ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

ManiSonS: A New Visualization Tool for Manifold Clustering

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

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

Abstract. Manifold learning is an important theme in machine learning. This paper proposes a new visualization approach to manifold clustering. The method is based on pie charts in order to obtain meaningful visualizations of the clustering results when applying a manifold technique. In addition to this, the proposed approach extracts all the existing relationships among the attributes of the different clusters and find the most important variables of the manifold in order to distinguish among the different clusters. 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

Many manifold learning methods have been developed in the last decade, and it has become a hot topic. These dimension reduction methods can be approached from the point of view of either unsupervised learning or supervised learning. They can be divided into linear and non linear methods. Recent research has focused on nonlinear manifolds, and the long list of "manifold learning" algorithms provide sophisticated examples of dimension reduction [1, 2]. In the context of machine learning, manifold methods may be viewed as a preliminary feature extraction step, after which pattern recognition algorithms are applied. In this paper, clustering algorithms are used after applying the manifold technique. Moreover, as it is well-known, data visualization can greatly enhance the understanding of multivariate data structures, and hence cluster analysis and data visualization often go hand in hand. So that, to visualize the clustering results after applying the manifold can be of great interest. Therefore, this paper presents a new use, in manifold field, of the *Sectors on Sectors (SonS)* visualization technique, recently proposed by the authors in [3], in order to show the results of the clustering carried out on the manifold. Supervised approaches were used because this paper deals with two classification problems where the information about the label of each pattern is available. In order to solve these problems, both linear and nonlinear methods have been used. The

*This work was partially financed by the Spanish Ministerio de Economía y Competitividad (MINECO) and FEDER funds in the framework of the project *EFIS: Un Sistema Inteligente Adaptativo para la Gestión Eficiente de la Energía en Grandes Edificios*, with reference IPT-2011-0962-920000.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

methods used to solve these problems were Linear Discriminant Analysis (LDA) [4], Neighborhood Components Analysis (NCA) [5] and Maximally Collapsing Metric Learning (MCML) [6].

2 ManiSonS: Sectors on Sectors (SonS) applied to Manifold visualization.

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 [3]. This method is based on the well-known pie chart visualization. Each cluster is represented by one 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.

Due to the importance of obtaining knowledge about the Manifold, the use of new visualization methods is paramount. *ManiSonS* makes possible to obtain visual information about the clustering in the reduced space. Figure 1[1] 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 found in the manifold:** 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 1, 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. This radius corresponds to the relative value of each variable, with respect to the sum of all of them[2]. That is, let X be a centroid corresponding to one cluster, so that,

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

[1] Figures included in this paper are available in color at http://idal.uv.es/ManiSonS

[2] 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 on its range, that is, the relevance is not higher even though the variable has a higher range. The use of scaled variables, guarantees that the radius is a measure of the relevance of the variable within the cluster.

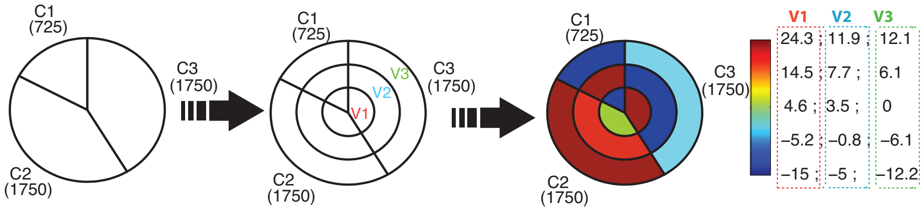

Fig. 1: The three steps followed to create the *SonS* visualization method. From left to right: producing as many sectors as clusters; splitting each sector according to the attributes; and color coding to identify real values.

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 1, 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[3]. 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 1, right).

3 Results

3.1 Data sets

The first data set is a synthetic data set created to show the performance of the proposed visualization method. The data consists of three clouds of points defined by six coordinates. The first three coordinates contain the most relevant information about the three different clusters, while the remaining three provide irrelevant information or noise, that is, very small values that barely provide information. Table 1 shows the variation ranges of each one of the variables.

[3] This is automatically extensible to other measures such as the median, which is a much more adequate prototype measure in presence of outliers, for instance. Therefore, SonS is not restricted to the use of a particular prototype measure.

Coordinate	max.	min.	mean	σ
1st	-14.0083	-34.9812	-18.1518	6.8112
2nd	15.9955	0.0119	12.2174	5.6212
3rd	30.9962	10.0240	17.4124	6.2271
4th	0.0200	0.0100	0.0151	0.0029
5th	0.0200	0.0100	0.0150	0.0028
6th	0.0200	0.0101	0.0153	0.0028

Table 1: Information about the variable ranges of the synthetic data set.

Moreover, as a real example, the *seeds data set*[4] was used, which contains X-ray images of wheat. The examined group comprised kernels belonging to three different varieties of wheat: “Kama”, “Rosa” and “Canadian”, 70 elements each, randomly selected for the experiment. Studies were conducted using combine harvested wheat grain coming from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. The data set will be used for clustering tasks. To construct the data, seven geometric parameters of wheat kernels were measured (Area, Perimeter, Compactness, Length of kernel, Width of kernel, Asymmetry coefficient, Length of kernel groove). All of these parameters were real-valued continuous.

3.2 Performance evaluation

3.2.1 Synthetic Data Set

As previously mentioned, three different manifolds have been used to tackle this problem (LDA, NCA and MCML) because they are supervised techniques, that is, they make possible to introduce labels in the learning procedure. Moreover, they support exact out-of-sample extension, that is, they learn an explicit function between the data space and the low-dimensional latent space, with the same number of patterns in both spaces. After applying the different dimensionality reduction techniques (the data was reduced to two dimensions), a clustering based on k-means algorithm was performed.

Since the three manifolds produced the same success rate (100%), after applying the clustering algorithm, only the results obtained by NCA are shown. Figure 2 shows the results provided by the *ManiSonS* visualization method after applying the clustering algorithm on the manifold.

Figure 2 provides relevant information about the clustering carried out in the reduced space. For example, it can be seen which variables are important to separate between clusters. To this end, those variables that take high values for one cluster and low values for the other one, must be sought. For example, V1 is the most important variable to separate clusters C1 and C2, since it takes high values for C2 and low values for C1. In order to separate between C1 and C3, the two variables of the manifold are useful. The input pattern will belong to C1

[4] http://archive.ics.uci.edu/ml/datasets/seeds. (*Last checked November 2012*)

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

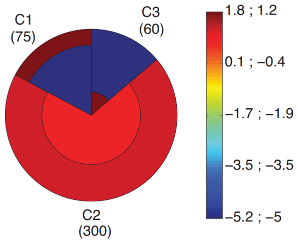

Fig. 2: *ManiSonS* visualization method applied to the *synthetic data set*.

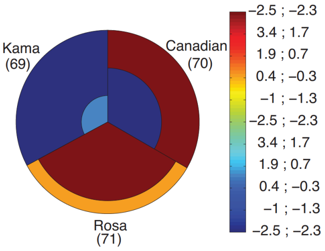

Fig. 3: *ManiSonS* visualization method applied to the *Seeds data set*.

if V1 takes low values or also if V2 takes high values. Otherwise, it will belong to C3. To separate between C2 and C3, the most relevant variable becomes V2.

3.2.2 Seeds Data Set

In this section, the three different manifolds that were used in the previous section are used again to tackle this problem. The dimensionality reduction technique finally presented is LDA (again the data was reduced to two dimensions) since it provided the highest success classification rate (96.67%). Figure 3 shows the results provided by the *ManiSonS* visualization method after applying the clustering algorithm on LDA manifold.

As shown in Figure 3, it is possible to characterize each cluster by means of the variables of the manifold. For example,"Kama" cluster is characterized because it is the only cluster in which the 2nd variable takes minimum values (dark blue). "Rosa" cluster is characterized because it is the only cluster in which 1st variable takes maximum values (dark red). The same occurs in "Canadian" cluster, but with the 2nd variable. Besides characterizing each cluster using the values that one variable in particular can take (or using the values of several variables in other possible problems) it can be determined which variables, or planes in the dimensional space of the manifold, are relevant to separate between clusters. For example, in order to distinguish between patterns belonging to "Kama" from "Rosa" variety, the 1st variable is very relevant. If the 1st variable takes low values, the wheat will belong to "Kama" variety, and if it takes high values it will belong to "Rosa" variety. For distinguishing between "Kama" and

"Canadian" variety, the 2nd variable takes the highest relevance. If this variable takes low values, the wheat will belong to "Kama" variety, while if it takes high values it will belong to "Canadian" one. Finally, for distinguish between "Rosa" and "Canadian" variety it should be checked the 1st variable. If it takes low values, the wheat will belong to "Canadian" variety, while if it takes high values the wheat will belong to "Rosa" one.

4 Conclusion

In this paper, a method called *ManiSonS*, which is based on *SonS* method applied to manifold clustering, has been presented by means of two examples (one synthetic and one real), demonstrating its applicability in order to extract rules from the visualization of the manifold clustering. The proposed method has shown to be a very useful tool when visualizing the clustering carried out on a manifold since it is possible to infer relationships among features and clusters. Moreover, it makes possible to determine which variable, or planes in the dimensional space of the manifold, are relevant to separate between clusters. The proposed graphical procedure helps to extract knowledge and interpretation and to obtain a better understanding about the results of the manifold.

This method can be used even with data sets with a large number of variables due to the fact that dimensionality reduction will make possible to represent the results of the clustering in the low-dimensional space without overloading the graph.

References

[1] J.A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction.* Information Science and Statistics. Springer, 2010.

[2] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000.

[3] 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.

[4] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 2004.

[5] Jacob Goldberger, Sam Roweis, Geoff Hinton, and Ruslan Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press, 2004.

[6] Amir Globerson and Sam Roweis. Metric learning by collapsing classes. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 451–458. MIT Press, Cambridge, MA, 2006.