PROCEEDINGS

of the Union of Scientists - Ruse

Book 5

Mathematics, Informatics and Physics

Volume 9, 2012

RUSE

The Ruse Branch of the Union of Scientists in Bulgaria was founded in 1956. Its first Chairman was Prof. Stoyan Petrov. He was followed by Prof. Trifon Georgiev, Prof. Kolyo Vasilev, Prof. Georgi Popov, Prof. Mityo Kanev, Assoc. Prof. Boris Borisov, Prof. Emil Marinov, Prof. Hristo Beloev. The individual members number nearly 300 recognized scientists from Ruse, organized in 13 scientific sections. There are several collective members too – organizations and companies from Ruse, known for their success in the field of science and higher education, or their applied research activities. The activities of the Union of Scientists – Ruse are numerous: scientific, educational and other humanitarian events directly related to hot issues in the development of Ruse region, including its infrastructure, environment, history and future development; commitment to the development of the scientific organizations in Ruse, the professional development and growth of the scientists and the protection of their individual rights.

The Union of Scientists – Ruse (US – Ruse) organizes publishing of scientific and popular informative literature, and since 1998 – the "Proceedings of the Union of Scientists- Ruse".

BOOK 5

"MATHEMATICS, INFORMATICS AND PHYSICS"

VOLUME 9

CONTENTS

NONHIERARCHICAL METHOD FOR CLUSTERING

Georgi Krastev

Angel Kanchev University of Ruse

Abstract: *The paper presents a method for nonhierarchical clustering. The method is called method of the minimal spanning tree. Two algorithms for constructing a minimal spanning tree are described. On the basis of the proposed algorithm, a program for nonhierarchical clustering has been developed, which possesses a huge variety of possibilities for visualizing the obtained results.*
Keywords*: Nonhierarchical Clustering, Minimal Spanning Tree*

INTRODUCTION

There has been an obvious trend to discover new algorithms for clustering a set of objects in recent years. Methods for linear clustering, for determination the density of the clusters, for estimating the significance of the formed clusters, and so on, were created. All that complemented the cluster analysis, whose major advantage lies in the spontaneous classification without the need of training series. The applications in the sphere of the technical diagnostics refer to the traditional research, connected with multi-dimensional analysis.

By the nonhierarchical methods for clustering, the objective is to divide the set of objects in K clusters in such a way, that the objects belonging to the same cluster are located closely to each other and the individual clusters are well separated. As all the K clusters are obtained simultaneously, the classification does not have a hierarchical character. Usually, the computational work by that type of classification is longer and more complex, and it is based on some heuristic algorithms, described in the literature on operational research, for example the procedures MASLOC, RELOCATE, ISODATA [1, 2, 3, 6, 7, 8]. The nonhierarchical clustering is used very rarely in the technical diagnostics, although it can also be very interesting in relation to some more peculiar sets of data.

ALGORITHMS FOR THE NONHIERARCHICAL CLUSTERING

One of the simplest algorithms for clustering is the following one. Let us imagine a group of seven islands, which we want to connect via bridges in such a way, that one can reach from every island to each other island in the group. As the building of bridges is connected with certain expenses, the objective that we are going to have, will be that the set of bridges, being constructed, must have a minimum price. We will accept that for each pair of islands we know, if it can be connected directly via a bridge and how much it costs. Figure 1 represents two possible configurations. The variant *a* is the better solution than variant *b* in view of the condition for the minimum length of the road network. This type of graphs are called trees, and the tree, about which the sum of the values of the separate links is the minimum one, is called a minimal spanning tree. Such a type of graphical representation is also the optimal solution of the problem.

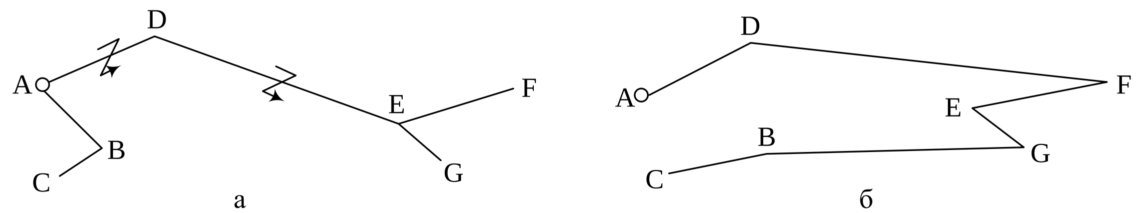


Fig. 1. A spanning tree for connecting objects according to the graph theory. The variant *a* is a minimal spanning tree. The arrows show a possible separation of clusters.

An interesting type of nonhierarchical clustering could be considered usingi graph theory. The method is called the method of the *minimal spanning (covering) tree.*

In that case, the objects, which must take part in the clustering, can be analyzed as points of a graph or a network. The links between the separate points have values, equal to the distance (Euclidean or some other measure of similarity) between them. The isolation of clusters happens with the aid of some approaches of the graph theory.

Expressed by means of the terms of the graph theory, the above task will look in the following way: It is given a non-oriented graph *G(V, E)*, in which the points of the multitude *V* are the islands, and the rib *(i, j)* exists when it is possible to build a bridge between the islands *i* and *j*. The weight *f(i, j)* determines its eventual price. We are looking for a covering (comprising) tree *T(V, D)* of *G* with a minimum amount of weights of those, which are involved in the *D* ribs.

A spanning tree *T* with such a characteristic is called a minimal spanning tree.

We will build a minimal spanning tree by choosing ribs consequently and by including them into the tree. The choice of ribs is made in such a way that the ribs chosen up to that moment establish the multitude *X*, which is a part of some minimal spanning tree.

Conditions: Let us assume that *X* is a multitude of ribs, which are a part of some minimal spanning tree *T* of graph *G* and *S* is such a multitude of points, that there is no rib from *X*, which connects a point from *S* with a point from *V-S*. Let us assume that *e* is a rib with a minimum weight among all the ribs, connecting a point from *S* with a point from *V-S*. Then, if we add the rib *e* to *X*, we will receive a multitude of ribs *X*', which is a part of some minimal spanning tree *T'*.

According to these Conditions, it follows that the minimal spanning tree can be obtained by means of a greedy algorithm.

ALGORITHM FOR BUILDING A MINIMAL SPANNING TREE

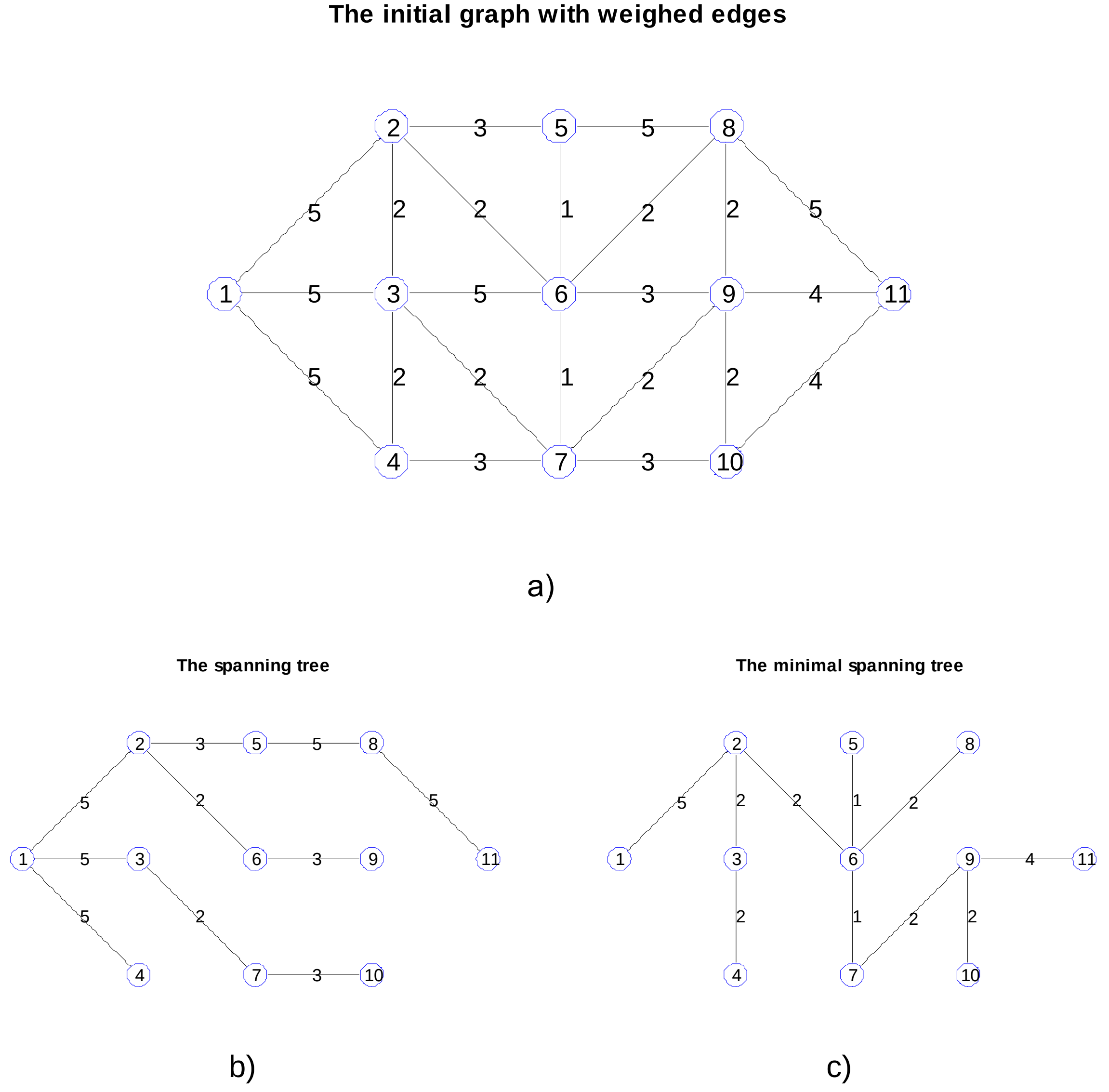
```
X = { };
while (|X| < n-1)
  { Choose a multitude of points S, such
     that the ribs of  X  do not connect to a point from S with a point from V-S.
     Find the lightest rib e, connecting a point from S with a point from V-S.
     Add the rib e to X.
   }
```

Choosing the multitude *S* in a different way, we receive different algorithms for building a minimal covering tree [4, 5].

By the Algorithm of Prim, the multitude of ribs *X* is a tree, and the multitude of points/peaks *S* contains the points of *X*. Initially *X* contains a random starting point from the graph. For finding the lightest rib between *S* and **V-S** the algorithm supports a priority tail, where all the points from *V-S* are located, which are adjacent to some point from *S*. The priority of any point/peak, according to which the pyramid is built, is the weight of the lightest rib from this point to any point from *S*. That reminds to the Algorithm of Dijkstra (where the clues which are used, are the lengths of the roads instead of weights of ribs). As in the Algorithm of Dijkstra, as well, each point has a previous one *prev[v],* which is the other end of the lightest rib from *v* to a point from *S*. The pseudo-code of the Algorithm of Prim is almost identical with the pseudo-code of the Algorithm of Dijkstra.

Another strategy is applied by the Algorithm of Kruskal. Instead of increasing of a single tree, this algorithm tries to put the possibly lightest rib into the tree on every step. Initially the ribs are ordered according to increasing lenghts. The multitude *X* is empty and the points of the graph are presented as trivial trees (without ribs). For each rib from the sorted list, it is consequently checked if its ends are in different trees. If that is so, then the rib is added to the multitude *X* and both trees are united into a common tree. The multitude *S* is in a non-declared way, the multitude of points of the one of the two uniting trees.

By implementing the Algorithm of Kruskal for presenting the separate trees, special data structures are used, designated for work with non-crossing multitudes. The obtained results are shown in Figure 2.



Number of edges on the spanning tree = 10
The total weight = 38
Number of edges on the minimal spanning tree = 10
The total weight = 23

Fig. 2. Nonhierarchical clustering via a minimal spanning tree

CONCLUSION

The discovering of important factors, creating the appropriate distribution of analytical data in the researched objects, is an important result both from the clustering analysis and from the methods for pattern recognizing, as well.

By the graph-theoretical formation of clusters other approaches are possible, as well, which are actually nonhierarchical.

The presented method for nonhierarchical clustering and its appropriate software are especially useful for large masses of experimental data.

REFERENCES

[1] Diday, E., Orders and overlapping clusters by pyramids, Multidimensional Data Analysis (eds. J. De Leeuw, W. J. Heisen, J. J. Meulman and F. Critchley), DSWO Press, Leiden, 201-234, 2006.

[2] Mfoumoune, E., Les aspects algorithmiques de la classification ascendante pyramidale et incrémentale, Thèse de doctorat, Université Paris 9 Dauphine, 2008.

[3] Rosch, E., Principle of categorization, E. Rosch, B. Lloyd edits, Cognition and Categorization , pp 27-48, Hillsdale, N.J.: Erlbaum, 2008.

[4] Breiman, L., J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Wadsworth, 2004.

[5] Noirhomme-Fraiture, M., Visualization of Large Data Sets: the Zoom Star Solution, Journal of Symbolic Data Analysis, vol 1, July 2002.

[6] Rasson, P. J., S. Lissoir, Symbolic Kernel discriminante analysis, Proc. NTTS'08 Sorrento, Italy, Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg), 2008.

[7] Michalski, S., R. E. Stepp, Learning from Observation: Conceptual Clustering, Machine Learning (eds. R. S. Michalski, J. G. Carbonell, and T. M. Mitchell), Springer-Verlag, 331-363, 2004.

[8] Wille, R. Knowledge Acquisition by methods of formal concepts analysis, Data Analysis, Learning symbolic and Numeric Knowledge, Diday edit, Nova Science Publishers, 2009.

CONTACT ADDRESS

Assoc. Prof. Georgi Nikolov Krastev, PhD,
Department of Computer Systems and Technologies, FEEEA
Angel Kanchev University of Ruse
7017 Ruse, Studentska Str. 8, Bulgaria
Phone: (++35982) 888 672
E-mail: gkrastev@ecs.uni-ruse.bg

НЕЙЕРАРХИЧЕН МЕТОД ЗА КЛЪСТЕРИРАНЕ

Георги Кръстев

Русенски университет “Ангел Кънчев”

Резюме: *В статията е представен метод за нейерархично клъстериране. Методът се нарича метод на минимално разклоненото (покриващо) дърво. Представени са два алгоритъма за построяване на минимално покриващо дърво. По предложения алгоритъм е разработена програма за нейерархично клъстериране с богати възможности за визуализиране на получените резултати.*

Ключови думи*: Нейерархично клъстериране, Минимално покриващо дърво*

ISSN 1314-3077

9 771314 307000