TreeSOM: Cluster Analysis in the Self-Organizing Map

Elena V. Samsonova[1], Joost N. Kok[2] and Ad P. IJzerman[1]
[1]Leiden/Amsterdam Center for Drug Research, [2]Leiden Institute of Advanced Computer Science
Leiden University, Leiden, The Netherlands
{elena.samsonova,joost}@liacs.nl, ijzerman@chem.leidenuniv.nl

Abstract - *We present the TreeSOM method and a set of tools to perform unsupervised SOM cluster analysis, determine cluster confidence and visualize the result as a tree facilitating comparison with existing hierarchical classifiers. We also introduce a distance measure for cluster trees that allows to select a SOM with the most confident clusters.*

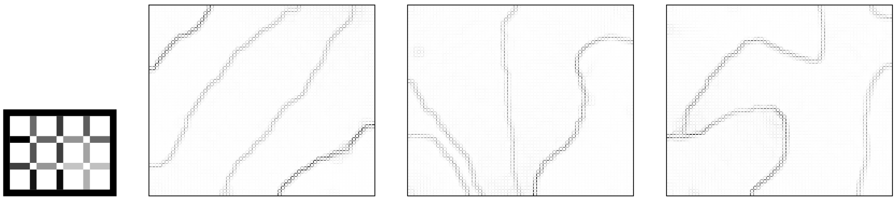
Key words - self-organizing map, hierarchical clustering, tree, reliability, visualization, tool

1 Introduction

Problems of ordering high-dimensional data in a small number of dimensions are frequently encountered. Such data are often noisy or incomplete, so that classical clustering methods such as linkage or multidimensional scaling, cannot be used. Kohonen self-organizing map (SOM) [1] can be used as a clustering method that addresses these issues. It is capable of mapping high-dimensional data onto a low-dimensional grid, placing similar data elements close together, forming clusters. However, different map initializations and input order of data elements, may result in different clusterings [2], as is illustrated in figure 1. Ideally, a large number of SOMs with varying random seed needs to be created, their clusterings analyzed, and only those clusters occurring in a majority of cases should be chosen.

In this paper we experiment with *TreeSOM* — an unsupervised method for cluster analysis and confidence testing for SOMs [3]. When used for clustering, SOM can be represented as a tree [4] allowing for easy comparisons with the outcomes of hierarchical classifiers widely used in various domains. Moreover, a tree representation allows to solve the problem of cluster confidence testing taking advantage of consensus tree building methods, developed and implemented independently of SOM (e.g., [5] which we use in this paper; other methods often produce similar results). A *consensus tree* represents an "average" of a set of trees with frequencies of occurrence of its branches compared to the set of all trees representing reliable clusters as subtrees. *TreeSOM* makes one further step in selecting one of the SOMs as the best representative of the consensus. Such combination of a consensus tree providing a cluster hierarchy, and a cluster map revealing spatial ordering of clusters, allows to view the clustering from different perspectives leading to reliable conclusions.

The rest of the paper is organized as follows. In section 2 we briefly outline the *TreeSOM* algorithm and describe the visualization methods. In section 3 we present a case study applying *TreeSOM* onto three example data sets. And finally, in section 4 we draw some general conclusions on SOM clustering tendencies.



(a) (b) (c)

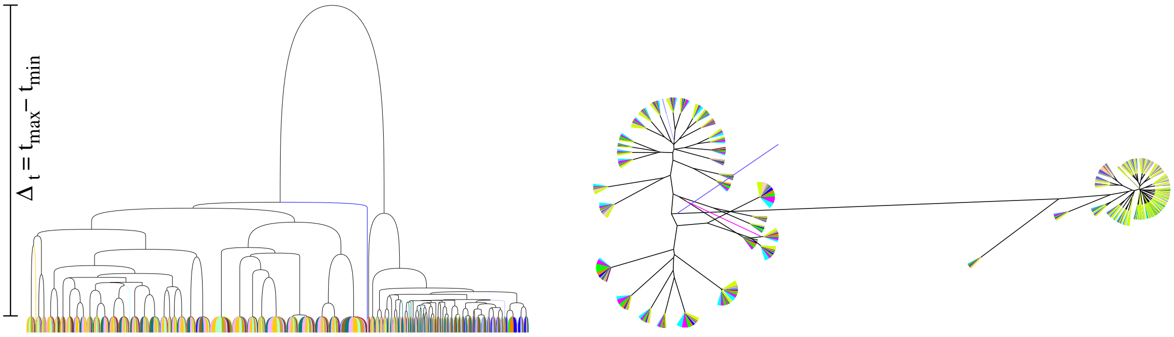
Figure 1: Examples of SOMs trained with identical parameters but different random initializations (*abalone* data set, see case study in section 3 for details and algorithm parameters). The SOMs reveal five sharply separated clusters on each map. The inset at left shows the details of this visualization. On maps (a) and (b) the clusters can be ordered linearly, but on map (c) it is no longer possible since the central cluster is adjacent to every other cluster on the map

2 SOM Clustering and Confidence Testing

Cluster Discovery When self-organizing maps are used for clustering, finding clusters on the SOM becomes a crucial task. Several fairly complex approaches have been developed, e.g., [4], [6]. The SOM representation in figure 1 is similar to the popular *umat* visualization [7]. Here the nodes are shown as large white boxes surrounded by edges shaded to indicate their lengths, with white standing for zero length, and black for the largest distance between any two adjacent nodes on the map. Using this representation, we can define a cluster as a group of nodes surrounded by an uninterrupted border of a given shade or darker representing distances equal to or greater than a given distance threshold. Thus, each node within a cluster must be connected to at least one other node within the same cluster with an edge that is shorter than the distance threshold. To determine the distribution of training data over the SOM clusters, each data item is assigned to the node that is most similar to it. Then, two data elements belong to the same cluster either if they are assigned to the same node, or if the corresponding nodes belong to the same cluster.

SOM as a Tree The SOM cluster analysis yields a series of nested clusterings that allows to represent cluster development as a tree. At each threshold in the clusterings series one or more clusters is split into several subclusters that is represented as a node in the tree. Thus, the sum of all branch lengths on the path from the root to the last node is the same for each path, and equals the difference between the maximal and the minimal threshold values: $\triangle_t = t_{max} - t_{min}$. The tips show the individual elements found in the corresponding clusters. Figure 2 shows a traditional hierarchical tree (a) and an alternative space-efficient "unrooted" or circular tree (b) representing the same hierarchy, popular in, e.g., bioinformatics. Each subtree is drawn within a sector proportional to the number of its leaves. The tree may be further "fanned out" for clarity. This representation is particularly beneficial for displaying "flat" hierarchies. Since in clustering analysis the emphasis lies upon the relations between the clusters rather than on an exact hierarchy, the root may be omitted.

Clustering Confidence Cluster trees of a large number of SOMs can be used with consensus tree methods to determine confident clusters represented as subtrees of a consensus tree. Confidence of each cluster is shown on the branch leading to the corresponding node with dash lengths, where a solid line (dashes of full length) stands for full confidence. Branch lengths in a consensus tree are not



(a) a hierachical representation

(b) a space-efficient circular representation

Figure 2: SOM as a tree (*abalone* data set, see case study in section 3 for details). Individual data elements are shown at tree tips. (a) A traditional hierarchical tree representation. (b) An alternative "unrooted" or circular tree representation. The root is found in the middle of the longest branch

related to clustering thresholds but reflect the distances between the corresponding nodes (clusters) and their siblings (nodes with the same parent): assume set A is split into $B_1, \ldots, B_n$, then the branch AB_j has length $|AB_j| = \frac{1}{n-1} \sum_{i \neq j} |B_i B_j|$. The distance between two sets P and Q equals the average distance between each element from P and from Q.

The Most Representative SOM Cluster confidence analysis leads to a final tree converting a spatial ordering of clusters inherent to a SOM, into a hierarchy. Although it is desirable in many cases, in many other cases it is not, as it lacks the information on proximity of clusters to one another. To solve this problem, *TreeSOM* uses a distance measure for cluster trees allowing to select an individual tree, and hence a SOM, as the best representative of the consensus by comparing nodes of the two trees. Together with a consensus tree, it provides full information on SOM clustering.

3 Case Study

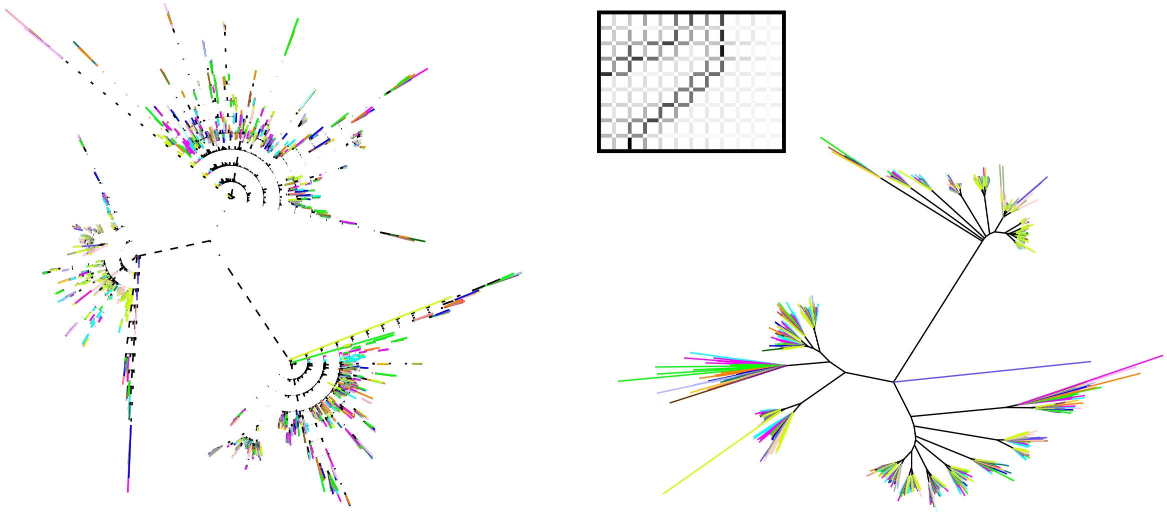
In this paper we illustrate the TreeSOM method on three test cases: distribution of abalone age groups (the *abalone* data set), distribution of protein localization sites in yeast (the *yeast* data set), and study of voting behavior of different countries during the yearly EuroVision Song Contests (the *song contest* data set). The former two data sets were obtained from the public database of the UCI Machine Learning Repository [8], and the latter one was kindly gathered and provided by Tim Cocx and is available from the authors upon request. These data sets were selected because of the difference in size, cardinality and attribute ranges (see table 1). In all cases Euclidean distance measure was used to train two SOM series (large and small maps) of 100 SOMs each.

All the SOMs were trained in two phases using Gaussian neighborhood and linear decrease of learning rate and radius. Phase-specific parameters were: (1) starting learning rate 0.2, starting radius 9, and a small iteration count (the actual iteration counts depend on the data set size and are listed within each test case); (2) starting learning rate 0.02, starting radius 3, and a large iteration count. The consensus tree algorithm [5] was used in PHYLIP [9] implementation. All the figures were generated by the TreeSOM software. Color versions and additional figures are available as supplementary material.

data set	size	cardinality	attribute ranges	contents
abalone	4177	8	1: 0, 1, 2 2-8: (0,2.9)	predicting the age of abalone from physical measurements [10]
yeast	1484	8	[0,1]	predicting localization site of protein in yeast from various characteristics [11]
song contest	38	38	[0,1]	discovering voting similarities among the countries (see below)

Here *size* is the number of vectors in the data set and *cardinality* is the number of attributes in each vector.
The *song contest* data reflects voting behavior of each country with respect to other countries over 47 years of the EuroVision Song Contest history (1957-2003). Each value v_{ij} represents the mean voting percentage of country i with respect to country j: $v_{ij} = \frac{p_{ij}}{n_j} / \sum_k \frac{p_{kj}}{n_j}$ where p_{ij} is the total number of points that country i awarded to country j, and n_j is the number of times that country j participated in the Contest.

Table 1: Data sets used in the case study



(a) 70x59 SOM cluster consensus tree; the maps are shown in figure 1

(b) 12x9 SOM cluster consensus tree

Figure 3: SOM cluster maps and consensus trees of the *abalone* data set

3.1 The Abalone Age Case

The SOM series used for this data set measure 70x59 and 12x9 units. The two phases counted 50,000 and 300,000 iterations respectively.
The consensus cluster tree of the large SOM (figure 3a) only shows three major clusters with fairly poor confidence as is evident from the short dashes used to draw the branches. In fact, confidence of the upper branch is far below 50% as dash length only constitutes a fraction of each dash period.

On the other hand, the cluster consensus tree of the small SOM (figure 3b) reveals a fully confident clustering. In fact, as is illustrated in figure 3b, all the SOMs converged to the same map producing identical cluster trees, a highly unusual but evidently not impossible situation. The map no longer shows five clusters as in figure 1, but just three, and the cluster tree reveals further cluster subdivision. Comparing these two SOM series, we can conclude that where a smaller SOM produces a confident clustering, a larger SOM presents several possible spatial arrangements of the training set, offering alternative data clusterings. Since SOM is a topology-preserving mapping, the various clusterings must be supported by some aspects of the data. Thus, a smaller SOM may be used to determine the most likely clusters, and a larger SOM to discover other possible relationships in the data.

3.2 The Case of Protein Localization in Yeast

The SOM series used for this data set measure 37x28 and 12x9 units. The two phases counted 10,000 and 200,000 iterations respectively.

The consensus tree of the large SOM (figure 4a) shows no large clusters, but many small ones with low confidence. However, the small SOM yields a consensus tree with clear clusters (figure 4b). Cluster maps and trees in figure 4c-d illustrate this result. The large maps in figure 4c do not show any clusters besides two small ones containing a mix of yeast localization sites. However, all maps contain areas predominantly "populated" by the data with the same localization site, even though they do not form strict clusters (see supplementary material). Cluster trees in this figure are drawn with uniform branch lengths emphasizing tree structure. They show three main clusters with very few subtrees indicating a "flat" hierarchy. However, these clusters appear to have different data such that their consensus fails to preserve them. It is also evident in large distances found between each individual SOM and the consensus, ranging between 0.85 and 0.9. The small SOMs in figure 4d also fail to show clusters, but their cluster trees reveal much more nested hierarchies, covering a wide range of distances to the consensus — from 0.09 to 0.987.

This test case allows to draw the same conclusions as for the abalone case, that smaller SOMs result in better defined and more confident clusters, whereas larger SOMs offer alternative views at data relationships.

3.3 The EuroVision Song Contest Case

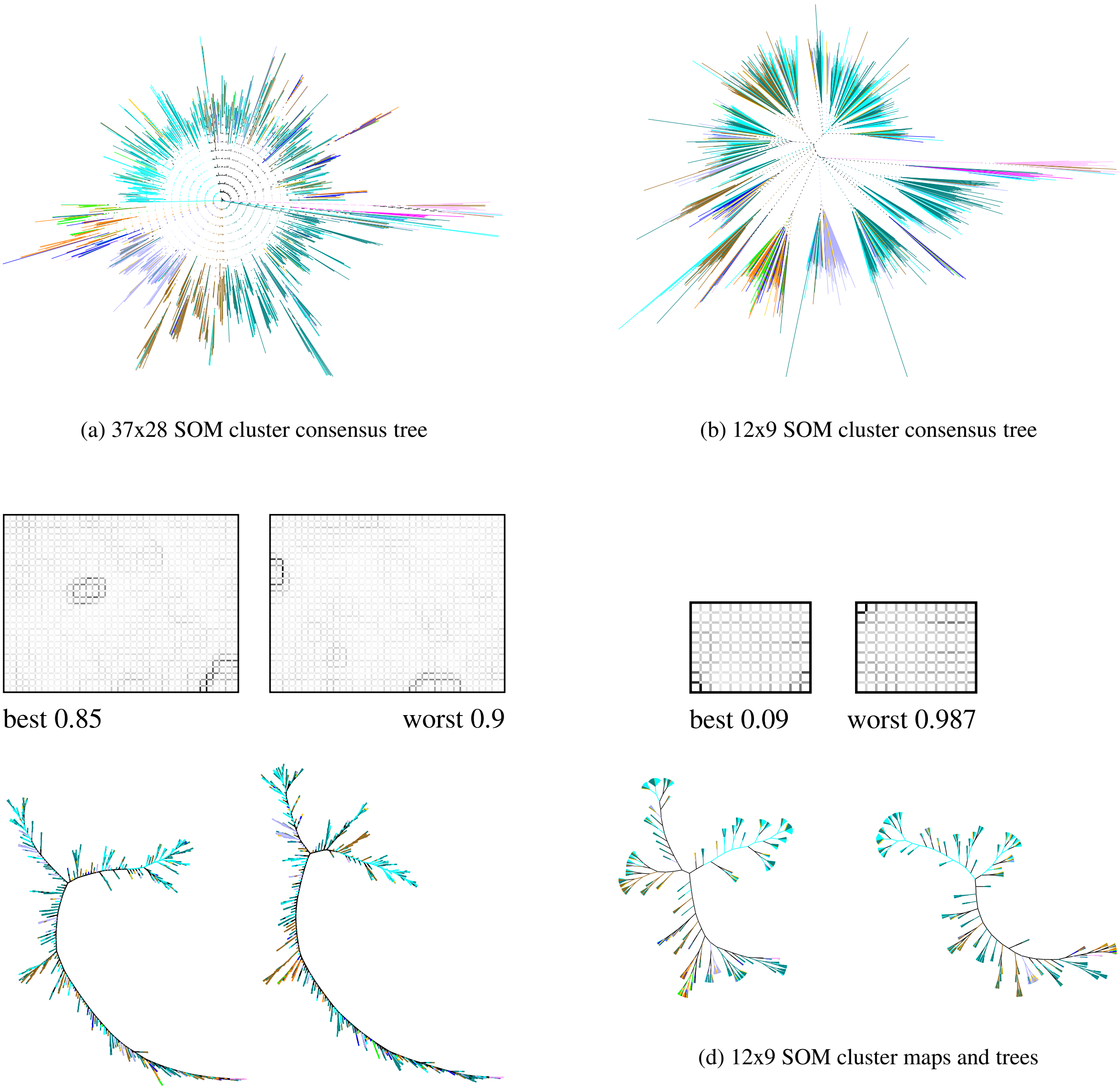
The SOM series used for this data set measure 12x9 and 4x3* units. The two phases counted 10,000 and 200,000 iterations respectively.

As shown in figure 5, the SOM cluster consensus trees are very similar to each other. In this case, just like in the previous test cases, the smaller SOM yields better defined clusters. Figure 6 shows the best and worst representative SOMs and their cluster trees in both series. They cover a comparable range of distances to the consensus, but the small SOMs show a much higher cluster separation than the large ones. Indeed, their maps also feature more very dark borders.

4 Conclusions

In this paper we demonstrated the use of *TreeSOM* on three data sets of different size and contents revealing that smaller SOMs tend to yield better defined and more confident clusters, whereas larger

*Starting radii from the standardized parameter set are too large for this particular SOM, therefore smaller radii were used: 3 and 2 respectively.



(a) 37x28 SOM cluster consensus tree

(b) 12x9 SOM cluster consensus tree

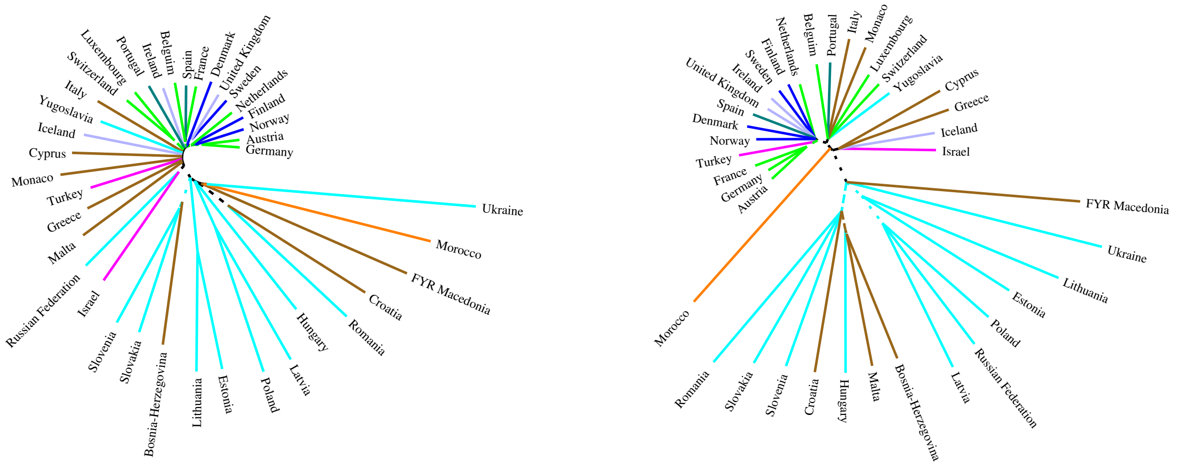
(c) 37x28 SOM cluster maps and trees

(d) 12x9 SOM cluster maps and trees

Figure 4: SOM cluster maps and trees of the *yeast* data set. All tree branches are drawn with the same length revealing tree structure

SOMs show relationships supported by only some aspects of the data. We do not believe that this effect should be dismissed as overlearning, since the data vector cardinality in the first two examples is significantly smaller than the number of data vectors. Such analysis enables the user not only to isolate confident clusters, but also to estimate cluster variability and explore other, possibly weaker supported relationships in the data that may be of relevance to the problem at hand.

Tools and supplementary material are available from:
http://web.inter.nl.net/users/Elena.Samsonova/resources.shtml#TreeSOM.



(a) 12x9 SOM cluster consensus tree

(b) 4x3 SOM cluster consensus tree

Figure 5: SOM cluster consensus trees of the *song contest* data set

Acknowledgments We thank Bas Groenendijk for his contribution to the implementation and testing of the software, the UCI Machine Learning Repository for the *abalone* and *yeast* data sets and Tim Cocx for the *song contest* data set. We are also grateful to the Netherlands Organisation for Scientific Research (NWO) for the financial support.

References

[1] T. Kohonen (1997), *Self-Organizing Maps*, **vol. 30** of *Springer Series in Information Sciences*, Springer, second edition.

[2] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen (1995), *SOM_PAK: the self-organizing map program package*, second edition.

[3] E. V. Samsonova, T. Bäck, J. N. Kok, and A. P. IJzerman (2005), Reliable hierarchical clustering with the self-organizing map. In *Proceedings of the 6th International Conference on Intelligent Data Analysis*. September 8-10, 2005, Madrid, Spain, in press.

[4] J. Himberg (2000), A SOM based cluster visualization and its application for false coloring. In *IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)*, **vol. 3**, p. 3587–3592.

[5] T. Margush and F. R. McMorris (1981), Consensus n-trees, *Bulletin of Mathematical Biology*, **vol. 43**, p. 239–244.

[6] J. Vesanto and E. Alhoniemi (2000), Clustering of the self-organizing map, *IEEE Transactions on Neural Networks*, **vol. 11**, p. 586–600.

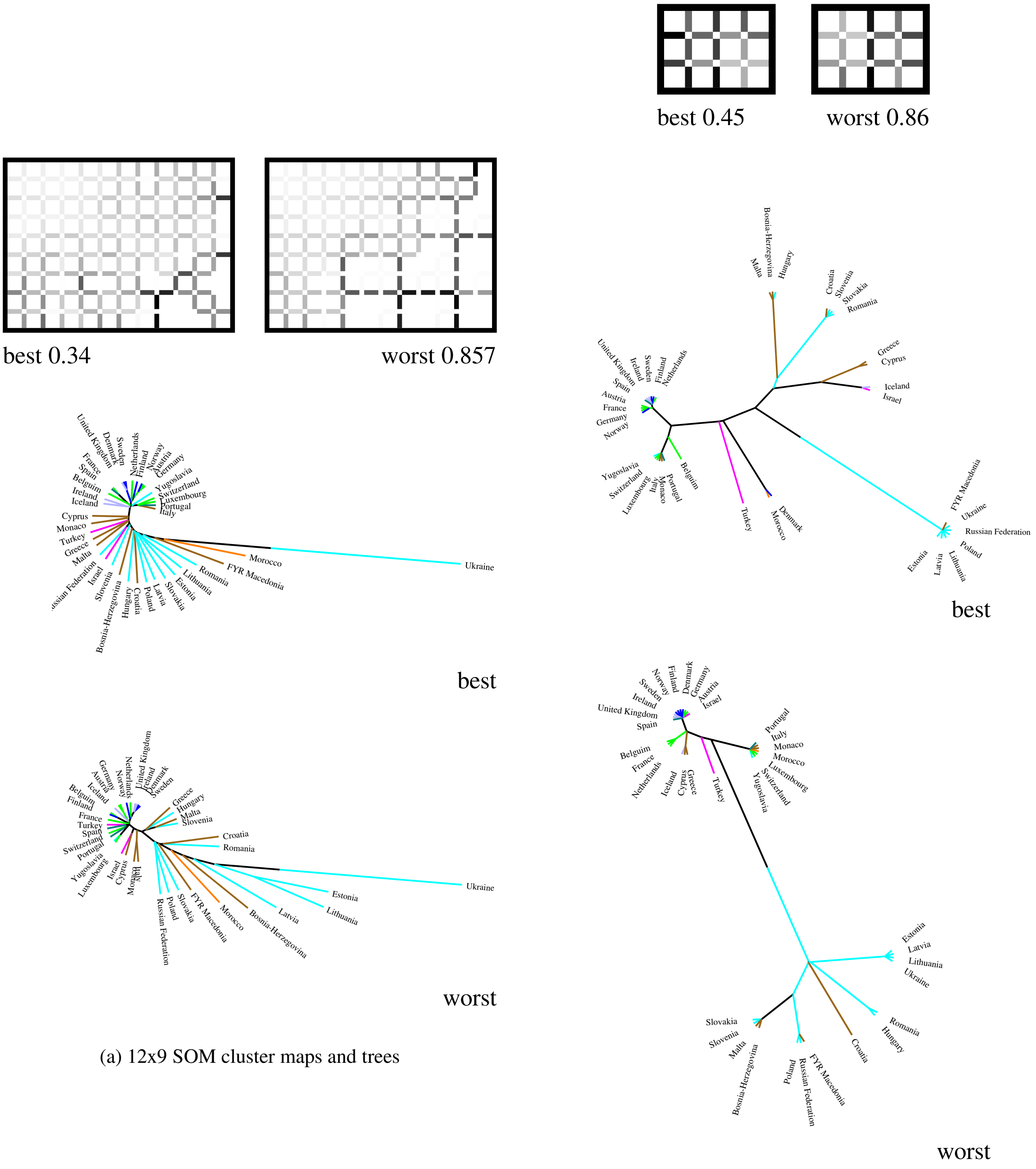
[7] M. A. Kraaijveld, J. Mao, and A. K. Jain (1992), A non-linear projection method based on Kohonen's topology preserving maps. In *Proceedings of the 11th International Conference on Pattern Recognition (11ICPR)*, p. 41–45, Los Alamitos, CA. IEEE Comput. Soc. Press.

[8] C. Blake and C. Merz (1998). UCI repository of machine learning databases, University of California, Irvine, Dept. of Information and Computer Sciences, *http://www.ics.uci.edu/∼mlearn/MLRepository.html*.

[9] J. Felsenstein (1989), PHYLIP – phylogeny inference package (version 3.2), *Cladistics*, **vol. 5**, p. 164–166, *http://evolution.gs.washington.edu/phylip.html*.

[10] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford (1994), The population biology of abalone (*Haliotis* species) in Tasmania. I. Blacklip abalone (*H. rubra*) from the North coast and islands of Bass Strait, Technical Report 48, Sea Fisheries Division, Hobart, Tasmania 7001, Australia. ISSN 1034-3288.

[11] P. Horton and K. Nakai (1996), A probablistic classification system for predicting the cellular localization sites of proteins. In D. J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith, editors, *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, p. 109–115, St. Louis, MO, USA. AAAI.



(a) 12x9 SOM cluster maps and trees

(b) 4x3 SOM cluster maps and trees

Figure 6: Cluster maps and trees of the *song contest* data set