MDSonS: A Hierarchical Clustering Visualization Tool

José M. Martínez-Martínez*, Pablo Escandell-Montero, Emilio Soria-Olivas, José D. Martín-Guerrero, and Juan Gómez-Sanchis

IDAL, University of Valencia - Electronic Engineering Department
Av de la Universidad, s/n, 46100, Burjassot, Valencia - Spain
idal@uv.es
http://idal.uv.es/

Abstract. This paper proposes a new visualization approach for hierarchical cluster analysis, based on the Sectors on Sectors (SonS) visualization published in [7] by the authors. The method presented in this paper makes use of Multidimensional Scaling (MDS) to improve SonS visualization by means of representing distances between pairs of clusters. The proposed method is based on pie charts, and the main advantage is that it can extract all the existing relationships among centroids' attributes at any hierarchy level as well as representing the distances among all clusters. The methodology is tested on one synthetic data set and one real data set.

Keywords: Visual Data Mining, Hierarchical Clustering, Data Visualization

1 Introduction

Data visualization can greatly enhance our understanding of multivariate data structures, and hence cluster analysis and data visualization often go hand in hand. Results from partitioning cluster analysis can be visualized by projecting the data into a two-dimensional space. Cluster membership is usually represented by different colors and glyphs, or by dividing clusters into several panels of a trellis display [4]. In addition, silhouette plots [9] provide a popular tool for diagnosing the quality of a partition. Sometimes, the high dimensional data sets involve some level of hierarchical structure making difficult the use of the same visualization tools [4],[11]. Regarding hierarchical clustering, it is difficult to find methods for visualizing their results. Hierarchical cluster analysis is almost always accompanied by a dendrogram. Cutting the dendrogram at a specific level results in a clustering. Another visualization tool is the so-called treemap [10]. A treemap works by dividing the display area into a nested sequence of rectangles whose areas correspond to an attribute of the dataset. In some works, treemaps

* This work has been partially supported by the *Spanish Ministry of Economy and Competitiveness* with the Project RTA2014-00025-C05-05

are used to visualize hierarchical clustering [12], [6], [8], [1]. Also, other popular tools as convex cluster hulls or silhouettes are specific to clustering [4]. Despite the fact that the dendrogram is an excellent tool to determine the number of clusters in a given hierarchical data set, treemaps are very useful when visualizing the hierarchy, and convex cluster hulls and silhouettes give information about how the centroids partition the input space, and how well each object lies within its cluster, respectively; nevertheless, these techniques do not provide any information about the values of the attributes in each cluster centroid and the relationships among them. This drawback is solved by the visualization method proposed in this paper, which is also able to visualize hierarchical structures. In this paper a visualization technique for hierarchical cluster analysis, based on Sectors on Sectors (SonS) visualization, published by the authors in a previous work [7], is proposed. The method proposed in this paper is based on the use of Multidimensional Scaling, which makes possible to represent the distances among all cluster centroids'.

2 Sectors on Sectors (SonS)

Sectors on Sectors (SonS) is a visualization method that extracts visual information of data groups by representing the number of instances in each group, the value of the centroids of these groups of data and the existing relationships among the several groups and variables. This method is based on the well-known pie chart visualization. Each cluster is represented by a slice of a circle (pie sectors). The arc length of each pie sector is proportional to the number of patterns included in each cluster. By means of new divisions in each pie sector and a color bar with the same number of labels as attributes, the existing relationships among centroids' attributes of the different clusters can be inferred. Figure 2 represents the three steps followed to create the *SonS* visualization method; which are stated as follows:

Sectors on Sectors (SonS) is a visualization method that extracts visual information of data groups by representing the number of instances in each group, the value of the centroids of these groups of data and the existing relationships among the several groups and variables. Figure 2 represents the three steps followed to create the *SonS* visualization method; which are stated as follows:

1. **Division of one circle on several sectors depending on the number of clusters:** First of all the circle is divided into several pie segments or sectors corresponding to each cluster. The arc length of each sector is proportional to the number of patterns included in each cluster. The number of patterns belonging to each cluster is shown within parentheses. In this way, the significance of each cluster is easily recognizable (Figure 2, left).
2. **Division of the pie sectors depending on the number and the value of attributes:** After the first step, each sector is divided into as many subsectors as variables presented in the problem. The inner part corresponds to the first variable, and going outwards, the next variables are appearing.

Each one of these parts vary its radius, corresponding to the relative value of each variable, with respect to the sum of all of them[1]. That is, let X be a centroid corresponding to one cluster, so that,

$$X = \{x_1, x_2, \ldots, x_N\} \tag{1}$$

Then, the radius of each subsector (corresponding to each centroid attribute) is calculated as follows:

$$r_i = \frac{|x_i|}{\sum_{i=1}^{N} |x_i|}, i = 1 \ldots N \tag{2}$$

In this way the bigger the radius corresponding to each variable, the higher the weight of the variable and therefore, the more relevant the feature. This is a good method to identify the relevance of each variable within each cluster in a straightforward way (Figure 2, middle).

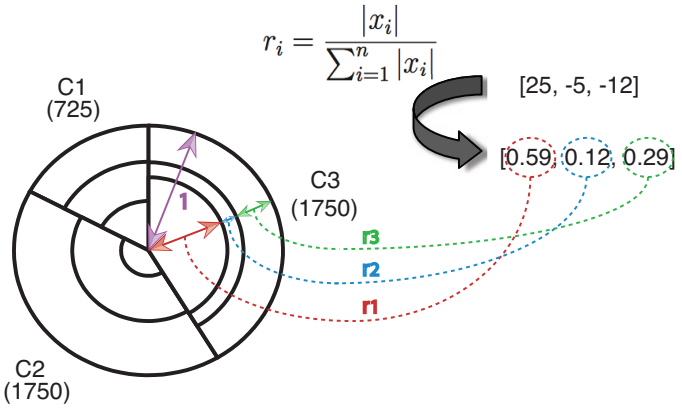


Fig. 1: Example of how the method computes the radii of the different subsectors in order to represent the relevance of the features in each cluster. After applying Eq. 2 to the vector $[25, -5, -12]$, the relevance of each attribute is obtained. The sum of all these "transformed" attributes is equal to 1.

In this example, it is shown one centroid with the values $[25, -5, -12]$. After applying Eq. 2 to this vector, the relevance of each attribute is obtained. Notice that the relevance for the first attribute (inner subsector) is 0.59,

[1] Each variable is scaled between $[0, 1]$ before carrying out the clustering in order to avoid a biased model. Moreover, the scaling makes that the relevance of each variable (represented by the size of the radius) is independent of its range.

0.12 for the second one (subsector in the middle) and 0.29 for the third one (outer sector) and that the sum of all these "transformed" attributes is equal to 1.

3. **Color coding for identifying the real value of features:** Attached to the graph, there is a color bar with the same number of column labels as variables (each column label for each variable). The mean value of the variables of each class (normally, the centroid) is codified by means of colors[2]. The value of the color for the first feature (inner subsector), is given by the first column label, the second feature by the second column label and so on. In this way, it is possible to know the exact value of each variable for each cluster centroid (Figure 2, right).

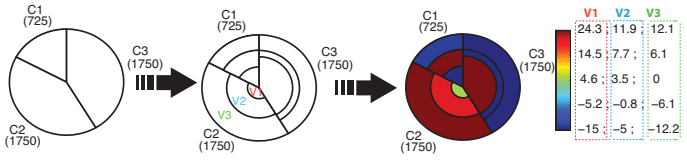


Fig. 2: The three steps followed to create the *SonS* visualization method.

3 SonS improvement: Multidimensional Sectors on Sectors (MDSonS)

The method proposed in this paper, is an improvement of the SonS technique [7], called Multidimensional Sectors on Sectors (MDSonS). The visualization method is different from that proposed in [7] due to the need of accommodating the information provided by Multidimensional Scaling (MDS) to the new visualization. Once decided the structure of the clustering (number of clusters in each hierarchical level), the visualization graph is produced in three steps for each hierarchy level (see Fig. 3):

1. **Representation of the different clusters and their size:** First of all, as many circles as clusters are drawn. The area of each circle is proportional to the number of patterns included in each cluster and the distance among circles is proportional to the distance among clusters' centroids. The distances among centroids are computed by MDS. MDS produces a representation of the similarity (or dissimilarity) between pairs of objects in a multidimensional space as distances between points of a low-dimensional space [2]. The number of patterns belonging to each cluster is shown within parentheses (Fig. 3, left).

[2] This is automatically extensible to other measures.

2. **Division of the circles depending on the number and the value of attributes:** In order to represent the value of the attributes for each cluster centroid, each circle corresponding with each cluster is divided into several sectors, which correspond to each variable. The first variable is the one that starts with a vertical line in the top middle of the circle, and the rest of the variables is appearing sequentially counter clockwise. The arc length of each sector corresponds to the relative value of each variable, respect to the sum of all of them[3]. In this way, the bigger the arc length of a given variable, the more relevant the variable. With this method it can be identified the relevance of each variable, within each cluster in a straightforward way. (Fig. 3, middle)
3. **Color coding for identifying the real value of features:** Attached to the graph, there is a color bar with the same number of labels as variables (each label for each variable). The mean value of the variables of each class (normally, the centroid) is codified by means of colors[4]. The value of the color for the first feature, is given by the first column label, the second feature by the second column label and so on. In this way, it is possible to know the exact value of each variable for each cluster centroid (Fig. 3, right).

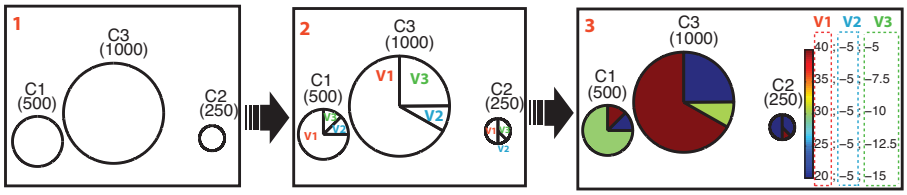


Fig. 3: The three steps followed to create the MDSonS visualization method.

The description for the first hierarchy level can be extended to the rest of levels. The main advantage of the proposed visualization technique is that it is possible to observe relationships among different variables in the same cluster and relationships among the same variables in different clusters, in the different levels of the hierarchy; but specially, what is remarkable in comparison to SonS is the representation of the distances among clusters centroids.

[3] Each variable is standardized to zero mean and unit variance before applying the clustering algorithm in order to avoid a biased model. Moreover, the standardization makes that the relevance of each variable (represented by the size of the radius) is independent of its range.

[4] This is automatically extensible to other measures

4 Analysis and results

4.1 Data Set

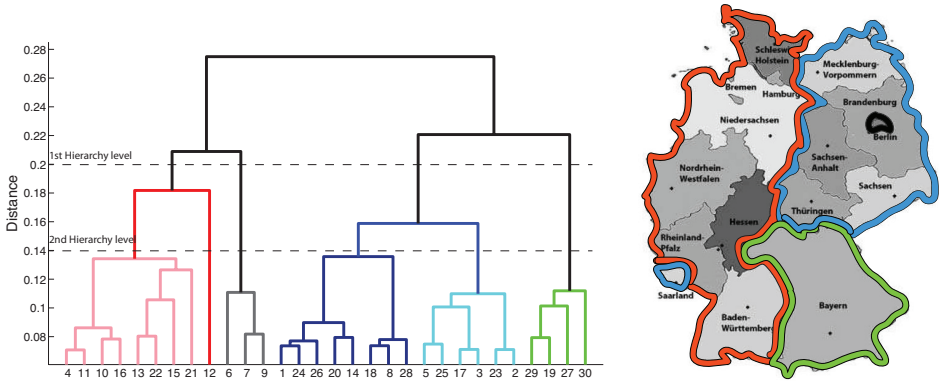
To test both methods, a data set of the German parliamentary elections of September 18, 2005 was used. The data, extracted from package *flexclust* of "R" program [3], consist of the proportions of "second votes" obtained by the five parties that got elected to the first chamber of the German parliament for each of the 299 electoral districts. It should be emphasized that the proportions do not sum the unity because parties that did not get elected into parliament have been omitted from the data set. Before Election day, the German government comprised a coalition of Social Democrats (SPD) and the Green Party (GRUENE); their main opposition consisted of the conservative party (Christian Democrats, UNION) and the Liberal Party (FDP). The latter two intended to form a coalition after the election if they gained a joint majority, so the two major "sides" during the campaign were SPD+GRUENE versus UNION+FDP. In addition, a new "party of the left wing" (LINKE) canvassed for the first time; this new party contained the descendents of the Communist Party of the former East Germany and some left-wing separatists from the SPD in the former West Germany.

4.2 Performance Evaluation

Fig. 4a shows the dendrogram obtained for the *"German Elections"* data set. The number of clusters and the hierarchy can be visually established by the expert. In particular, analyzing the dendrogram when the distance is 0.2 (higher dashed line), four different clusters (level 1) are obtained, which are represented in red, black, blue and green colors. The second level can be obtained at a the distance around 0.14 (lower dashed line). In this way, the red and blue clusters are now divided into two new ones represented in different shades of the same color that its parent.

Each electoral district belongs to one of the 16 German federal states. After carrying out the clustering, the state corresponding with each pattern of the different clusters was analyzed. Therefore, the most predominant states in each cluster can be found in order to check if each cluster corresponds with different German areas. The conclusion is that the four clusters (first level in the hierarchy) correspond with four different regions, namely, West Germany (without Saarland), East Germany (without Berlin and without Bayern) together with Saarland, Bayern and Berlin represented in Fig. 4b in red, blue, green and black, respectively.

Saarland's behaviour (located in the southwest of Germany, at the French border) may attract some attention because they voted in a similar way to the Eastern states. This is most likely due to the fact that Oskar Lafontaine, one of the two leaders of LINKE, is a former prime minister of Saarland as pointed out in [5]. Another striking state is Berlin, which exhibits very diverse voting behaviour and thus spreads over the rest of the clusters except some patterns, which form a different group because they are quite far from the other clusters.



(a) (b)

Fig. 4: (a) Dendrogram corresponding with the clustering of the *German Elections* data set. Higher dashed line represents the distance of the first hierarchical level. Lower dashed line represents the distance of the second hierarchical level. (b) Map of Germany with the 16 different German federal states. The four regions corresponding with the clustering in the first hierarchy level are delimited in different colors.

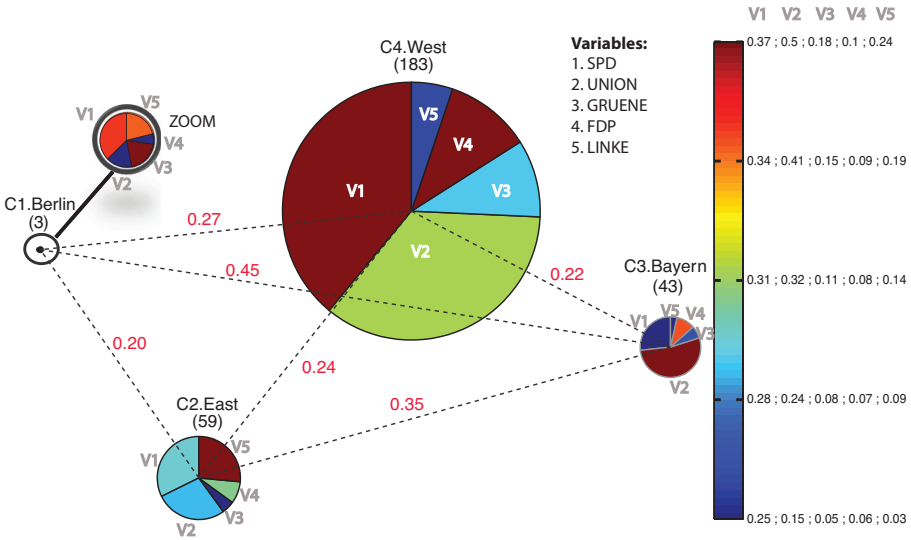


Fig. 5: MDSonS representation for the *German Elections* data set.

Fig. 5 shows the representation produced by MDSonS (only for the first hierarchy level). This visualization is appropriate for this kind of data because a long arc corresponding to each variable means a large number of votes. In this way, it is easily recognizable which party has the strongest performance but also the exact value of each party (information provided by the color bar and its labels).

In Fig. 5, it can be seen that there are four different clusters corresponding with the geographic areas marked with several colors in Fig. 4b. From now on, the red area will be called *"West"*, the blue one *"East"*, the green one *"Bayern"* and the black one *"Berlin"*.

The cluster with the largest number of patterns is *"West"*, specifically 183 (shown within parentheses), hence the corresponding circle is the biggest one. The most relevant features in this cluster are those corresponding with SPD and UNION (1st and 2nd features respectively). The sectors corresponding with these parties, which are in opposite wings, are quite similar in arc length; but if the 1st and 2nd column labels are checked it can be seen that the most voted party is SPD with a value of 0.37. The rest of parties in this cluster are not significantly relevant.

In cluter *"Bayern"* occurs the same as in the explained previously. The parties with the biggest support are SPD (0.25) and UNION (0.5), but in this case the most supported party is UNION instead SPD. The most similar cluster to *"Bayern"*, is *"West"*; it should be emphasized that these two clusters are the only ones that have the largest support in the two first parties, receiving the other three parties much lower support. This assumption can be proved checking all the distances (red numbers) between the cluster *"Bayern"* and the other clusters since the minimum distance appears between the two mentioned clusters.

The cluster *"Berlin"* has only three patterns, the rest of Berlin patterns (nine) spread over the rest of the clusters. Thus, these three patterns corresponding with cluster *"Berlin"* cannot be considered as a global behavioral pattern of Berlin. It should also be emphasized that the closest cluster to *"Berlin"* is *"East"*, inferred from the red numbers. In fact these are the two most similar (closest) clusters among all of them. Notice that these two clusters are the only ones where LINKE has an important relevance. This makes sense since LINKE party contained the descendents of the Communist Party of the former East Germany and some left-wing separatists from the SPD in the former West Germany.

In the case of cluster *"East"*, the parties with strongest performance are SPD (0.3), UNION (0.25) and LINKE (0.24). As commented previously, LINKE has a significant performance compared with other clusters.

In the next hierarchy level new conclusions can be drawn following the same procedure. In addition to the conclusions of the features within each cluster, information can also be extracted about the relationship of features in the different clusters. For example, SPD presents the biggest support in cluster *"West"* (0.37), UNION in cluster *"Bayern"*, GRUENE in cluster *"Berlin"* (0.18), FDP in cluster *"West"* (0.1) and LINKE in cluster *"East"* (0.24).

Fig. 6 shows the clustering visualization achieved by the SonS method (only first hierarchy), in which each sector corresponds to a cluster and each subsector to a variable, in order to compare the differences of both visualization tools. Some similar conclusions can be obtained; however, the information related to the distances between different clusters is not available. This information provides great utility to the method, and on certain scenarios may be very important. It also helps to contrast hypotheses. For example, it is known by intuition that the clusters *"Bayern"* and *"West"* are similar since these two clusters are the only ones that have the largest support in the first two parties, receiving the other three parties much lower support, as mentioned previously. However, representing the distances among clusters it can be proved. Moreover, it can also be proved the hypothesis that clusters *"Berlin"* and *"East"* are very similar because these two clusters are the only ones where LINKE has an important relevance. This information, which is not available in the SonS method, provides an essential aid to the problem understanding. In addition, the result of the method is more intuitive to analyze, and since it presents information in a less compact design, it allows a neat representation of a larger number of variables.

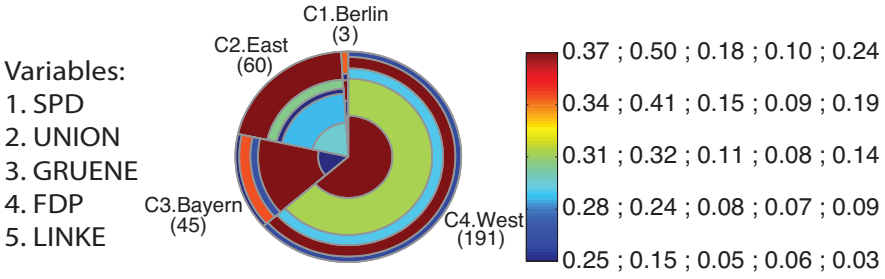


Fig. 6: *SonS* visualization method for the *German Elections* data set.

5 Conclusions

A new visualization method (MDSonS) for hierarchical clustering has been proposed in this paper. Its performance has been assessed by means of a real example, which demonstrates its applicability. MDSonS has shown to be a very useful tool when visualizing hierarchical clustering since it is possible to infer relationships among features, clusters and levels of the hierarchy. At the same time, this new approach entails a new improvement with regard to the SonS method [7], which consists of carrying out an MDS of the centroids; and drawing each cluster in the location provided by MDS. Therefore, MDSons makes possible to know the similarity between clusters at a glance. This new method is easier to interpret for an untrained reader due to the fact that SonS graph is somewhat

overladen of sectors and sub-sectors. Therefore, the MDSonS is able to represent more features than the SonS with an acceptable interpretability. Moreover, MDSonS makes possible to know the similarity between clusters at a glance by representing the distances. For that purpose, in the 2D space of representation, the circles corresponding to each cluster are shown separated by a distance proportional to the actual distance between the centroids in the high-dimensional space.

References

1. Baehrecke, E., Dang, N., Babaria, K., Shneiderman, B.: Visualization and analysis of microarray and gene ontology data with treemaps. BMC bioinformatics 5(1), 84 (2004)
2. Borg, I., Groenen, P.J.: Modern Multidimensional Scaling. Theory and Applications. Springer (2005)
3. Chambers, J.M.: Software for Data Analysis: Programming with R. Springer, New York (2008)
4. Chen, C.h., Hrdle, W., Unwin, A.: Handbook of Data Visualization (Springer Handbooks of Computational Statistics). Springer-Verlag TELOS, Santa Clara, CA, USA (2008)
5. Kesselman, M., Krieger, J., Joseph, W.: Introduction to Comparative Politics: Political Challenges and Changing Agendas. Wadsworth (2009)
6. Makanju, A., Brooks, S., Zincir-Heywood, A., Milios, E.: Logview: Visualizing event log clusters. In: Privacy, Security and Trust, 2008. PST '08. Sixth Annual Conference on. pp. 99 –108 (oct 2008)
7. Martínez-Martínez, J.M., Escandell-Montero, P., Soria-Olivas, E., Martín-Guerrero, J.D., Martínez-Sober, M., Gómez-Sanchis, J.: Sectors on sectors (sons): A new hierarchical clustering visualization tool. In: Computational Intelligence and Data Mining, 20011. CIDM '11. IEEE Symposium on. pp. 304–309 (15 2011-april 2011)
8. McConnell, P., Johnson, K., Lin, S.: Applications of Tree-Maps to hierarchical biological data. Bioinformatics 18(9), 1278 (2002)
9. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20(1), 53–65 (1987)
10. Shneiderman, B.: Tree visualization with tree-maps: A 2-d space-filling approach. ACM Transactions on Graphics 11, 92–99 (1991)
11. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, Fourth Edition. Academic Press (2008)
12. Thomas, J.J., Tajudin, D.A.: Visualizing the examination timetabling data using clustering method and treemaps. In: Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications (June 2006)