Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 921–930
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Clustering with a minimum spanning tree of scale-free-like structure

Niina Päivinen *

Department of Computer Science, University of Kuopio, P.O. Box 1627, FIN-70211 Kuopio, Finland

Received 10 September 2004
Available online 11 November 2004

Abstract

In this study a novel approach to graph-theoretic clustering is presented. A clustering algorithm which uses a structure called scale-free minimum spanning tree is presented and its performance is compared with standard minimum spanning tree clustering and *k*-means methods. The results show that the proposed method is a potential clustering procedure after some further analysis is done.

Keywords: Graph-theoretic clustering; Minimum spanning tree; Scale-free networks

1. Introduction

The goal of this study was the clustering of real-world data using methods based on graph theory.

A minimum spanning tree (MST) of a weighted graph connects all the given data points at the lowest possible cost (Sedgewick, 1984). An MST can be used in clustering: if the weights of the edges represent the distances between the data points, removing edges from the MST leads to a collection of connected components which can be defined to be clusters. It might be possible to use some other kinds of networks in clustering; in this study, one network model was considered.

Different irregular network architectures have been proposed in the literature. One of the oldest is the random graph model of Erdös and Rényi. It has been used in different application fields as an idealized model along with networks with regular structure. Two newer models are small-world and scale-free networks. Small-world networks lie somewhere between regular and random networks and the name analogy derives from the small-world phenomenon (Watts and Strogatz, 1998). Whereas the probability that a vertex has k links follows a Poisson distribution in random

* Tel.: +358 17 16 2172; fax: +358 17 16 2595.
E-mail address: niina.paivinen@cs.uku.fi

0167-8655/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.09.039

networks, scale-free networks follow a power law $P(k) \sim k^{-\gamma}$. The exponent γ has had values of $\gamma = 2.1$–2.4 for many real-world cases. (Strogatz, 2001).

A scale-free structure emerges in a network when it is growing by adding new vertices, and the new vertices are preferably attached to vertices which are already highly connected (Barabási and Albert, 1999). Both of these ingredients are necessary if a scale-free structure is wanted (Barabási et al., 1999). The situation becomes a little different if each vertex has some initial fitness which affects to the connection-making process. In this situation scale-free behavior can emerge also (Ergün and Rodgers, 2002).

2. Materials

The most important selection criterion for the data was the continuity of the attributes. In addition the selected datasets were all studied before. The standardized deviates of original attribute values were used in this study.

Three datasets from UCI Machine Learning Repository (Blake and Merz, 1998) were used along with a dataset consisting of intracranial EEG measurements from rats.

Fisher's iris plant dataset contains 150 instances and three (continuous) attributes measured from three different iris plant species. One class is known to be linearly separable from the other two classes, the latter are not linearly separable from each other.

The thyroid dataset has 215 instances, three classes and five attributes, all continuous. The aim is to predict the patient's thyroid class: "normal" (150 instances), "hyper" (35 instances) and "hypo" (30 instances).

The Pima Indians diabetes dataset contains 768 instances, eight attributes (not all of them continuous), and two classes: tested positive for diabetes and tested negative. From this database only 400 instances (152 positive, 248 negative) and six attributes (the continuous-valued ones) were used.

The EEG dataset consists of 400 samples, of which 80 represent epileptic seizure activity and 320 normal electrical activity of the brain. All six attributes were continuous and they were selected in such a way that they were not statistically strongly correlated with each other. This data has been studied before and it has been noted that it is difficult to cluster (Päivinen and Grönfors, 2004).

Fig. 1 shows a scale-free tree which consists of iris dataset elements. The blue vertices represent the species *Iris setosa*, orange vertices *I. virginica*, and the green vertices *I. versicolor*. Figs. 2–4 represent the same kind of structures from the other datasets.

3. Methods

Clustering algorithms based on graph theory can be used to detect clusters of different shapes and sizes, a feature that is not common among clustering methods. An example of this approach is a minimum spanning tree (MST) clustering (see Algorithm 1). The data must have well-separable clusters in order that they can be recognized with the MST clustering. On the other hand, the method does not need any parameters like the number of clusters or some other a priori information about the underlying data which can be considered as an advantage of the method (Theodoridis and Koutroumbas, 2003).

Algorithm 1. Minimum spanning tree clustering

procedure MST Clustering (V: set of data points)
  construct a fully connected graph G of V
    such that the edge weights are the distances between data points
  construct a minimum spanning tree T of G
  find all inconsistent edges of T
  remove the inconsistent edges to get a set of connected components
  define the connected components as clusters

The definition of the "inconsistent" edges causes problems in the MST clustering algorithm. One way to overcome these probelms is to extend the MST method by defining regions of influence of the vertex pairs (Theodoridis and Koutroum-

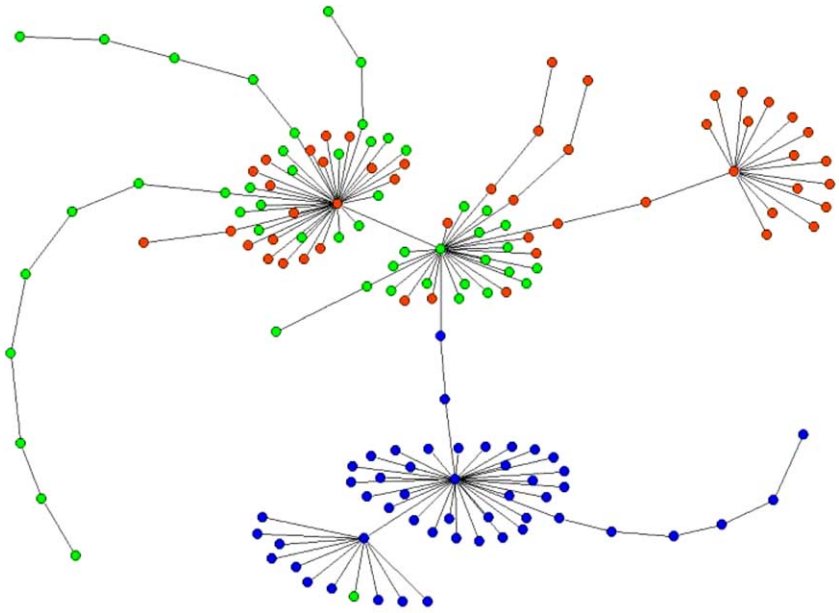

Fig. 1. A scale-free network from iris dataset (blue: *setosa*, green: *versicolor*, orange: *virginica*). (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

bas, 2003). This leads to an another kind of clustering algorithm which can be effective in some cases (Osbourn and Martinez, 1995).

Since scale-free networks have some vertices which are highly connected but the majority of them are not, this leads to a "natural" clustering: highly connected vertices can be thought to be "cluster centers", and all the vertices connected to this hub belong to the same cluster. For example, the iris plant Fig. 1 seems to contain five clusters and some long out-of-cluster branches. For reference, a MST created from the same iris data can be seen in Fig. 5.

Prim's algorithm for constructing a minimum spanning tree (Aho et al., 1983) was modified to produce a spanning tree which has a scale-free structure. This structure was named as *scale-free minimum spanning tree* (SFMST). In Algorithm 2, where the construction of an SFMST is presented, V means the set of vertices (data points), D is a square matrix containing the distances between the vertices, W is the weight matrix whose initial values are the reversed distances between the vertices, E is the set of edges not in the SFMST (initially all the possible edges between the vertices in V), S is a set containing the vertices of the SFMST, and P contains the edges of the SFMST. Throughout the algorithm the edges are identified as ordered pairs of the end point vertices of the edges.

The main differences when compared with original Prim's algorithm are the use of "reversed" distances and updating them during the construction process. Updating is not always necessary; namely, if a vertex has only one or two edges, it might not be reasonable to give it the reward for large connectivity. The updating is controlled by a constant which tells how many edges does a vertex have to have before its weight can get a bonus for connectivity.

Algorithm 2. Scale-free minimum spanning tree construction

procedure Scale-Free Minimum Spanning Tree (V: set of vertices)

set E to contain all the possible edges between the vertices in V

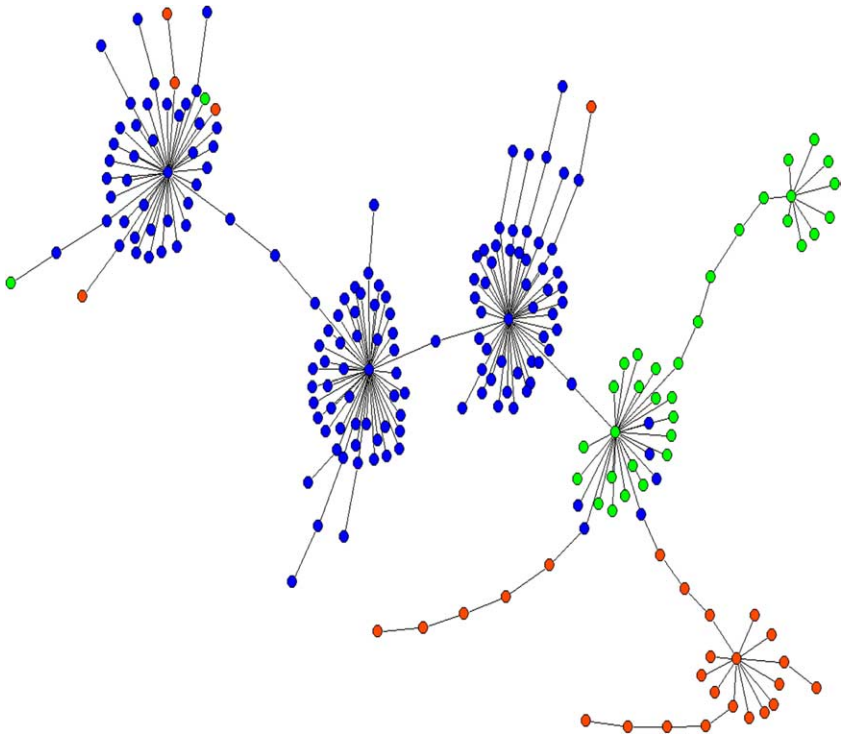

Fig. 2. Thyroid (blue: normal, green: hyper, orange: hypo). (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

calculate distance matrix D // distances between vertices
$W = \lceil \max(D) \rceil - D$; // reversed distances
select $(u,v) \in E$ which has the greatest weight
$S = \{u,v\}$; // vertices of SFMST
$P = \{(u,v)\}$; // edges of SFMST
$E = E \backslash \{(u,v)\}$; // unused edges
while $|S| \neq |V|$ **do**
  select $(u,v) \in E$, $u \in S$, $v \notin S$, which has the greatest weight
  $S = S \cup \{v\}$;
  $P = P \cup \{(u,v)\}$;
  $E = E \backslash \{(u,v)\}$;
  update weights in E if necessary
end

The line "update weights in E if necessary" is probably the hardest line of the algorithm to implement. Every link has to have an effect to the weight. Linear dependence—every link adds same constant to the weight—causes one vertex to gain nearly all the edges, resulting to a star-like network with one hub. The updated weight was defined to depend non-linearly on the number of links in such a way that the connectivity bonus increases slowly with the number of links and starts to decrease when the number of links is big enough. Specifically, if n is the number of links from a vertex v and it is greater than a pre-defined threshold value, then the weights of all possible links which have v as one end vertex are set to $w_{\text{new}} = w_{\text{old}} + nc^n$, where w_{old} is the old weight of the link and c is a constant, $0.5 < c < 1$.

The two constants, the minimal amount of edges needed before a bonus for connectivity is awarded and the multiplying constant in weight-updating, determine the final form of the spanning tree. The values used in this study were chosen based on numerical experiments. The threshold value for the minimum number of edges a vertex has to have in order to get the connectivity bonus was set at three, meaning that if a vertex has at

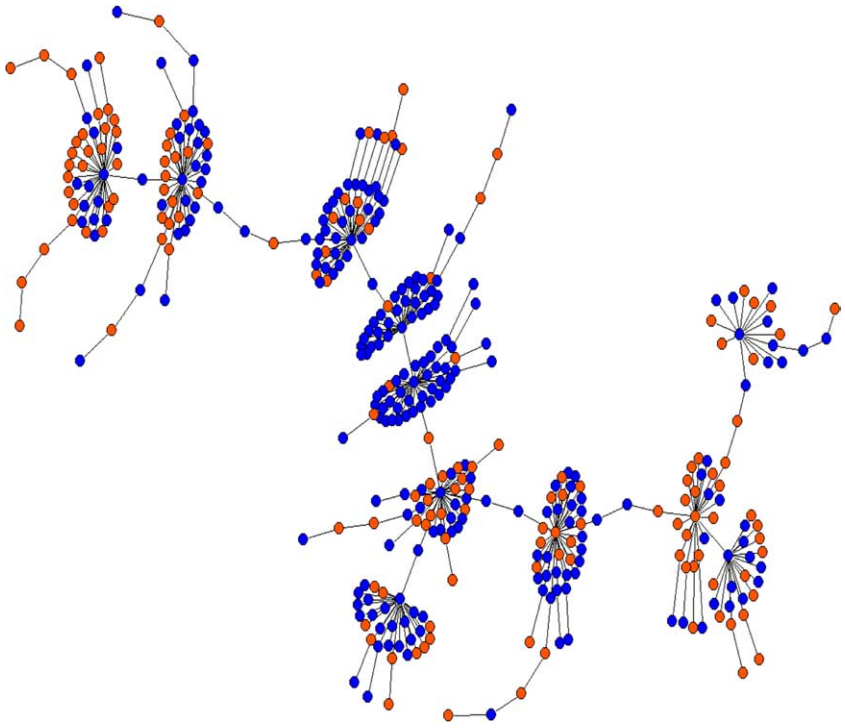

Fig. 3. Pima indians (blue: negative, orange: positive). (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

least three links, it is made more attractive in the linking process. The value $c = 1$ for a multiplying constant leads to a unipolar network with one hub because then the connectivity bonus depends linearly on the number of links. The smaller the value of c, the less hubs does the resulting network have; at value $c = 0.5$ the network looks nearly like a basic MST. The value $c = 0.9$ was used in this study; then a vertex has to have ten links before the connectivity bonus starts to decrease. This follows from the fact that the zero of the derivative of the weight-updating function is at $n = -1/\ln c$.

The proposed algorithm does not produce the usual scale-free network. The SFMST is not random since at each step the "fittest" vertex is added to the tree, and its structure is that of a spanning tree, meaning that it has no cycles and there is only one path between any two vertices. The construction makes the SFMST minimal with respect to the distances between the vertices when the scale-free property has to be maintained.

4. Results

For each dataset three different clustering methods were tested: SFMST, MST and k-means.

In both MST and SFMST methods Euclidean distance was used as the distance measure. A vertex was defined to be a hub if it had at least four links, and an SFMST cluster was defined to be a hub and all the vertices that connect directly to it. If two hubs were connected to each other or there was only one linking vertex between the hubs, they were defined to be in the same cluster. In addition, a branch was defined to be a chain of (one or more) vertices originating from a hub, in such a way that the vertex connected directly to the hub vertex does not belong to a branch. In MST clustering inconsistent edges were defined as follows. For each vertex in the minimum spanning tree, an average length of edges that lie at most two steps apart from the vertex, m, is calculated along with the standard deviation, σ. If for

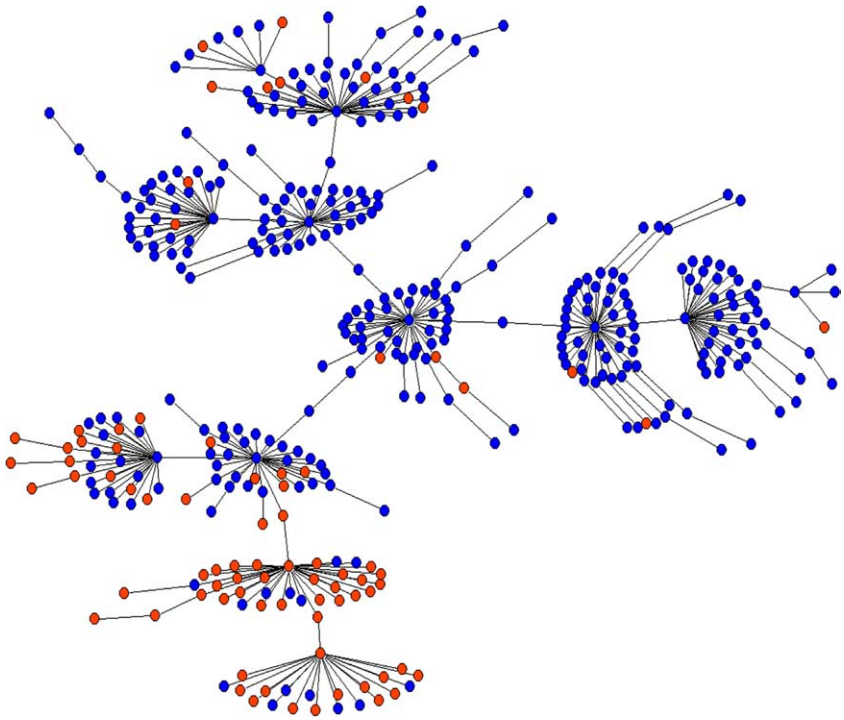

Fig. 4. EEG (blue: normal, orange: epileptic). (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

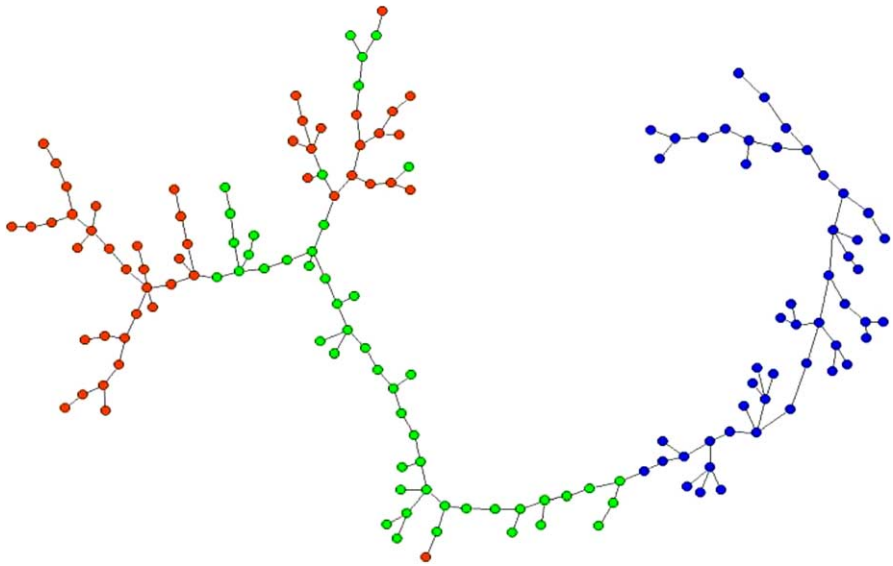

Fig. 5. An MST from iris dataset (blue: *setosa*, green: *versicolor*, orange: *virginica*). (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

some edge its length l satisfies $|m - l| > q\sigma$ for a pre-defined positive constant q, the edge is defined to be inconsistent and is thus removed from the MST. The constant q had values around two in this study. In addition, k-means procedure with squared Euclidean distance as the distance measure was used with two different values of k for each dataset.

For the iris plant dataset, the SFMST method produced three clusters and eight branches (see Table 1). The first cluster, C_1, which contains the majority of *setosa* vertices, has two hubs. The second cluster also has two hubs, one with *versicolor* vertex as the center and the other *virginica* vertex. The third cluster has one hub and contains only species *virginica*. The species *setosa* is known to be linearly separable from the other two, and it can be seen that the SFMST method produced one cluster and one branch, C_1 and B_1, which together contain all but one of these vertices. The remaining *setosa* vertex is connected to cluster C_1 and to the hub of cluster C_2 which means that it is assigned to cluster C_2. Cluster C_1 contains also one *versicolor* vertex. The MST method produces a tree in which *setosa* vertices are at the one end of the tree (see Fig. 5). The edge between *setosa* and *versicolor* vertices is the longest one in the tree, and so removing it separates *setosa* vertices from the other vertices. However, using the previously given definition for inconsistent edges, all *setosa* vertices are not in the same cluster. The k-means method separates *setosa* vertices well from the other vertices.

For the thyroid dataset the SFMST method produced four clusters and numerous branches

Table 1
Results for iris dataset

	Setosa	Versicolor	Virginia
SFMST			
C_1	44	1	0
C_2	1	35	28
C_3	0	0	17
B_1	5	0	0
B_2	0	1	0
B_3	0	0	2
B_4	0	0	1
B_5	0	7	0
B_6	0	2	0
B_7	0	4	0
B_8	0	0	2
$k = 5$			
C_1	0	19	2
C_2	0	2	27
C_3	22	0	0
C_4	0	29	21
C_5	28	0	0
MST			
C_1	1	45	30
C_2	0	0	1
C_3	0	4	7
C_4	0	1	0
C_5	0	0	8
C_6	0	0	1
C_7	0	0	3
C_8	36	0	0
C_9	13	0	0
$k = 3$			
C_1	33	0	0
C_2	0	46	50
C_3	17	4	0

Table 2
Results for thyroid dataset

	Normal	Hyper	Hypo
SFMST			
C_1	37	1	2
C_2	96	20	0
C_3	0	9	0
C_4	0	0	12
B_n	17	5	16
$k = 5$			
C_1	64	14	0
C_2	0	16	0
C_3	86	1	8
C_4	0	0	22
C_5	0	4	0
MST			
C_1	118	8	30
C_2	7	24	0
C_3	18	3	0
C_4	2	0	0
C_5	1	0	0
C_6	1	0	0
C_7	1	0	0
C_8	1	0	0
C_9	1	0	0
$k = 3$			
C_1	0	16	0
C_2	150	19	8
C_3	0	0	22

of which most were singleton-vertex. All the branch vertices B_n are presented as one row in Table 2. Cluster C_2 has three hubs, other clusters have one hub each. Clusters C_3 and C_4 contain only hyper and hypo vertices, respectively. Cluster C_1 has all kinds of vertices, mostly normal. Cluster C_2 has two hubs which have only normal vertices and one hub which has mostly hyper vertices connected directly to it. It seems that the hypo vertices can not be found using MST clustering. On the other hand, k-means succeeds to separate different kind of vertices quite well.

Results for the pima indians diabetes dataset are in Table 3. The SFMST produced five clusters and lots of branches. Clusters C_1 and C_3 have one hub, clusters C_2 and C_5 have two hubs and the last cluster C_4 has five hubs. This dataset was found not to be well-separated, the positive cases did not differ greatly from the negative cases at least with the used distance function. One must take into account that not all of the measured attributes were used since they were not continuous-valued; this may affect the clustering results.

The SFMST method for the EEG dataset, in Table 4, resulted in two big clusters and some branches; cluster C_1 has four hubs, cluster C_2 has seven. Epileptic vertices are mainly situated in cluster C_1 or attached directly to it. The MST method finds some of the epileptic vertices but most of them are blended in with the normal ones.

Table 3
Results for pima indians diabetes dataset

	Negative	Positive
SFMST		
C_1	9	7
C_2	15	27
C_3	26	12
C_4	133	38
C_5	32	37
B_n	33	31
$k = 9$		
C_1	27	23
C_2	13	26
C_3	8	9
C_4	56	20
C_5	33	38
C_6	3	1
C_7	23	22
C_8	80	7
C_9	5	6
MST		
C_1	224	138
C_2	2	0
C_3	9	1
C_4	1	0
C_5	1	0
C_6	1	1
C_7	0	1
C_8	1	1
C_9	9	10
$k = 5$		
C_1	102	9
C_2	72	43
C_3	19	39
C_4	47	52
C_5	8	9

Table 4
Results for EEG dataset

	Normal	Epileptic
SFMST		
C_1	56	57
C_2	217	13
B_n	47	10
$k = 9$		
C_1	37	1
C_2	84	3
C_3	46	2
C_4	46	9
C_5	38	11
C_6	39	4
C_7	3	42
C_8	16	8
C_9	11	0
MST		
C_1	306	45
C_2	7	35
C_3	1	0
C_4	1	0
C_5	1	0
C_6	1	0
C_7	1	0
C_8	1	0
C_9	1	0
$k = 5$		
C_1	9	44
C_2	101	0
C_3	51	19
C_4	55	3
C_5	104	14

With the k-means procedure one does not seem to find seizure vertices properly.

The comparison between the results of different methods is not a simple task. The k-means method assigns every data point to a cluster whereas the SFMST method produces, in addition with the distinct clusters, some branches which could be regarded to belong to the hub from which they are originated, or as separate clusters or even uncertain points. The results should be carefully reviewed with the application area on mind—then the interpretation of the structure of the SFMST may reveal a new viewpoint to the application. The results of the MST method are even more difficult to interpret because of the many singleton-vertex clusters it produces.

5. Discussion

One drawback of the presented procedure is that the algorithm is quite time-consuming; clearly algorithm development and analysis for faster computing times is needed. Slow computing times restrict the practical amount of data points. Based on this it can be argued if the link distribution of produced trees really follow a power law.

The dependence on distance function or an similarity measure is an open question; maybe non-continuous attributes can be used along with the continuous ones if the distance measure is selected accordingly. In addition some other similarity measure might suit better for clustering purposes.

When selecting the attributes to be used in clustering one should pay attention to their statistical properties. Strongly correlated attributes and greatly differing value ranges may harm the clustering performance and lead to confusing results. In this study, only the EEG dataset had statistically uncorrelated attributes; in the case of the other datasets, the attributes provided in the dataset were used as such.

The SFMST clustering method produced a few clusters which may contain more than one hub and several branches. The meaning of branches is still unclear, and the many hubs inside one cluster may mean that the cluster contains several subclusters.

So far only weighted non-directed trees have been used in clustering. Weights are of course essential, but the use of directed arcs and a general graph structure may bring forth some new information which can be used in clustering.

In conclusion it might be said that the proposed SFMST method still needs some development and analysis but the effort invested to it could turn out to be profitable.

Acknowledgement

The author wishes to thank professors Tapio Grönfors and Seppo Lammi, from Department of Computer Science, University of Kuopio, for continuous encouragement and advisement and for commenting the manuscript as well. Compliments for providing the EEG dataset go to Jari Nissinen, Markku Penttonen and Asla Pitkänen from A.I. Virtanen Institute for Molecular Sciences, University of Kuopio.

The network pictures in this document were created with Pajek—Program for Large Network Analysis (Batagelj and Mrvar, 2004).

References

Aho, A.V., Hopcroft, J.E., Ullman, J.D., 1983. Data structures and algorithms. Addison-Wesley, Reading, Massachusetts.

Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (5439), 509–512.

Barabási, A.-L., Albert, R., Jeong, H., 1999. Mean-field theory for scale-free random networks. Physica A 272 (1–2), 173–187.

Batagelj, V., Mrvar, A., 2004. Pajek—program for large network analysis. http://vlado.fmf.uni-lj.si/pub/networks/pajek/. Accessed 4th March 2004.

Blake, C., Merz, C., 1998. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html. Accessed 4th March 2004.

Ergün, G., Rodgers, G., 2002. Growing random networks with fitness. Physica A 303 (1–2), 261–272.

Osbourn, G.C., Martinez, R.F., 1995. Empirically defined regions of influence for clustering analyses. Pattern Recognition 28 (11), 1793–1806.

Päivinen, N., Grönfors, T., 2004. Minimum spanning tree clustering of EEG signals. In: Tanskanen, J.M., (Ed.), Proceedings of the 6th Nordic Signal Processing Symposium (NORSIG 2004), pages 149–152.

Sedgewick, R., 1984. Algorithms, corrected ed. Addison-Wesley, Reading.

Strogatz, S.H., 2001. Exploring complex networks. Nature 410 (6825), 268–276.

Theodoridis, S., Koutroumbas, K., 2003. Pattern recognition, second ed. Academic Press, Amsterdam.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393 (6684), 440–442.