Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

Minimum Spanning Tree Clustering of EEG Signals

Niina Päivinen[†] *and Tapio Grönfors*

University of Kuopio
Department of Computer Science
P.O.Box 1627
70211 Kuopio
Finland
[†]Tel. +358-17-16 2172, Fax: +358-17-16 2595
[†]e-mail: niina.paivinen@cs.uku.fi

ABSTRACT

In this study minimum spanning tree (MST) clustering is used to cluster EEG signals which contain epileptic seizures. Three strategies to get a clustering from the MST are presented and tested. As a reference, standard k-means clustering method is used on the same data and the results are compared. The results show that MST clustering is a promising method but further research is still needed.

1. BACKGROUND

The problem behind this study is the detection of epileptic seizures from pre-recorded electroencephalogram (EEG) signals using computational methods. Six features have been calculated from the data so that the measurements are presented as time-dependent vectors in six-dimensional space. The goal is to differentiate seizures from normal electrophysiological activity. To achieve this, pattern recognition methods are used.

Minimum spanning trees (MSTs) are often mentioned as a possible means to obtain clustering; see for example [1, 2]. Since the miminum spanning tree contains all the elements to be clustered and is connected, removing edges from the tree leads to a collection of disjoint subtrees which can be treated as a clustering of the original data. The minimum spanning tree problem is much discussed in computer science. It has many possible applications in different fields and it has originally risen from practical grounds [3]. The well-known Prim's and Kruskal's algorithms for constructing MSTs can be found, for example, in [4].

MST clustering is an unsupervised classification method. Clusters of different shape and size can be detected by MST-based clustering. However, there are some difficulties with this method. For example, the selection of the method for removing inconsistent edges from the spanning tree is somewhat problematic. Some encouraging results have recently been obtained with gene expression data [5].

2. MATERIAL AND METHODS

EEG potentials collected from ten different rats were used. Each recording consists of 200 seconds before seizure onset and 180 seconds after, totalling 380 seconds. The test database consisted of 6460 vectors of which 300 represented seizure.

A set of time-dependent features were calculated from the recordings and the best six seizure-detecting features among them were selected using statistical methods [6]. The selected features were the third Hjorth parameter (complexity), median frequency, spectral skewness [7], fractal dimension [8], approximate entropy [9], and maximal Lyapunov exponent [10]. All features were continuous-valued and their standardized deviates (z-values) [11] were used throughout the study.

The third Hjorth parameter, complexity, is a time-domain feature which depends on the variance of the original signal and its first two derivatives. Median frequency and spectral skewness are calculated in the frequency domain. Median frequency is the frequency which divides the area of the amplitude spectrum into equal halves and spectral skewness is the usual statistical skewness applied to the spectrum.

The remaining three features are nonlinear and they all are measures of complexity. Fractal dimension is the approximated dimension of the original signal embedded in the phase space and its values are between one and two. Approximated entropy measures the amount of regularity in the data – a smaller value indicates more regular behavior. Maximal Lyapunov exponent is a measure of chaos: positive value means that the system exhibits sensitive dependence on initial conditions.

Table 1. Cluster sizes: total (during-seizure)

Cluster	Strategy 1	Strategy 2	Strategy 3	k-means
1	392 (78)	321 (43)	351 (45)	42 (1)
2	1	18	42 (35)	23 (8)
3	1	42 (35)	1	44
4	1	13 (2)	1	48 (12)
5	1	1	1	44 (9)
6	1 (1)	1	1	48 (5)
7	1 (1)	2	1	18 (5)
8	1	1	1	89 (1)
9	1	1	1	44 (39)

Prim's algorithm was used to construct the minimum spanning tree. Euclidean distance function was used and in addition some other metrics were tested (correlation, Mahalanobis and cityblock) [12].

To get a clustering from the MST, a strategy for removing inconsistent edges is needed. Three different strategies for defining inconsistent edges were used. The first strategy is to remove k longest edges from the spanning tree to get a clustering to $k+1$ clusters. In the second strategy, an edge is defined as inconsistent if it is of "considerably different" length when comparing with its neighboring edges. For each node in the minimum spanning tree an average length of its edges, m, is calculated along with the standard deviation σ. If for some edge its length e satisfies $|e-m| > q\sigma$ the edge is treated as inconsistent. Here q is a positive constant (the values around 1.5 have been used in this study) [2]. The third strategy is similar to the second except that all the edges that lie at most two steps away from the current node are taken into account when calculating the mean and the standard deviation. The constant q had slightly greater values than in the second strategy.

The strategy for inconsistent edge removal is crucial with respect to the clustering result. Sometimes the globally best clustering (the first strategy) does not give satisfying results though it complies with the intuitive definition of clustering: the distance between clusters must be greater than the distances between the elements inside the clusters. The second and third strategies produce a local clustering, meaning that not all the edges of the tree are considered when defining inconsistent edges.

3. RESULTS

From the test database vectors, 400 (of which 20 % were representing seizures) were selected for each test. All three methods for removing inconsistent edges were tested and as a reference k-means procedure was used as well.

Table 1 contains the cluster sizes for a selected test case using all four clustering methods. The first number indicates the total number of elements in that cluster and the number in parentheses tells the number of during-seizure elements (when absent, none of the elements were during-seizure). Several numbers of clusters were tested and the results with nine clusters were selected to be shown here because of the favourable results with second and third strategies.

The first strategy, removing the longest edges from the MST, did not give satisfying results. Except for one cluster, all clusters were singleton and thus not very useful. Two of the singleton clusters contained a during-seizure observation. Removing the nodes in the singleton clusters from the MST and running the same procedure again did not change the general performance.

The second and third strategies performed a little better. Small clusters were present but none of them contained during-seizure elements. Removing these small clusters and clustering again using the third strategy led to a two-cluster situation where these clusters were the same as the original first and second cluster. When the standard k-means method was used to produce two clusters, the first cluster had 252 elements (of which 15 were during-seizure) and the second cluster had 141 elements (65 during-seizure). Generally k-means method produced better results when decreasing the number of clusters whereas MST clustering strategies behaved the opposite.

Table 2 shows means and standard deviations for the edge lengths: for all the original edges (distances between all the nodes), for the edges in the MST, and for the removed edges (all the three strategies).

Table 2. Edge length statistics

	mean	std
all	2.89	1.18
MST	0.80	0.37
Strategy 1	2.51	0.90
Strategy 2	0.77	0.26
Strategy 3	1.96	1.28

The MST clustering produced numerous small clusters (one to a few elements). This means that some leaves and tips of branches are somewhat further from the main tree than the other nodes. These elements might be considered as outliers.

As a conclusion one might say that the EEG signals were difficult to cluster with the MST but the standard

k-means method did not give satisfying results either. One might prefer the results from the MST, strategy 3, since it is natural to guess that the second cluster is the during-seizure -cluster. It seems that the k-means method tends to make all the clusters about the same size. Therefore it is really difficult to know which cluster contains the during-seizure data.

3.1 Clustering of the seizure data

A close examination of the results form the selected test case revealed that 31 during-seizure elements, which were defined as the "seizure core", were always in the same cluster. Thus the next step was to cluster only the 80 during-seizure elements. Strategy 1 was rejected because of its poor performance in the first tests.

The during-seizure elements were numbered according to table 3. The 80 seizure elements were originated from ten different recordings, eight elements from each recording. The six coodinates of each element were calculated using the data obtained during a one-second period of the original recording.

Table 3. During-seizure element numbering

signal number	200 s	201 s	$\cdots$	207 s
1	11	21	$\cdots$	81
2	12	22	$\cdots$	82
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
10	20	30	$\cdots$	90

Table 4. Cluster sizes: total (seizure core)

Cluster	Strategy 2	Strategy 3	k-means
1	41 (23)	46 (31)	13 (1)
2	32 (7)	26	8
3	2	1	14 (10)
4	1	3	15 (14)
5	1	2	13
6	2 (1)	1	7 (6)
7	1	1	10

When comparing the results shown in table 4, it can be seen that using the third strategy in MST clustering the seizure core elements lie all in the same cluster. Using Fisher's linear discriminant function, maximal Lyapunov exponent was discovered to be the main discriminating feature between the seizure core and the other seizure elements.

Fig. 1 shows a clustering of the seizure data based on the minimum spanning tree. Clusters were formed using strategy 3, see table 4. The edges were added to the tree in such a way that for each node its leftmost edge was added before the other edges, meaning that the edges were always added from left to right. The node numbers 11–90 mean element numbers and the boldfaces indicate the seizure core elements. The six edges marked with dashed lines are removed to get a clustering to seven clusters.

The elements 11–20 represent the time 200–201 seconds, meaning the first spike of the seizure. All these elements are situated in the branches of the spanning tree and none of them are in the seizure core.

When examining closer the seizure core elements, it can be noted that two sample recordings did not contain any of them. Fig. 2 shows two recordings of which the first one contains several seizure core elements and the second one contains none. Some elements have been labeled in the figures. Based on a visual inspection, it seems that the seizure core elements consist mainly of one dominant frequency. In the minimum spanning tree it means that the elements which have one dominant frequency are situated near the root and the elements where several frequencies are present are near the tips of the branches.

4. DISCUSSION

The results from this study are encouraging enough to continue working with the MST clustering. The method clearly has some good points which make it superior to k-means method in some cases. Although the results gotten with the EEG signals were not excellent they were not poor either. With some other database the results might be of a different quality.

Since the definition of the inconsistent edges is an essential part of the MST clustering, work in this area might lead to better results. Different distance functions may also play a significant role. There is always room for algorithm development.

ACKNOWLEDGEMENTS

We thank Jari Nissinen, Markku Penttonen and Asla Pitkänen from A.I. Virtanen Institute for Molecular Sciences, University of Kuopio, for the original annotated recordings. We also thank Seppo Lammi from Department of Computer Science, University of Kuopio, for his versatile statistical expertise.

REFERENCES

[1] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. John Wiley & Sons, Inc., New York, second edition, 2001.

[2] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern recognition*. Academic Press, Amsterdam, second edition, 2003.

[3] R.L. Graham and Pavol Hell. On the history of the minimum spanning tree problem. *Annals of the History of Computing*, 7(1):43–57, January 1985.

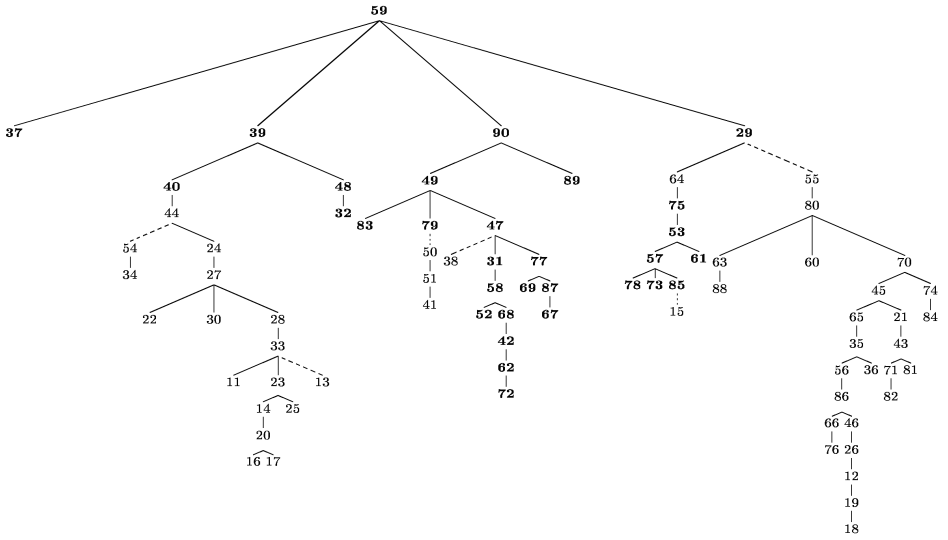

Fig.1. A minimum spanning tree

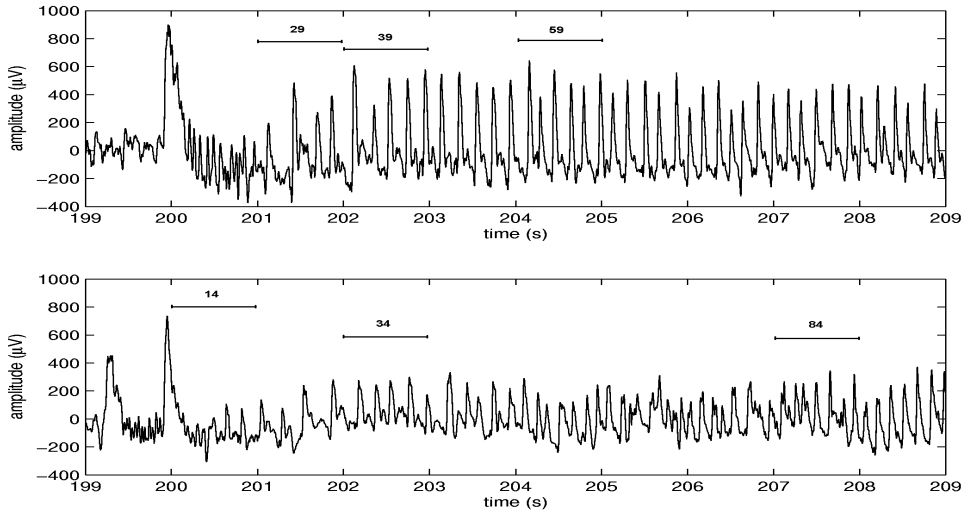

Fig.2. Two sample recordings with MST node numbers

[4] Alfred V. Aho, John E. Hopcroft, and Jeffrey D. Ullman. *Data structures and algorithms*. Addison-Wesley Series in Computer Science and Information Processing. Addison-Wesley, Reading, Massachusetts, 1983.

[5] Ying Xu, Victor Olman, and Dong Xu. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18(4):536–545, April 2002.

[6] Niina Päivinen, Seppo Lammi, Asla Pitkänen, Jari Nissinen, Markku Penttonen, and Tapio Grönfors. Epileptic seizure detection: A nonlinear viewpoint. *Manuscript submitted to Computer Methods and Programs in Biomedicine*, 2003.

[7] Rangaraj M. Rangayyan. *Biomedical signal analysis. A case-study approach*. IEEE Press Series on Biomedical Engineering. IEEE Press/Wiley, New York, 2002.

[8] A. Accardo, M. Affinito, M. Carrozzi, and F. Bouquet. Use of the fractal dimension for the analysis of electroencephalographic time series. *Biological Cybernetics*, 77(5):339–350, 1997.

[9] Steven M. Pincus, Igor M. Gladstone, and Richard A. Ehrenkranz. A regularity statistics for medical data analysis. *Journal of Clinical Monitoring*, 7(4):335–345, October 1991.

[10] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*. Cambridge Nonlinear Science Series 7. Cambridge University Press, Cambridge, 2002.

[11] P. Armitage, G. Berry, and J.N.S. Matthews. *Statistical methods in medical research*. Blackwell Science Ltd., Oxford, fourth edition, 2002.

[12] D.R. Wilson and Tony R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.