ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Extended visualization method for classification trees

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

IDAL, Intelligent Data Analysis Laboratory
University of Valencia - Electronic Engineering Department
Av de la Universidad, s/n, 46100, Burjassot, Valencia - Spain

Abstract. Classification tree analysis is one of the main techniques used in Data Mining, and nowadays there is a lack of a visualization method that support this tool. Therefore, graphical procedures can be developed in order to help simplify interpretation and to obtain a better understanding. This paper proposes a method for representing the input data for each class presented in each terminal node. For this purpose, the new visualization method *Sectors on Sectors (SonS)* is used. The methodology is tested in two real data sets.

1 Introduction

Data visualization can greatly enhance our understanding of multivariate data structures, and hence Data Mining techniques and data visualization often go hand in hand [1]. Because of this, it is not surprising to think about using data visualization in classification trees in order to simplify its interpretation.

Classification and Regression Trees (CARTs) are analytic procedures for predicting the values of a response variable from input variables [2, 3]. When the response variable of interest is categorical in nature, the technique is referred to as Classification Trees; if it is continuous in nature, the method is referred to as Regression Trees [4].

For classification problems, in which we focus on this paper, the goal is to find a tree where the terminal tree nodes are relatively "pure" i.e., contain observations that (almost) all belong to the same category or class. However, not always happen that the terminal nodes are pure. Because of this, we propose a visualization tool in which we can extract the maximum information by means of representing the input data organization for each class presented in each terminal node as well as the number of patterns belonging to each class presented in each terminal node. The proposed graphical procedure helps to simplify interpretation even for complex trees and helps to the understanding of what is happening in the terminal nodes.

*This work has been partially supported by the Projects CSD2007-00018 and UV-INV-AE11-41271.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

2 Sectors on Sectors (SonS)

The *Sectors on Sectors (SonS)* visualization method, proposed in [5], is based on the well-known pie chart visualization, and it focuses on visualizing hierarchical clustering. In the case presented in this paper, the *SonS* method is used in order to visualize the input space in the terminal nodes of the classification tree. Fig. 1 represents the three steps followed to create the *SonS* visualization method.

1. *Division of one circle on several sectors depending on the number of different classes in a terminal node:* First of all the circle is divided into several pie segments or sectors corresponding to each class. The area of each sector is proportional to the number of patterns included in each class. The number of patterns belonging to each class is shown within parentheses. In this way, the significance of each class is easily recognizable (Fig. 1 on the left).

2. *Division of the pie sectors depending on the number of variables:* After the first step, each sector is divided into as many subsectors as variables we find in the problem. The inner part corresponds to the first variable, and as we go outwards we encounter the next variables. In the original method, presented in [5], each one of these parts vary its radius in order to represent the relevance of each variable, but for the sake of simplicity, this step has been omitted (Fig. 1 in center).

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 variables of each class centroid (mean value) is codified by means of colors. 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 class centroid (Fig. 1 on the right).

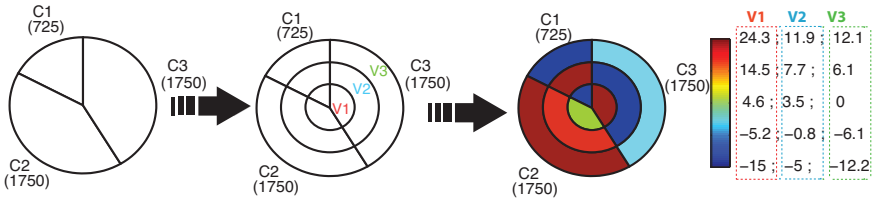

Fig. 1: The three steps followed to create the *SonS* visualization method.

3 Results

3.1 Data sets

Two real examples were used to assess the performance of the proposed method. The first data set used is the *"Iris flower data set"* [6]. The data set contains 3

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

classes of 50 instances each, where each class refers to a type of iris plant (*Setosa, Versicolor, Virginica*). One class is linearly separable from the other two; the latter are not linearly separable from each other. The input variables correspond with *sepal length, sepal width, petal length and petal width* (all in centimetres).

The second data set used is about Italian olive oils. This data set contains information about the percentage composition of fatty acids found in the lipid fraction of Italian olive oils [7]. The data set consists of 572 samples and 9 variables. There are eight variables that are fatty acids measured in $\% \times 100$ (i.e.,‰), and one variable that contains information about the classes. The classes refer to nine collection areas: three from the Northern region of Italy (Umbria, East and West Liguria), four from the South of Italy (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 areas in Italy based on their combinations of the fatty acids [7].

3.2 Performance evaluation

3.2.1 Example 1: Iris flower data set

For the classification tree training, 10-fold cross-validation has been used with the whole data set to compute the cost vector. The tree is pruned in that level that produces the smallest tree that is within one standard error of the minimum-cost tree [2, 4]. Fig. 2 [1] shows the classification tree obtained for the *"Iris flower data set"*. In each terminal node, the *SonS* graph has been drawn unless all the patterns included in the terminal node belong to the same class (as occurs in terminal node labelled as "*Setosa*"). As shown in Fig. 2, the 3rd variable separates *Setosa* class from others (*petal length*). If it takes a value less than 2.45, it means that the input pattern will belong to this class, and it will belong to any of the other two classes otherwise. The classification tree indicates that in order to differentiate between *Versicolor* and *Virginica* classes, the last variable (*Petal Width*) must be taken into account. If this variable is less than 1.75, the input pattern will belong to the *Versicolor* class; and if it is greater than or equal to 1.75 the pattern will belong to the *Virginica* class. However, as extracted from the *SonS* graph, in the terminal node corresponding with *Versicolor*, there are 5 patterns belonging to *Virginica* class. Looking at the last variable (outer subsector), which distinguish between the *Versicolor* and *Virginica* classes, along with the last column of the color bar, it can be seen that the sector corresponding to *Virginica* takes a value of 1.5, whereas the *Versicolor* class takes a value of 1.3. Thus, we could say that, *Versicolor* class corresponds with a value less than 1.5 instead of 1.75, as classification tree indicates. In the terminal node corresponding to *Virginica*, it can be observed that just one pattern belonging to *Versicolor* class has been erroneously included.

[1] Figures corresponding to classification trees and *SonS* graphs are available in color at http://idal.uv.es/SonStree

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

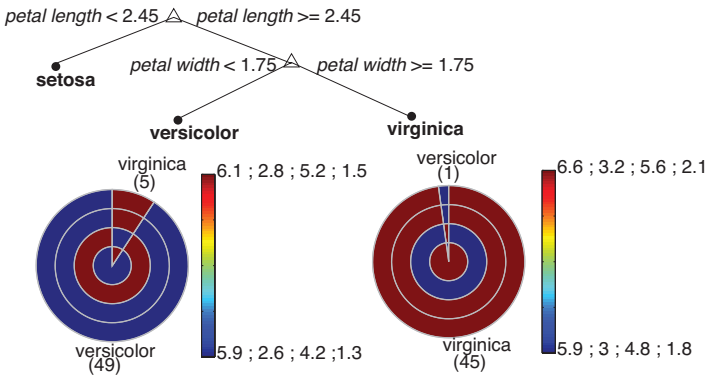

Fig. 2: Classification tree obtained for the *"Iris flower data set"* with the *SonS* graph in the terminal nodes.

3.2.2 Example 2: Italian olive oils data set

The classification tree training was carried out in the same way as in section 3.2.1. Fig. 3 shows the tree obtained for the *"Italian olive oils data set"*.

To extract the most significant conclusions, special attention will be paid to those terminal nodes where there is a considerable number of patterns erroneously included (more than 20%). Therefore, Fig. 3, only shows the *SonS* graphs that follow this rule. The first *SonS* graph that attracts some attention is the corresponding to the first Calabria terminal node (1st chart starting from the right) because more than the 30% of the patterns are wrong. This chart has one sector corresponding to Calabria (9 patterns), another one corresponding to Sicily (3 patterns) and finally one corresponding to South Apulia (1 pattern). In order to distinguish among these groups of patterns, new decision rules must be established. For example, Calabria and Sicily are easily distinguishable by means of the 4th variable because Calabria presents a maximum value (7352), indicated by deep red color, and Sicily presents a minimum value (7103), indicated by blue. Notice that other variables also present maximum values in one of these regions, and minimum values for the other one, but the 4th variable presents the widest range (in relative values) between the maximum and minimum values. Therefore, the procedure to follow is to choose an intermediate value (7227.5) to separate between these two regions. Hence, the new rule is that if the 4th variable is less than 7227.5, the patterns will belong to Sicily; and if they are greater than or equal to 7227.5 will belong to Calabria. For distinguishing South Apulia from others regions, a similar procedure can be followed, but since only one pattern is affected, an ad-hoc definition of a rule might be pointless.

Another terminal node that presents a large number of patterns erroneously included is the corresponding to the second Calabria terminal node (2nd chart starting from the right). In this case, low values of the 8th variable separate Calabria from Sicily.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

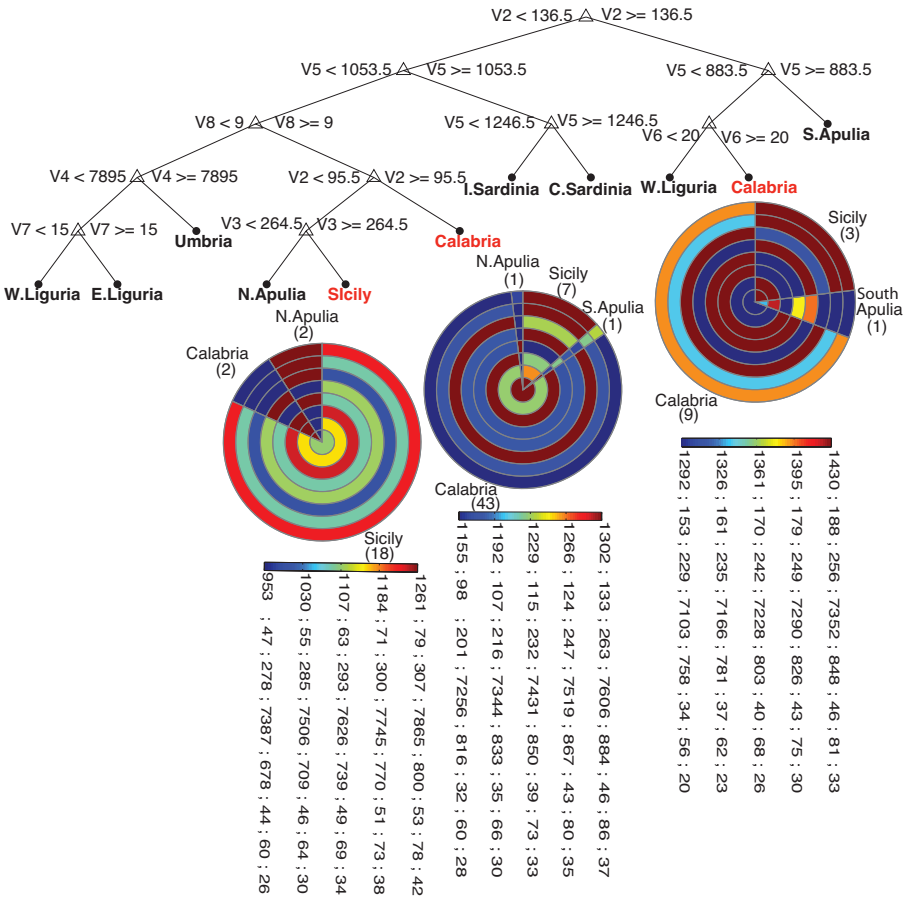

Fig. 3: Classification tree obtained for the *"Italian olive oils"* data set with the *SonS* graph in the terminal nodes.

The last terminal node to consider is that corresponding to Sicily (3rd chart starting from the right). In this case, the 5th variable takes relevance in order to distinguish among the regions in this terminal node. Notice that, if this variable takes low values (blue) the olive oil will belong to North Apulia, if it takes intermediate values (green) the olive oil will belong to Sicily, and finally, if it takes high values (red) the olive oil will belong to Calabria. This is the only variable that distinguishes among the three classes included in this terminal node. It is worth mentioning that a deeper tree (which would have less generalization ability) could achieve to separate these classes. Our approach allows to extract, visually, this separation as well as gain knowledge about the problem while preserving the generalization capabilities of the tree.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

4 Conclusion

In this paper, the performance of the *SonS* method applied to classification trees has been shown by means of two real examples, demonstrating its applicability. The proposed graphical procedure helps to extract knowledge and interpretation and to obtain a better understanding even for complex trees. This method is capable of providing visual information of the input patterns belonging to a terminal node in the decision tree, so that it will be possible to extract information about the values of their variables and obtain information about the patterns erroneously included. Therefore, new decision rules can be established visually in order to distinguish them. One limitation is that this approach is not very efficient when dealing with a large number of features.

Another advantage of this method is that it could also be used to build shallow classification trees. That means that, to draw a very deep tree is not necessary because we can extract the same conclusions visually (starting in previous nodes). That is, if the nodes of the tree are removed at some level, it will be possible to establish the rules visually without needing to build deep trees. Moreover, the *SonS* graphs could be used in other nodes (not only in terminal nodes) in order to obtain visual information about how the classification tree evolves.

Another interesting use of the original *SonS* method, in classification trees, could be to carry out a clustering algorithm with the data included in each terminal node and visualize the result. In this case, visual information about the different clusters obtained in each terminal node would be extracted.

References

[1] Chun-houh Chen, Wolfgang Hrdle, and Antony Unwin. *Handbook of Data Visualization (Springer Handbooks of Computational Statistics)*. Springer-Verlag TELOS, Santa Clara, CA, USA, 2008.

[2] Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, 1 edition, January 1984.

[3] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, corrected edition, July 2003.

[4] Ethem Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2nd edition, 2010.

[5] José M. Martínez-Martínez, Pablo Escandell-Montero, Emilio Soria-Olivas, José D. Martín-Guerrero, Marcelino Martínez-Sober, and Juan Gómez-Sanchis. Sectors on Sectors (SonS): A New Hierarchical Clustering Visualization Tool. In *Computational Intelligence and Data Mining, 20011. CIDM '11. IEEE Symposium on*, pages 304–309, April 2011.

[6] A. Frank and A. Asuncion. UCI machine learning repository [http://archive.ics.uci.edu/ml], 2011.

[7] 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.