Four Metrics for Efficiently Comparing Attributed Trees

Andrea Torsello [1] Džena Hidović [2] Marcello Pelillo [1]
(1) Dipartimento di Informatica
Università Ca' Foscari di Venezia
Via Torino 155, 30172 Venezia Mestre, Italy
(2) School of Computer Science
University of Birmingham
Edgbaston, Birmingham, B15 2TT, UK

Abstract

We address the problem of comparing attributed trees and propose four novel distance metrics centered around the notion of a maximal similarity common subtree, and hence can be computed in polynomial time. We experimentally validate the usefulness of our metrics on shape matching tasks, and compare them with edit-distance.

1 Introduction

Graph-based representations have long been used with considerable success in computer vision and pattern recognition in the abstraction and recognition of objects and scene structure. As a consequence, the problem of how to measure the similarity or distance of pictorial information using tree abstractions has been a widely researched topic of over twenty years. Most metrics on trees found in the literature are defined in terms of edit-distance [8]. Zhang and Shasha [10] have investigated a special case of edit-distance which involves trees with an order relation among sibling nodes in a rooted tree, proving that it can be computed in polynomial time. Nonetheless, in the general case the problem has been proven to be NP-complete [11].

Recently, a new and more principled approach to the definition of distance measure between structures has emerged. In [1], Bunke and Shearer introduce a distance measure on unattributed graphs based on the maximum common subgraph and prove that it is a metric. In [9] Valiente extended this work introducing a bottom-up distance measure between trees. While this measure can be calculated in polynomial time both on ordered and unordered trees, it is limited to rooted and unattributed trees.

In this paper we propose four distance measures, two normalized and two non-normalized, for trees equipped with either symbolic or continuous-valued attributes. All our measures fulfill the properties of a metric, and can be computed in polynomial time.

2 Distance Metrics

Let T_1 and T_2 be two trees with node sets V_1 and V_2, respectively. Any bijection $\phi : H_1 \to H_2$, with $H_1 \subseteq V_1$ and $H_2 \subseteq V_2$, is called a *subtree isomorphism* if it preserves both the adjacency relationships between the nodes and the connectedness of the matched subgraphs.

Let σ be any similarity measure on the nodes of trees to be compared, possibly using the value of the attributes associated with the nodes, we define the similarity induced by the isomorphism ϕ as: $W_\sigma(\phi) = \sum_{u \in H_1} \sigma(u, \phi(u))$. The isomorphism ϕ is called a *maximum similarity subtree isomorphism* if $W_\sigma(\phi)$ is largest among all subtree isomorphisms between T_1 and T_2. The maximum similarity subtree isomorphism can be computed in polynomial time. For the rest of the paper we will omit the subscript σ when the node-similarity used is clear from the context.

Now, given a set S, a non-negative function $d : S \times S \to \mathbb{R}$ is a *metric* on S if the following properties hold for any $x, y, z \in S$.

1. $d(x, y) = 0 \Leftrightarrow x = y$ (identity and uniqueness)
2. $d(x, y) = d(y, x)$ (symmetry)
3. $d(x, y) + d(y, z) \geq d(x, z)$ (triangular inequality).

Furthermore, if the function satisfies $d(x, y) \leq 1$ it is said to be a *normalized metric*. In the rest of the paper, we shall assume that all similarity functions are of the form $\sigma(x, y) = 1 - d(x, y)$, where d is a normalized metrics.

For any two trees T_1 and T_2, we define the following metrics

$$d_1(T_1, T_2) = \max(|T_1|, |T_2|) - W(\phi_{12}) \quad (1)$$

$$d_2(T_1, T_2) = |T_1| + |T_2| - 2W(\phi_{12}) \quad (2)$$

$$d_3(T_1, T_2) = 1 - \frac{W(\phi_{12})}{\max(|T_1|, |T_2|)} \quad (3)$$

$$d_4(T_1, T_2) = 1 - \frac{W(\phi_{12})}{|T_1| + |T_2| - W(\phi_{12})} \quad (4)$$

where ϕ_{12} is the maximum similarity common subtree isomorphism between T_1 and T_2, and $|T|$ is the number of the nodes of tree T. The calculation of ϕ_{12} and, consequently, the optimal value of $W(\phi_{12})$, is going to be different for rooted and unrooted trees. Nevertheless, once the optimal similarity is at hand, the definition of the distance and the analysis of its properties are independent on whether the trees are rooted or not.

The first two metrics are unbounded and provide an absolute measure of dissimilarity between two attributed trees,

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

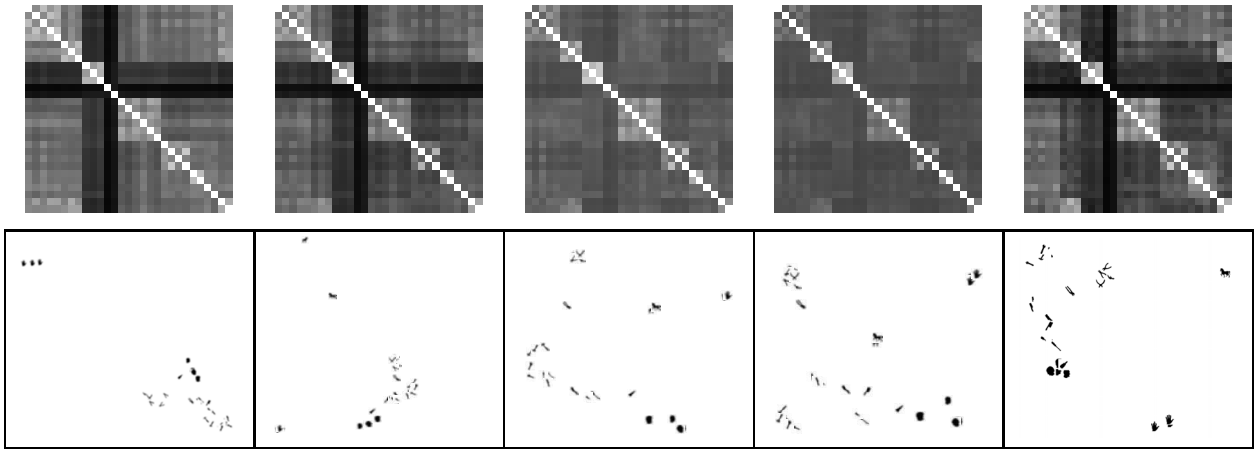

Figure 1. Top row: Distance matrices. bottom row: Multidimensional scaling. On each ro, left to right: $d1$, $d2$, $d3$, $d4$, and edit-distance.

in the sense that a particular perturbation on a tree "moves" it in tree-space by a distance which is independent of the whole tree mass. In some applications it is useful to have a metric which is bounded from above and provides a measure of relative dissimilarity. For these reasons, we have introduce the last two metrics, which are the normalized versions of the first two. For the proofs that all these measures are indeed metrics, we refer the reader to [7].

3 Experimental Results

We evaluated the new metrics on three different tree-based shape representations. The first is the shock tree representation used by Pelillo, Siddiqi and Zucker in [3], which is based on the shape's skeleton. skeletal points are grouped in so-called shock-classes, which distinguish between the cases where the local bitangent circle has maximum, minimum, constant, or monotonic radius. The groups are then abstracted using a rooted tree where node adjacency reflects the adjacency of shock-groups in the skeleton, and the distance from the root is related to the distance from the shape barycenter. Here, we used the same node-distances employed in [3], defined in terms of length, distance to the border, propagation speed, and curvature of the corresponding skeletal branch.

We compared our distance metrics with edit-distance. To approximate the edit-distance we used the relaxation labeling algorithm presented in [6] with the following costs: we defined the cost of matching node u to node w to be equal to the distance between their attributes, while the cost of removing any node to be equal to 1. Note that, with these costs, edit-distance is not normalized.

Our shape database contained 29 shapes from 8 different classes. To better visualize the distances we performed 2D multidimensional scaling (MDS) on the five matrices. Figure 1 shows, on the distance matrices obtained using our metrics and edit-distance and the resut of applying MDS. In the distance matrices, lighter colors represent lower distances while darker colors represent higher distances. As can be seen, the same block structure emerges in all five matrices. In particular, the main diagonal blocks are almost identical in all five cases, while the off-diagonal blocks present a wider variation. Essentially, the most significant differences among the five metrics are the dark bands clearly visible in the non-normalized matrices.

In order to assess the ability of the distances to preserve class structure, we performed pairwise clustering. In particular, we used two pairwise clustering algorithms: Shi and Malik's Normalized Cut [5], and Pavan and Pelillo's Dominant Sets [2]. Figure 2 shows the clusters obtained with Normalized Cut, displayed in order of extraction, while Figure 3 presents the clusters obtained with the Dominant Sets approach. While the performance of the clustering algorithms, on this shape recognition task, varied significantly, the dependency on the choice of the distance measure was

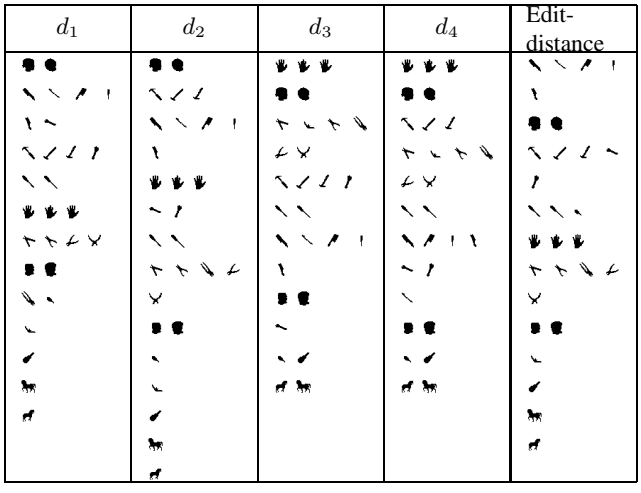

Figure 2. Clusters obtained with Normalized Cut in the first experiment.

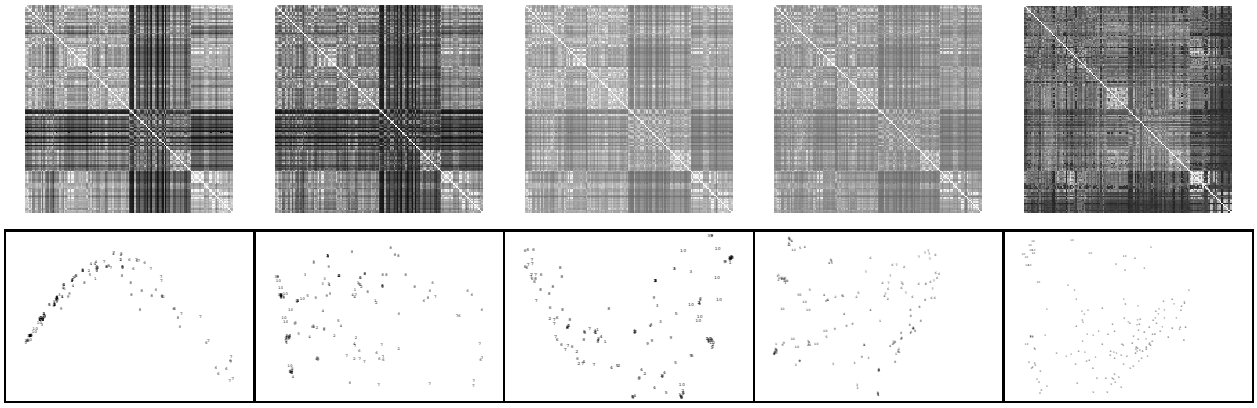

Figure 4. Second experiment. Top row: Distance matrices. Bottom row: Multidimensional scaling from the second experiment. For each row, left to right: $d1$, $d2$, $d3$, $d4$, and edit-distance. The numbers in each MDS plot represent class labels.

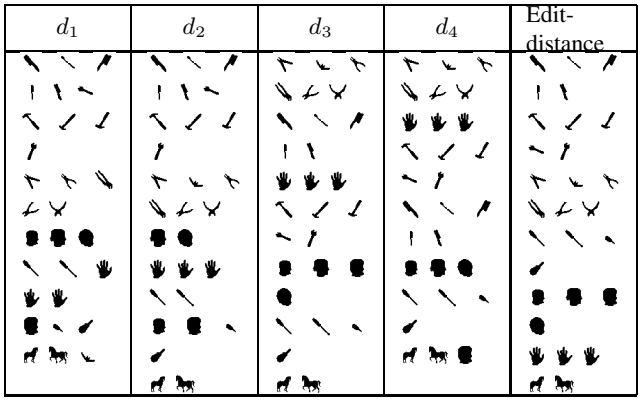

Figure 3. Clusters obtained with Dominant Sets in the first experiment.

less pronounced. Nonetheless, some differences can be observed. In particular, we notice how Normalized Cut exhibits a well-known tendency to over-segment the data, a behavior particularly visible on the non-normalized metrics d_1 and d_2. A particularly interesting example is from the classification of the two horses: the shock-tree representation of the horses have the highest average number of nodes of all shape classes, and they present the highest variation in terms of number of nodes. For this reason, as can be seen by looking at the MDS results, the non-normalized measures strongly separate the two instances, while the normalized versions are able to keep them close together. The clusters obtained with the Dominant Sets approach are much better, with our normalized metrics providing results almost identical to edit-distance.

Our second set of experiments used a larger database of shapes abstracted again in terms of shock-trees. Here, however, we attribute the trees with the proportion of the shape boundary generating the corresponding shock-group. The database consisted of 150 shapes divided into 10 classes of 15 shapes each, and presented a higher structural noise than the previous one. Here the node distance and node-matching cost for edit-distance was defined as the absolute difference between the attributes, while the node removal cost was the value of the attribute itself. With this edit costs edit-distance is a normalized metric.

Figure 4 shows the resulting distance matrices and MDS. All measures extract the same block structure, with non-normalized metrics showing the same off-diagonal dark bands as in the previous experiments. In particular, the metrics d_1 and d_2 do not distribute the shapes uniformly, but, rather, on a tight band along a curve. There are two reasons for this behavior: First, the metrics are inherently non-Euclidean, while MDS performs an "optimal" embedding on a Euclidean space; Second, the metrics d_1 and d_3 take the tree-similarity, which is smaller than the cardinality of the smallest tree, and balances it against the cardinality of the maximum tree. The other two proposed metrics balance the weight against the average cardinality, thereby providing a "tighter" measure.

Next, we applied the same clustering algorithms used previously. In order to assess the quality of the groupings, we used two well-known cluster-validation measures: the standard misclassification rate and the Rand Index. The latter measure is calculated as follows: We count the number of pairs of shapes that belong to the same class and that are clustered together and the number of pairs of shapes belonging to different classes that are in different clusters. The sum of these two figures divided by the total number of pairs gives us the Rand index. The higher the value, the better the classification. Table 1 summarizes the clustering results. The Dominant Sets method provides better results in this case as well, while the different metrics generate clusters with comparable validation measures.

The last set of experiments was performed on a tree representation of Northern Lights [4]. As in the previous experiments, the representation used is derived from the morphological skeleton, but the choice of structural representation was different from the one adopted for shock-graphs, and

	Misclassification rate		Rand index	
	Normalized Cut	Dominant Sets	Normalized Cut	Dominant Sets
d_1	25.3%	20.7%	90.1%	90.8%
d_2	28.7%	22.7%	90.1%	90.8%
d_3	23.3%	21.3%	90.3%	90.8%
d_4	22.7%	20.7%	90.5%	90.8%
edit	22.7%	24.0%	90.4%	90.8%

Table 1. Validation measures of clusters obtained in the second experiment.

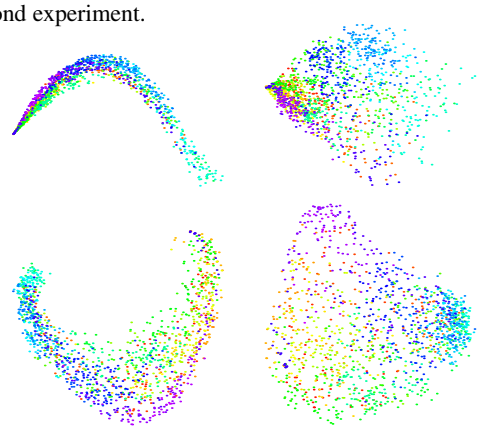

Figure 5. Multidimensional scaling of the distances obtained with our metrics from the third experiment. Top to bottom, left to right: $d1$, $d2$, $d3$, and $d4$.

the extracted trees tend to be larger.

The database consisted of 1440 shapes. Using our metrics we were able to extract the full distance matrices within a few hours, but it was infeasible to compute edit-distance on the entire database. For this reason, in order to be able to compare the results with edit-distance, we also performed experiments using a smaller database consisting of 50 shapes. The calculation of edit-distance, even on this reduced database, took a full weekend. Figure 5 displays the results of applying MDS to the distance matrices obtained with our measures. Here the hue of the point varies uniformly from red on the first shape to purple on the last. While there is no clear separation, there is a clear locality in shape-space of trees with similar indices.

In this case, we did not have the ground truth for the class memberships. We opted for a standard measure that favors compact and well-separated clusters: the Davies-Bouldin index. Let e_i be the average distance between elements in class i, and d_{ij} the average distance between elements in cluster i and elements in cluster j The Davies-Bouldin index is $DB = \frac{1}{c}\sum_{i=1}^{c} \max_j R_{ij}$, where c is the number of clusters and $R_{ij} = \frac{e_i+e_j}{d_{ij}}$ is the cluster separation measure. Clearly, lower values correspond to better separated and more compact clusters. Table 2 provides the values of the Davies-Bouldin index on the extracted clusters. As was the case with the previous experiments, the results are clearly comparable.

	Normalized Cut		Dominant Sets	
	50 trees	1440 trees	50 trees	1440 trees
d_1	0.0270	0.0159	0.0695	0.0057
d_2	0.0232	0.0135	0.0670	0.0055
d_3	0.0486	0.0165	0.0723	0.0074
d_4	0.0349	0.0155	0.0670	0.0068
edit	0.0232	—	0.0635	—

Table 2. Davies-Bouldin index of clusters obtained in the third experiment.

4 Conclusions

In this paper we have presented four novel distance measures for attributed trees based on the notion of a maximum similarity subtree isomorphism, and that can be computed in polynomial time. We have experimentally validated their usefulness by comparing them with edit-distance on three different shape recognition tasks. Our experimental results show that, in terms of quality, the proposed metrics compare well with edit-distance, their computation being, however, orders of magnitude faster.

References

[1] H. Bunke and K. Shearer. A graph distance metric based on the maximal common subgraph. *PRL*, 19:255–259, 1998.

[2] M. Pavan and M. Pelillo. A new graph-theoretic approach to clustering and segmentation. In *CVPR* , Vol. I, pp. 145-152, 2003.

[3] M. Pelillo, K. Sidiqi, and S. W. Zucker. Matching hierarchical structures using association graphs. *TPAMI.*, 21(11):1105–1120, 1999.

[4] M. Peura. Attribute trees in image analysis: Heuristic matching and learning techniques. In *ICIAP*, pp. 1160–1165, 1999.

[5] J. Shi and J. Malik, Normalized cuts and image segmentation, *TPAMI.*, 22(8):888–905, 2000.

[6] A. Torsello and E. R. Hancock. Efficiently computing weighted tree edit-distance using relaxation labeling. In *EMMCVPR* LNCS 2134, pp. 438–453, 2001.

[7] A. Torsello, D. Hidović, and M. Pelillo. Polynomial time metrics for attributed trees. Technical report CS-2003-19, University "Ca' Foscari" of Venice, Italy.

[8] W. H. Tsai and K.-S. Fu. Error-correcting isomorphism of attributed relational graphs for pattern analysis. *IEEE Trans Syst. Man Cybern.*, 9:757–768, 1979.

[9] G. Valiente. An efficient bottom-up distance between trees, in *Proc. Int. Symp. String Proc. Inf. Ret.*, pp. 212–219, 2001.

[10] K. Zhang and D. Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18:1245–1262, 1989.

[11] K. Zhang, R. Statman, and D. Shasha. On the editing distance between unordered labeled trees. *Inform. Process. Letters*, 42:133–139, 1992.