications in the tree, and thus is more robust than the Robinson-Foulds distance.

IV. CONCLUSIONS

The Robison-Foulds distance is the most widely used measure for comparing leaf labeled trees, but lacks robustness in the face of very small changes. The bipartition dissimilarity measure introduced by Boc et al. [13] dissimilarity refines Robinson-Foulds distance by comparing the quality of the tree bipartitions instead of their quantity. We presented some results in this paper on sensitivity analysis of bipartition dissimilarity measure. By showing that bipartition dissimilarity measure reacts more moderately to a single tree rearrangement operation than Robison-Foulds distance, these results reduce the uncertainty of bipartition dissimilarity and offer deeper insights into behavior of this measure. A possible direction of research is to study the sensitivity of other generalizations of Robinson-Foulds distance, e.g., the measures introduced in [11, 12] for rooted trees. It would also be interesting to design other measures for comparing leaf labeled trees.

REFERENCES