Similarity Measurement on Leaf-labelled Trees

H. De Meyer
Department of Applied Mathematics and Computer Science,
Ghent University,
Krijgslaan 281 (S9),
B-9000 Gent, Belgium
Hans.DeMeyer@rug.ac.be

B. De Baets
Department of Applied Mathematics, Biometrics and Process Control,
Ghent University,
Coupure links 653,
B-9000 Gent, Belgium
Bernard.DeBaets@rug.ac.be

S. Janssens
Department of Applied Mathematics, Biometrics and Process Control,
Ghent University,
Coupure links 653,
B-9000 Gent, Belgium
Saskia.Janssens@rug.ac.be

Abstract

We outline a method for measuring in an efficient way an overall degree of similarity between different leaf-labelled trees. In particular, we consider rooted trees, either unordered or ordered and not necessarily carrying the same labels. The trees to be compared are first represented in the form of matrices whose entries are in the interval $[0, 1]$ and the actual comparison strategy then relies on a one-parameter class of fuzzy similarity measures that are applied to the matrices. Various tests have been carried out to compare the new method with existing ones and to measure its robustness with regard to varying parameters.

Keywords: Similarity; Leaf-labelled tree; Tree comparison.

1 Similarity measures for ordinary sets

The task of measuring similarity amounts to the quantitative comparison of instances of a given complex mathematical structure. It is often performed in a variety of disciplines such as numerical taxonomy, ecology, chemistry, information retrieval, psychology, citation analysis, automatic classification: in fact, everywhere where the degree of similarity or dissimilarity between the objects under study plays a role.

A common method for comparing objects is to select an appropriate set of features and to create for each object a binary string where a bit being set on implies the presence of a particular feature. A degree of similarity between two objects is then obtained by directly using one of the many available measures for comparing binary strings. Binary strings can be represented by ordinary subsets in a finite universe X of dimension n (the feature-space), and the most frequently used similarity measures are the ones based on cardinalities of the sets involved. By far the most popular of these similarity measures is known as Jaccard's coefficient [5], which for given ordinary sets A and B is defined as:

$$S_J(A,B) = \frac{\#(A \cap B)}{\#(A \cup B)}. \qquad (1)$$

Clearly, this similarity measure, which is in terms of strings the ratio between the count of the bits on in both object A and object B and the count of the bits on in objects A or B, is symmetric and reflexive.

Another well-known symmetric and reflexive similarity measure is the so-called simple matching coefficient [8], defined as:

$$S_M(A,B) = \frac{\#(A \Delta B)^c}{n}. \qquad (2)$$

Herein Δ denotes the symmetric difference and in terms of binary strings this coefficient measures the fraction of matching bits in objects A and B.

In a previous paper [1], we have derived a class of 28 similarity measures for ordinary sets in the form of a rational expression solely based on cardinalities of the sets involved. A survey of similarity measures for ordinary sets frequently used in practice can be found in [7]. Here we only retain (1) and (2) as they are prototypical for two major classes of similarity measures: those which include the double zeros and those not including double zeros. Over the years there has been much discussion as to which type of measure to use. Here, we do not want to add to that discussion and instead consider the following one-parameter family of similarity measures:

$$S_\mu(A,B) = \frac{\#(A \cap B) + \mu \#(A \cup B)^c}{\#(A \cup B) + \mu \#(A \cup B)^c}, \qquad (3)$$

where the parameter μ can take any value in $[0, 1]$. With μ varying from 0 to 1, the similarity mea-

sure S_μ gradually moves from Jaccard's coefficient towards the simple matching coefficient.

Let us remind that a similarity measure S is called T-transitive if for any subsets A, B and C of the universe X, it holds that:

$$T(S(A,B),S(B,C)) \leq S(A,C)\,. \qquad (4)$$

In this definition, T is a t-norm, i.e. an increasing, commutative and associative binary operation on $[0,1]$ with as neutral element 1. The four main t-norms are the minimum operator M, the algebraic product P, the Łukasiewicz t-norm W (defined by $\mathrm{W}(x,y) = \max(x+y-1,0)$) and the drastic product Z (defined by $\mathrm{Z}(x,1) = \mathrm{Z}(1,x) = x$ and $\mathrm{Z}(x,y) = 0$ elsewhere). Also, M-transitivity implies P-transitivity, P-transitivity implies W-transitivity and W-transitivity implies Z-transitivity. Furthermore, following a convention from fuzzy set theory, a measure that is at the same time reflexive, symmetric and T-transitive, is called a T-equivalence [2].

It is well known that the minimum operator M is the only idempotent t-norm and that M-equivalences are in one-to-one correspondence with partition trees [10]. Also, the t-norms P and W are prototypical examples of continuous Archimedean t-norms and P- and W-equivalences are important in view of their correspondence to pseudo-metrics on the underlying universe [2].

In [1] we have shown that Jaccard's coefficient (1) and the simple matching coefficient (2) share the same transitivity property. More precisely, they are both W-equivalences, and it can be proven that the same holds for any member S_μ of the family (3).

2 Fuzzy similarity measures

If instead of binary vectors we have to compare vectors whose components can be scaled to the real interval $[0,1]$, the need emerges to extend the previous similarity measures for covering the latter case. In fact, in the same way as binary vectors are related to ordinary subsets of a finite universe X, vectors with components in $[0,1]$ can be related to fuzzy subsets of X. We will reserve the term fuzzy similarity measure for reflexive and symmetric binary fuzzy relations on $\mathcal{F}(X) = [0,1]^X$, whereas T-transitive fuzzy similarity measures, with T a t-norm, can still be called T-equivalences.

Let A, B denote fuzzy sets in a finite universe $X = \{x_1, x_2, \ldots, x_n\}$, and let

$$a_i = A(x_i)\,, \quad b_i = B(x_i)\,, \qquad (5)$$

with $a_i, b_i \in [0,1]$ for $i = 1,2,\ldots,n$. If we want to generalize the similarity measures for ordinary (crisp) sets to their fuzzy counterpart, we need fuzzification rules that define the cardinality of fuzzy sets and translate the classical set operations.

In this paper, we define the cardinality of a fuzzy set A as:

$$\#A = \sum_{i=1}^{n} A(x_i) = \sum_{i=1}^{n} a_i\,, \qquad (6)$$

also known as the sigma-count of A [9]. Furthermore, we shall use the following fuzzification rules, for any $x_i \in X$:

$$\begin{aligned} A^c(x_i) &= 1 - a_i\,, \\ A \cap B(x_i) &= \min\{a_i, b_i\}\,, \\ A \cup B(x_i) &= \max\{a_i, b_i\}\,, \\ A \Delta B(x_i) &= \mid a_i - b_i \mid\,. \end{aligned} \qquad (7)$$

This set of rules is certainly not unique (for a more general setting, see e.g. [3]), but it has the major advantage that its application to the similarity measures belonging to the family (3) preserves the properties of reflexivity and transitivity.

Making use of the fuzzification rules (7), we can define the following one-parameter family of fuzzy similarity measures:

$$\tilde{S}_\mu(A,B) = \frac{\sum \min(a_i,b_i) + \mu\,(n - \sum \max(a_i,b_i))}{\sum \max(a_i,b_i) + \mu\,(n - \sum \max(a_i,b_i))}\,. \qquad (8)$$

It can be proven that for any $\mu \in [0,1]$, $\tilde{S}_\mu$ is a W-equivalence. In particular, this is the case for the fuzzified version of Jaccard's coefficient ($\mu = 0$):

$$\tilde{S}_0(A,B) = \frac{\sum \min(a_i,b_i)}{\sum \max(a_i,b_i)}\,, \qquad (9)$$

and for the fuzzified version of the simple matching coefficient ($\mu = 1$):

$$\tilde{S}_1(A,B) = 1 - \frac{\sum |a_i - b_i|}{n}\,. \qquad (10)$$

3 Tree comparison

Fuzzy similarity measures serve as ready-made tools for comparing finite vectors or matrices whose elements take values in the interval $[0,1]$. On the other hand, in many scientific fields there is a need for methods to compare mathematical objects that do not possess a direct representation as an n-tuple. In systematic biology, for example, many studies have been devoted to methods and algorithms for comparing the branching structure of classification trees or dendrograms (see e.g. [11] and references therein). In particular,

Zhong et al. [11] have developed a general comparison methodology for different leaf-labelled trees. They use a similarity measure for ordinary sets, to compare first pairs of subtrees (which are simply reduced to their respective leaf node sets) and they further propose a so-called webbing matrix method to calculate the overall similarity of two leaf-labelled trees.

In the present paper we propose a method based on fuzzy similarity measures for comparing two leaf-labelled trees by means of an overall similarity coefficient. We further assume that the tree is an unordered rooted tree with internal nodes possessing at least two children.

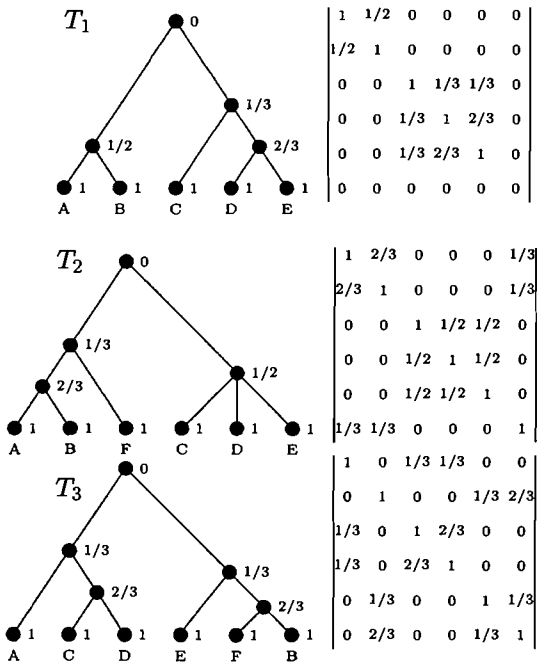

Figure 1: Three unordered trees to be compared and their corresponding matrices.

First, we attribute weights to the nodes of a tree as follows. Among all paths in the tree from the root to a leaf node and passing through a given node, we select a path with maximum length. If that path has length q (i.e. contains q edges) and the given node lies p edges away from the root, the weight p/q is attributed to that node. Note that the root has weight 0 and all the (labelled) leaf nodes have weight 1. This procedure of attributing weights to tree nodes is illustrated in Figure 1 on three example trees T_1, T_2, T_3. It is assumed that the common leaf node label set is $\mathcal{L}$ = {A,B,C,D,E,F}, though the leaf labelled F is missing in T_1. If we order the leaf node labels in a uniform but otherwise arbitrary way (here we choose the alphabetical order), we can unambiguously associate with any tree T a symmetric matrix $\hat{T}$ indexed by the ordered leaf labels. For any $X, Y \in \mathcal{L}$, the element $\hat{T}_{X,Y}$ equals the weight of the least common ancestor in the tree T of the two leaf nodes with respective labels X and Y. With any label that is not present in the tree, we associate a row and column of all zeros in the matrix. Finally, we remark that the tree is nothing but the partition tree associated with the matrix, and therefore the matrix itself is M-transitive. In Figure 1 the matrices corresponding to T_1, T_2 and T_3 are shown to the right of the trees.

An overall similarity coefficient between two trees T_i and T_j is obtained by first selecting a particular fuzzy similarity measure $\tilde{S}$ and then calculating $\tilde{S}(\hat{T}_i, \hat{T}_j)$, where the matrices are interpreted as fuzzy sets in the universe $\mathcal{L}^2$. For the previous example, we obtain with Jaccard's fuzzy similarity measure (9), the overall similarities: $\tilde{S}_0(\hat{T}_1, \hat{T}_2) = 25/36 = 0.694$, $\tilde{S}_0(\hat{T}_1, \hat{T}_3) = 17/43 = 0.395$, and $\tilde{S}_0(\hat{T}_2, \hat{T}_3) = 23/46 = 0.500$, results that are in agreement with those obtained in [11]. If, on the other hand, we use the fuzzy simple matching coefficient (10), we obtain: $\tilde{S}_1(\hat{T}_1, \hat{T}_2) = 97/108 = 0.898$, $\tilde{S}_1(\hat{T}_1, \hat{T}_3) = 82/108 = 0.759$, and $\tilde{S}_1(\hat{T}_2, \hat{T}_3) = 85/108 = 0.787$. For this small test set of trees, the order of the global similarities is independent of the choice of the fuzzy similarity measure within the family (3), but nonetheless it is clear that Jaccard's measure yields a much higher dispersion of the similarities.

The method described above can be extended for comparing rooted leaf-labelled trees that are ordered (i.e. branches emerging at internal nodes are ordered from left to right). First, we construct the (symmetrical) companion matrix as if the tree were unordered. Then, certain non-diagonal matrix elements acquire value 1 according to the following principle: for any $X, Y \in \mathcal{L}$, if from the branches emerging at the common ancestor node of X and Y, the one that contains leaf node X is at the left of the one that contains leaf node Y, then $\hat{T}_{X,Y} = 1$. Note that missing labels still correspond to a row and column of zeros solely. The non-symmetrical companion matrix of a given ordered leaf-labelled tree is M-transitive and therefore equivalent to a hierarchical tree of Hasse-diagrams, or Hasse-tree [6]. The arrows in the Hasse-diagrams are a graphical representation of the left-right relationship between branches.

The abovementioned example alone provides no decisive answer as to which fuzzy similarity measure to prefer for tree comparison. In an attempt to discriminate between the fuzzy similarity measures from family (3) and to verify whether the newly proposed method for tree comparison is sufficiently robust, an extensive series of tests has

been carried out.

In one of these tests, we considered the set of 26 different unordered leaf-labelled trees with four leaf nodes, which according to their topology are grouped into five classes. For the fuzzy similarity measures $\tilde{S}_\mu$ in (3) with $\mu = k/10, k = 0, 1, \ldots, 10$, we determined the (26×26)-matrix of overall similarities between the pairs of trees. Next, we calculated the M-transitive closure of the matrices and derived the associated partition tree [10]. The granularity of the clusters in the partition tree then gives an indication of how sensitive the fuzzy similarity measure $\tilde{S}_\mu$ is for comparing nearly similar trees. It has been observed that measure $\tilde{S}_0$ shows the finest granularity and that the coarsest granularity appears at the other end with measure $\tilde{S}_1$.

In another test setting, we made use of three sets of five trees each, essentially differing one from another with respect to the number of different labels occurring in the sets. On each of the three test sets the same procedure as explained before has been carried out. Hence, for each of the test sets and for each $\tilde{S}_\mu$ used, a partition tree was generated. It has been found that for the set of trees carrying more or less the same labels, the topology of the partition trees remains invariant with respect to μ. On the other hand, for the test set containing trees that share less labels, it has been observed that the topology of the partition tree changes significantly when moving from $\mu = 0$ to $\mu = 1$. Note that this is in agreement with what might be intuitively expected, since with increasing μ the similarity measure $\tilde{S}_\mu$ more predominantly takes into account the label context.

Yet another test aimed at verifying the robustness of the method against variations in the way of attributing weights to the internal nodes of a given tree. In fact, we considered two schemes for modifying the given tree such that all its leaf-nodes are at the same distance from the root. For a fixed similarity measure $\tilde{S}_\mu$, it was found that the different methods of attributing weights induce variations of the similarity coefficient within a 10% range. Also, changes in the order of similarities only occur when these similarities are initially very close. It follows that the proposed method is sufficiently stable with respect to these variations.

4 Conclusion

A new method for comparing ordered or unordered leaf-labelled trees has been introduced. The method proves to be robust and can be efficiently implemented to run in $\mathcal{O}(m^2)$ time where m denotes the cardinality of the label set.

It remains to be investigated how the proposed method can be generalized into a tool for comparing non-rooted trees and more general graphs as well. Also, the method could be adapted for obtaining instead of a degree of similarity a degree of inclusion between given trees, a coefficient that is relevant for retrieval algorithms in phylogenetic databases [4].

Acknowledgements

H. De Meyer is a Research Director of the Fund for Scientific Research - Flanders. This work is supported in part by the Bilateral Scientific and Technological Cooperation Flanders-Hungary BIL00/51.

References

[1] B. De Baets, H. De Meyer and H. Naessens, A class of rational cardinality-based similarity measures, *J. Comp. Appl. Math* (in press).

[2] B. De Baets and R. Mesiar, Pseudo-metrics and T-equivalences, *J. Fuzzy Math.* **5** (1997) 471–481.

[3] J. Fodor and M. Roubens, *Fuzzy Preference Modelling and Multicriteria Decision Support* (Kluwer Academic Publishers, Dordrecht, 1994).

[4] K. Herbert, H. Shan and J. Wang, Approximate searching in phylogenetic databases, in: *Proceedings of the CBGIST Symposium, Durham, USA* (2001) pp. 140–142.

[5] P. Jaccard, Nouvelles recherches sur la distribution florale, *Bull. de la Soc. Vaudoise des Sciences Naturelles* **44** (1908) 223–270.

[6] H. Naessens, B. De Baets and H. De Meyer, Generating Hasse trees of fuzzy preorder closures: an algorithmic approach, in: *Proceedings of the 1999 Joint Eusflat-Estylf Joint Conference, Palma, Spain* (1999) pp. 387–390.

[7] B. Sarker, The resemblance coefficients in group technology: a survey and comparative study of relational metrics, *Computers ind. Engng* **30** (1996) 103–116.

[8] R. Sokal and C. Michener, A statistical method for evaluating systematic relationships, *Univ. Kansas Sci. Bull.* **38** (1958) 1409–1438.

[9] L. Zadeh, Fuzzy sets, *Information and Control* **8** (1965) 338–353.

[10] L. Zadeh, Similarity relations and fuzzy orderings, *Inform. Sc.* **3** (1971) 177–200.

[11] Y. Zhong, C. Meacham and S. Pramanik, A general method for tree-comparison based on subtree similarity and its use in a taxonomic database, *BioSystems* **42** (1997) 1–8.