ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Growing Hierarchical Sectors on Sectors

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

University of Valencia - Electronic Engineering Department
st/ Dr Moliner 50, 46100, Burjassot, Valencia - Spain

Abstract.
Self-organizing maps are widely used in visual data mining. This paper proposes a new visualization approach for GHSOM algorithm, a hierarchical variant of SOM. The method is based on pie charts. That improves the visualization in hierarchical data structures making possible to extract all the existing relationships among the attributes of the neurons at any hierarchy level. The methodology is tested in one synthetic data set and one real data set. Achieved results show the suitability and usefulness of the proposed approach.

1 Introduction

The Self-Organizing Map (SOM) [1] is one of the most popular visualization tool. The SOM is a neural model that carries out a low-dimensional visualization of patterns defined in N-dimensional data sets. Two of the main limitations of SOM are the static architecture of the model and the difficulty for obtaining hierarchical realtionships [2]. Since hierarchical models can extract more information from a data set, SOM has been modified in several ways to deal with hierarchical frameworks, being specially remarkable the Growing Hierarchical Self-Organizing Map (GHSOM) [2, 3, 4]. The main problem of GHSOM is that it is not possible to visualize simultaneously the data information in each level. This paper presents a new visualization technique for GHSOM, that allows a simultaneous and compact visualization of the different hierarchy levels.

2 Growing Hierarchical Sectors on Sectors (GHSonS)

GHSonS is a new visualization method for GHSOM algorithm. Each neuron is represented by a slice of a circle (pie sectors). The area of each pie sector is proportional to the number of patterns assigned to each neuron (number of winners). 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 attributes at any hierarchy level can be extracted. The values of the different features should be scaled between $[0, 1]$, before carrying out the training and the visualization, in order to avoid a biased model. The graph is produced in three steps for each hierarchy level, starting from the first hierarchy level in which only one pie chart is used:

[*] This work has been partially supported by the *Spanish Ministry of Education and Science* with the Projects TIN2007-61006, CSD2007-00018 and DPI2010-19457, with the support of European FEDER funds in the last one.

1. *Division of one circle depending on the number of neurons in the present hierarchy*: The circle is divided into several sectors corresponding to each neuron. The area of each sector is proportional to the number of patterns assigned in each neuron. The number of patterns belonging to each neuron is shown within parentheses, thus showing the relevance of each neuron.

2. *Division of the pie sectors depending on the number and the value of variables*: Each sector, corresponding to each neuron, is divided into several parts which correspond to each variable. The inner part corresponds to the first variable and going outwards the next variables appear. Each one of these parts vary its radius. This radius corresponds to the relative value of each variable, with respect to the sum of all of them, in the corresponding pie segment.

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. The value of the first feature (inner area in one sector), is given by the first column label, the second feature by the second column label and so on. This way, the exact value of each variable for each neuron can be known.

The description for the first hierarchy level can be extended to the rest of levels. For the second hierarchy level, for instance, Fig. 2 shows that for each sector (in level 1) emerges a new pie chart with new values in the variables of the different neurons. If in a given data set more than two levels are present, a new pie will emerge from the sectors of the previous level. It should be emphasized that not always a new pie emerges from all the sectors, but it depends on the hierarchy given by GHSOM.

The main advantage of the proposed visualization technique is that it is possible to observe relationships among different variables in the same particular neurons and relationships among the same variables in different neurons, in the different levels of the hierarchy.

3 Results

3.1 Data sets

The first data set used is synthetic. It consists of three clouds of points defined by X, Y, and Z coordinates, as shown in Fig. 1. These three clouds of points can be divided into nine. The data set was labeled depending on the hierarchy of the data, i.e., three "super-classes" corresponding to the main three clusters and nine "sub-classes" corresponding to the 9 divided clusters.

As a real example it was used a data set that contains information about the percentage composition of fatty acids found in the lipid fraction of Italian olive oils [5]. The data set consists of 572 samples and 10 variables. The training variables are eight fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic,linolenic, arachidic, eicosenoic) measured in % × 100 (i.e.,‱). The other two variables contain information about the classes. There are two kinds of classes: three "super-classes" (regions of Italy): North, South, and the island of Sardinia; and

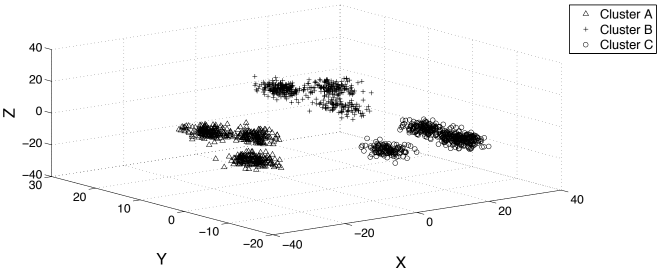

Fig. 1: Representation of synthetic data set. It is shown with different markers the points corresponding to the three main clusters.

nine collection areas: three from the Northern region (Umbria, East and West Liguria), four from the South (North and South Apulia, Calabria, and Sicily), and two from the island of Sardinia (inland and coastal Sardinia). The goal is to distinguish the oils from different regions and areas in Italy based on their combinations of the fatty acids [5].

3.2 Performance evaluation

3.2.1 Example 1: Synthetic data set

Fig. 1 shows three clusters after the first hierarchy level. After training the data set with the GHSOM algorithm, GHSonS is employed to produce the graph shown in Fig. 2[1].The training started with four neurons (four sectors in center pie) and the hierarchy grew up to two levels. In first hierarchy level, Fig. 2 shows one sector corresponding to the cluster “C”, one corresponding to “A” and other two corresponding to “B”. For distinguishing among “B” and the other two clusters the third variable (Z coordinate) is specially relevant. In Fig. 2 it is observed that this variable (outer subsector) takes high values for “B” and low values for “A” and “C”. This can be confirmed observing Z coordinate in Fig. 1. Once distinguished the cluster “B” from the two others, it should be distinguished between clusters “A” and “C”. For “A” and “C”, Fig. 2 show differences between variables 1 and 2 (X and Y coordinates) whereas the third variable remains almost constant. The most relevant difference is found in the first variable that takes the minimum value for the cluster “A” (about -15, see first column label) and the maximum value for the cluster “C” (about 22, see the first column label). As it can be observed in Fig. 1 the mentioned clusters are clearly distinguishable by means of the X coordinate. In order to distinguish among subclusters, a similar procedure can be carried out for the next hierarchy level (2).

[1] Figures corresponding to GHSonS graphs are available in color at http://idal.uv.es/ghsons

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

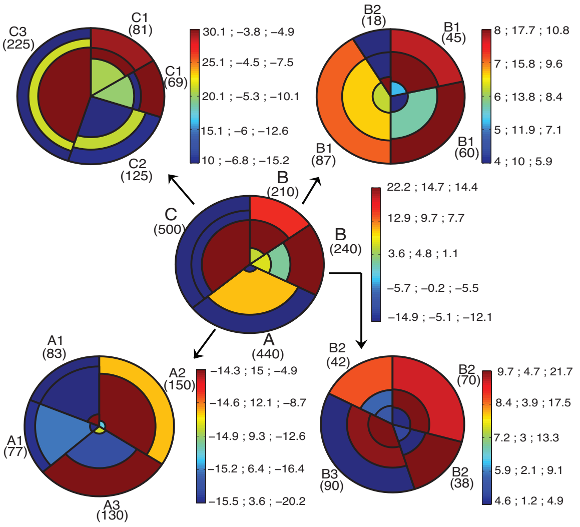

Fig. 2: Proposed visualization method for *synthetic* data set.

3.2.2 *Example2: Italian Olive Oils*

After training the mentioned data set with the GHSOM algorithm, two hierarchy levels are produced; the first level started the training with four neurons (four sectors in Fig.3). After this, the predominant region for each neuron, in the first hierarchy level, is checked. For the first hierarchy level (top pie chart, Fig. 3) there is one sector corresponding to the island of Sardinia, another to the South (specifically South Apulia), another to the North and finally another corresponding to the South again (specifically North Apulia, Calabria and Sicily). As it can be observed the radius of last variable, for some sectors, is very small. This fact makes dificult the visualization of the value in the mentioned variable. Because of this a zoom of the image has been carried out. The labels of the color bars have been removed for a better visualization; only the color bars are shown in order to indicate a qualitative value. Notice that although a given color may be very similar for two different variables, it does not mean a very similar value for the variables since each variable has its own range of values.

For distinguishing the oils from the different regions, in the first hierarchy level, the most important variables are the 8th (outer subsector in the circle) and the 6th (6th subsector starting from the inner one). If 8th variable is high, it involves that the oil belongs to the South and if it is low belongs to either the North or the Island of Sardinia (the same occurs with the 6th variable). For distinguishing between North and Sardinia the 5th and the 7th variables play a relevant role (high values for Sardinia and low for the North).

In the next hierarchy level three new GHSonS graphs were found which emerged from the previous sectors corresponding to Island of Sardinia, North and finally the South, specifically the sector which represented North Apulia,

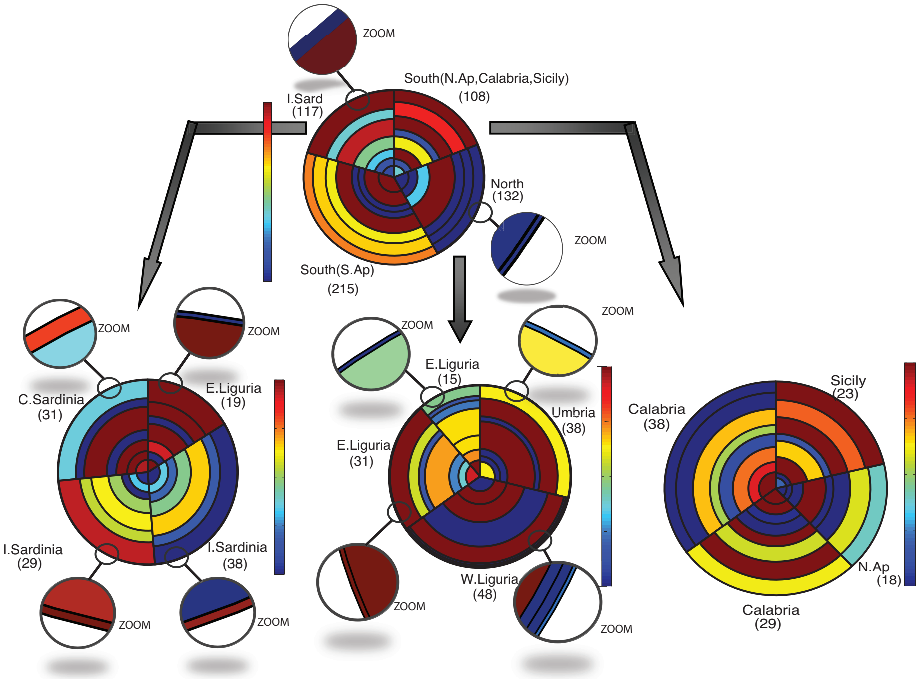

Fig. 3: Proposed visualization method for *Italian Olive Oil* data set.

Calabria and Sicily (Fig. 3).

The pie which emerged from the sector "I. Sard" has four sectors, one corresponding to Coast Sardinia, other two to the Island of Sardinia and finally another corresponding to East Liguria. Althoug it might be expected to find only neurons (sectors) corresponding to Island of Sardinia there is also one belonging to the North (E.Liguria); this is because the sectors were labeled with the name of the region which had the biggest number of patterns in this neuron; moreover the regions Island of Sardinia and the North are similar (they were only distinguished by means of two variables, the 5th and the 7th, in the previous hierarchy). However in the left pie (second level of hierarchy) East Liguria is easily distinguishable from the other sectors by means of the 8th variable (low values for East Liguria whereas for the rest of sectors it presents high values). In order to distinguish between the two "sub-classes" corresponding to the region Island of Sardinia (inland and coastal Sardinia) the variables that must be taken into account are the 1st, 2nd and 3rd (three inner subsectors); coastal Sardinia takes high values and inland Sardinia takes low values.

Regarding the "sub-classes" of the North (central pie of second hierarchy level, Fig 3), there are again four sectors, two corresponding with East Liguria, one corresponding with West Liguria and other one with Umbria. One of the two clusters corresponding with East Liguria is basically formed by oils from this area but the other also contains a considerable number of patterns belonging to Umbria. Low values of the variables 6th, 7th and 8th distinguish West Liguria from the rest of areas, but these variables present low relevance (very small radius). Actually, the 5th variable must be taken into account since it is more relevant and presents the maximum value for West Liguria and low values for the rest of Northern areas. Now for distinguishing between the rest of oils from these areas (East Liguria and Umbria), the 6th variable must be used (low values for East Liguria and high values for Umbria).The right pie in the second level of hierarchy (Fig. 3) describes the areas from the South. The 1st and the 2nd features distinguish Calabria from the other Southern areas; the oils from Calabria present high values in these variables, whereas for the other two Southern areas (Sicily and North Apulia) they present minimum values. In order to distinguish between the oils from Sicily and North Apulia it must be taken into account the variables 3rd and 6th. They present maximum values for Sicily and minimum for North Apulia. As it can bee seen in this pie, the sector corresponding with Sicily only presents 23 patterns (some of them actually belong to North Apulia) whereas it actually has 36. This is due to the fact that the rest of patterns spread over the rest of clusters; in particular, in one of sectors of Calabria (in the second hierarchy level, right pie) and one of the sectors (in the first hierarchy level) corresponding with the South (specifically to South Apulia).

4 Conclusion

A new visualization method for GHSOM algorithm has been proprosed. It has been shown the performance of this visualization tool by means of two examples (one synthetic and one real) demonstrating its applicability. The proposed method has shown to be a useful tool when visualizing hierarchical data since it is possible to infer relationships among features, neurons and levels of the hierarchy demonstrating its capacity for extract information from the data.

References

[1] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer Berlin, 3 edition, 2001.

[2] M. Dittenbach, D. Merkl, and A. Rauber. The growing hierarchical self-organizing map. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, volume 6, pages 15 –19 vol.6, July 2000.

[3] Risto Miikkulainen. Script recognition with hierarchical feature maps. *Connection Science*, 2:83–101, 1990.

[4] M. Dittenbach, A. Rauber, and D. Merkl. Uncovering hierarchical structure i data using the growing hierarchical self-organizaing map. *Neurocomputing*, 48(1):199–216(18), 2002.

[5] Armanino C. Lanteri S Forina, M. and E Tiscornia. *Classification of Olive Oils from their Fatty Acid Composition*, pages 189–214. Food Research and Data Analysis. Applied Science Publishers, London, 1983.