https://doi.org/10.52326/ic-ecco.2022/SEC.05

Analysis with Unsupervised Learning Based Techniques of Load Factor Profiles and Hyperspectral Images

Ştefan-Gheorghe Pentiuc [1], ORCID: 0000-0002-5239-9493
Elena Crenguța Bobric [2], ORCID: 0000-0002-6570-3095
Laura-Bianca Bilius [3], ORCID: 0000-0002-6081-3674
[1] MintViz Lab, MANSiD Research Center, Ștefan cel Mare University of Suceava, 13 Universitatii, Suceava 720229, Romania, pentiuc@usm.ro, crengutab@eed.usv.ro, laura.bilius@usm.ro

***Abstract*— The problem of obtaining an optimal partition consistent with a series of partitions resulting from the application of various clustering algorithms is NP complete. A heuristic method based on the concepts of central partition and strong patterns developed by Edwin Diday [3] is proposed. It is presented the experience regarding the use of analysis techniques based on unsupervised learning methods of load factor profiles and hyperspectral images.**

Keywords—machine learning; unsupervised learning; clustering algorithms; load factor profiles; hyperspectral images

I. INTRODUCTION

There are situations in which the data must be divided into disjoint groups that contain elements more similar to each other than to those in other groups. In general, it is required to obtain a partition into equivalence classes of the set of observations. In Machine Learning (ML), the partitioning of a finite set E without having a priori information on how to group its elements into classes corresponds to the context of unsupervised learning. There are several algorithms that achieve this, also called clustering algorithms. These algorithms receive as input the set E and the number M of classes in which the elements of E are to be grouped, and possibly a metric over E or a measure of similarity. Based on these the algorithm will produce at the output a partition P(E,M) into M equivalence classes . But not infrequently the application of several different algorithms or even the same algorithm, but with various input parameters to configure its execution, produce different partitions of E in M classes. The question arises as to which is the best partition among those obtained.

A solution is to obtain a consensus partition that derives from a set of clusters, so that it is better than the other partitions [1] or that fits better according to a certain criterion than any partition from the set of partitions in entry [2].

In the first part of the paper, a heuristic method is presented to find the best partition from the set of partitions provided by several clustering algorithms, starting from the partition of strong patterns, a notion introduced by Edwin Diday [3]. In the second part of the work, the practical validation of the method will be presented in 2 case studies of high current interest (electricity management and remote detection).

II. UNSUPERVISED LEARNING ALGORITHMS

In this paper we use the notion of pattern to represent the entities described by the values of their properties. A similar notion in ML is sample. A pattern is a vector:

$$\mathbf{x} = (x_1, x_2, \ldots, x_p)$$

where x_i are the observed or measured values of the entity's attributes, being called features. If x_i are real, then a pattern is a point in the set R_p.

Let there be a finite set of patterns E and we assume that it is required to find a partition P(E,M) of the set E in M classes. If M is not known, then a series of partitions of E, P(1), P(2), ... , P(n) can be requested, and by analyzing them, it will be decided which is the most suitable partition . A very useful representation of such a series of partitions is the pattern dendrogram which represents an indexed hierarchy of patterns (see Figure 1).

In a dendrogram, by practicing some horizontal cuts for various values of the α index of the hierarchy, successive partitions will be obtained, a class is made up of all the patterns located in the leaf nodes of the respective subtree. For example, in Figure 1, if α=4.5 is considered, then P(1) = { x_1, x_2, ..., x_5} is obtained, if α = 3.5, P(2) = { {x_5} , { x_1, x_3, x_3, x_4} } etc. To choose the

most suitable partition from a series of partitions provided by a hierarchical clustering algorithm, the parameters that can guide the user in selecting the most suitable partition can be calculated.

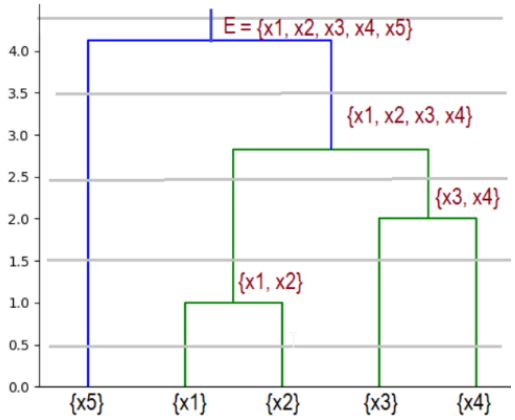

Figure 1. Example of a dendrogram.

However, our method does not refer to this kind of selection.

III. Best Partition

The problem we are addressing is how we choose from several partitions $P_i(E,M)$, with i=1,k, produced by several executions of some clustering algorithms with patterns from the set E. It can be the same algorithm, but executed with different parameters. For example, if we have a hierarchical clustering algorithm and execute it 4 times each time with a different linkage type (Ward, complete, average, or single linkage), it is possible to obtain 4 different partitions. What is the best partition, the one we should choose?

The problem consists in finding an optimal partition of a finite set E of patterns. This problem is NP complete and several heuristics have been proposed to solve it. In general, this problem is known as finding the consensus partition [1, 2] of several partitions.

In the approach we present, we will use the concepts of strong patterns and central partition, concepts defined by Edwin Diday and collaborators [3].

We assume that k clustering algorithms were executed that partitioned the set of patterns into M equivalence classes. All algorithms do not have to be distinct, but even if the same algorithm is executed several times, the executions are with different parameters. As a result of these executions, k partitions P_1 (M), P_2 (M), ... , P_K (M) resulted. Each partition in M equivalence classes, P_i (M), is represented by the equivalence relation $u_i(x,y)$ with x,y patterns in E. It is considered $u_i(x,y)=1$ if patterns x and y are in the same class of the partition P_i (M) and $u_i(x,y)=0$ otherwise.

The central partition is the partition P*(M) which is located at the minimum distance from the partitions P_1 (M), P_2 (M), ... , P_K (M). The distance between 2 partitions P_i(M) and P_j(M) can be calculated [3] like this:

$$d_C(P^i,P^j)=\frac{1}{2}\sum_{(x,y)\in E\times E}\left|u_i(x,y)-u_j(x,y)\right|$$

To define strong patterns, the following equivalence relation is considered:

$$w^K(x,y)=\begin{cases}1 & if \quad \sum_{i=1}^{K} u_i(x,y)=K\\ 0 & else\end{cases}$$

Two patterns will be considered strong if $w^k(x,y) = 1$. We will denote by S the number of strong patterns. The equivalence relation $w^k(x,y)$ determines a partition into S equivalence classes. Π(S) of E, called the partition of hard patterns.

IV. Proposed Method

The whole process is described in the Figure 2 in which each partition P_i(M) was represented by a vector of integers, $\Upsilon_i[n]$ where n is the number of patterns, and $\Upsilon_i[j]$ contains the index of the class to which the pattern j belongs in the partition P_i(M).

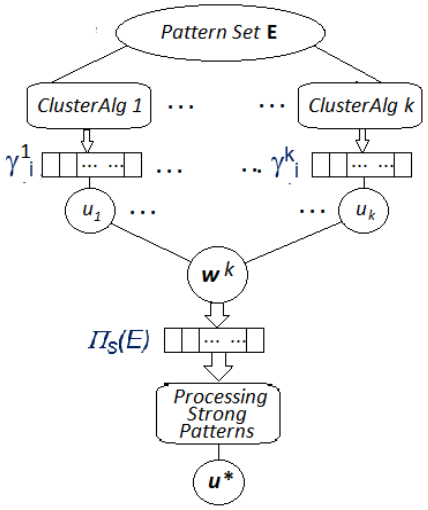

Figure 2. General flowchart of the method.

The partition Π(S) of E, in which the strong patterns are in fact the S equivalence classes, will constitute the input of the most suitable hierarchical ascending classification algorithm in the sense that is the hierarchical ascending classification algorithm that provided the closest partition P_j(M) to the central partition P*(M).

By applying the selected algorithm, the partitions Π(S-1), ..., Π(2) of E will be obtained (see Figure 3).

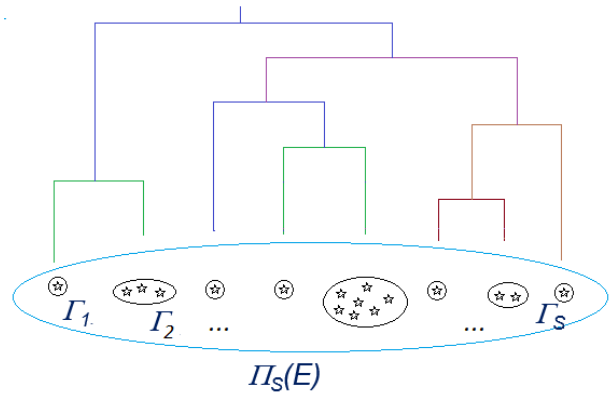

Figure 3. Strong patterns processing.

After applying this method, the user can choose the most suitable partition for the concrete problem to be solved, Π(m), with m in the interval [2, S}, where S is the number of strong patterns obtained after the k applications of some algorithms of clustering on the pattern set E.

V. LOAD PROFILE ANALYSIS

Load Profile represents the electricity consumption graph of a consumer in some period. As a rule, this graph represents the recording of the hourly consumption of electricity in 24 hours. In this case, a load factor will be represented by a pattern xz =(xz1, xz2, ..., xz24) corresponding to day z. The analysis was carried out on the daily Load Profiles from February to April [5] with the aim of determining the significant profiles of energy consumption.

By applying the proposed method, the partition Π(7) was selected, in which class C0 corresponded to some Load Profiles from the days when the substation point was revised, and class C2 contained only one pattern considered outlier. So a partition with 5 classes (Π(7)-{C0, C2}) was retained.

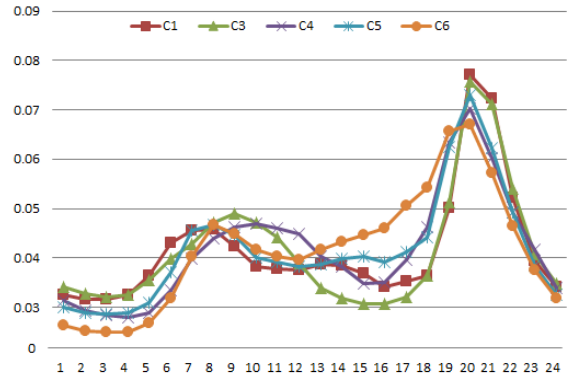

Figure 4. The Load Factor Profiles for the retained classes [6].

The result and the calendar distribution of the retained classes provided significant information [6] for the analysis of the electricity consumption from that substation point (see Figure 4). For example, load profiles from class C1 correspond to working days from the period starting on March 21, and load profiles from class C5 to working days from February to March 21. This fact also confirms the connection between the load profile and the day of the week, but also the atmospheric temperature.

VI. ANALYSIS OF HYPERSPECTRAL IMAGES

Hyperspectral imaging is an analytical technique based on spectroscopy that collects hundreds of images at different wavelengths for the same spatial area in the form of a 3D hypercube, where one dimension contains the spectral information and two dimensions contains the spatial details (see Figure 5 for representation of a Braila hyperspectral image taken from [10]) [11].

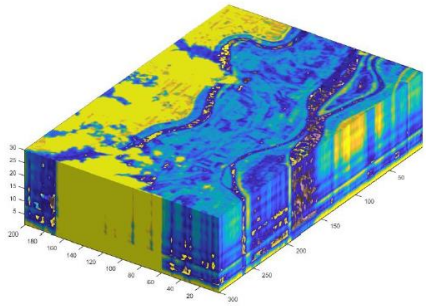

Figure 5. Three-dimensional representation of the hyperspectral image Braila data set.

Spectroscopy involves the study of light emitted or reflected by materials and its variation in energy with wavelength [12]. The property of absorption and emission of electromagnetic radiation varies depending on the material existing on the earth's surface from the visible spectrum to the NearInfraRed and the ShortWave InfraRed [13].

The classification of hyperspectral images involves the grouping of objects according to the characteristics they have by means of automatic learning algorithms, thus allowing the study of hard-to-reach areas, such as rocks, forests, relief, human settlements, roads [11]. The quality of the predictions depends on both the methods used and the spectral and spatial resolution because they are resource consuming. Spectral signatures represent the informational entities or shapes that characterize the pixels and represent the input data for machine learning algorithms. The class a shape belongs to is assigned based on its similarity to other shapes, e.g., the shapes they are most similar to [14].

Regarding the strong patterns, from the hyperspectral images the most representative pixels are chosen. These pixels are grouped using several unsupervised algorithms, such as hierarchical clustering, k-means and mean-shift [15]. We approached unsupervised learning because the

absence of groud truth for some hyperspectral images makes it impossible to use supervised learning algorithms. For each pixel, the labels assigned by the unsupervised learning algorithms were concatenated, thus building the multipartition matrix for all pixels. Strong patterns will be searched in the multi-partition matrix.

The goal is to find a consensus partition for the previously obtained multipartition, with the mention that this task is an NP-hard problem. In order to solve this task, the proposed heuristic involves the application of hierarchical grouping for strong patterns extracted from the multipartition matrix. In Figure 6 (right) is the graphical illustration of the extension of the classification to the entire Braila hyperspectral image.

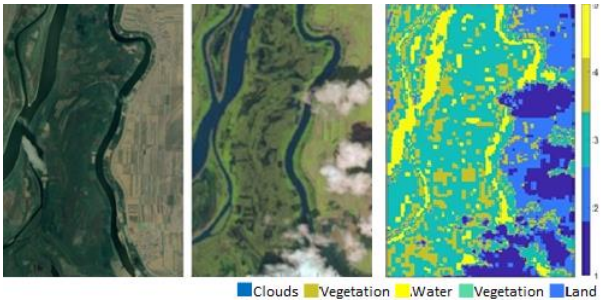

Figure 6. The true color of the Braila zone (an area of Braila's Small Island) in fact a satellite view available from Google Maps (left [1]), the true color provided by the EO-1–Hyperion satellite together (center [2]), and the hyperspectral image colored after clustering (right).

Since there is no general valid algorithm that generates a perfect partition, several partitions were built and the best consensus partition was sought by approaching the heuristic method that is achieved by hierarchical grouping of strong patterns. Strong patterns increased the classification accuracy because the obtained consensus partition is more homogeneous.

VII. Conclusions

The analysis starts from a set of raw data about which we have no a priori information about how it is structured. The analysis techniques are then the ones that take care of the context of unsupervised learning in Machine Learning.

The data set that includes n observations of p characteristics, thus resulting in a matrix of size nxp. These observations can be grouped unsupervised by several clustering algorithms. Of course, different algorithms, or even the same algorithm executed with different input parameters, will provide different partitions of the same data set. The problem arises of finding the best partition. One solution is given.

Two case studies are discussed. The first case study analyzes the outstanding patterns of Load Factor Profiles [3]. In the second case study [4], an attempt is made to classify and color the regions of an hyperspectral image, so that regions of the same nature (soil, water, vegetation etc.) are colored with same color.

Acknowledgment

We acknowledge the financial support provided by the project "Center for knowledge transfer to enterprises in the field of ICT - CENTRIC", Contract no. 5 / AXA 1 / 1.2.3 / G / 2018, Subsidiary contract no. 15.875 / 2021.

References

[1] Topchy, Alexander & Law, Martin & Jain, Anil & Fred, Ana. (2004). Analysis of Consensus Partition in Cluster Ensemble, Conference: Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), 1-4 November 2004, Brighton, UK225-232. 10.1109/ICDM.2004.10100.

[2] Vega-Pons, Sandro; Ruiz-Shulcloper, José (1 May 2011). "A Survey of Clustering Ensemble Algorithms". International Journal of Pattern Recognition and Artificial Intelligence. 25 (3): 337–372. doi:10.1142/S0218001411008683. S2CID 4643842.

[3] Celeux, G.; Diday, E.; Govaert, G.; Lechevalier, Y. and Ralambondrainy, H. (1989). Classification Automatique Des Donnees, Dunod, Paris.

[4] Scikit learn, Hierarchical Clustering https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

[5] Cartina, G., Grigoras, G., Bobric, E. C., & Comanescu, D. (2009, June). Improved fuzzy load models by clustering techniques in optimal planning of distribution networks. In PowerTech, 2009 IEEE Bucharest (pp. 1-6). IEEE.

[6] Crenguta Bobric and Stefan-Gheorghe Pentiuc 2018. "Clustering Techniques for Comparative Analysis of Load Factor Profiles", ATINER's Conference Paper Proceedings Series, ENG2018-0098

[7] Google Maps. Available online: https://www.google.ro/maps (accessed on 31 August 2020)..

[8] Earthexplorer.usgs.gov, U.S. Geological Survey, USGS. 2019. Available online: https://earthexplorer.usgs.gov (accessed on 31 August 2020).

[9] Laura Bianca BILIUS, and Pentiuc Ştefan Gheorghe. 2020. "Efficient Unsupervised Classification of Hyperspectral Images Using Voronoi Diagrams and Strong Patterns" Sensors 20, no. 19: 5684. https://doi.org/10.3390/s20195684

[10] USGS (U.S. Geological Survey), USGS EROS Archive - Earth Observing One (EO-1) – Hyperion, https://earthexplorer.usgs.gov/

[11] José Manuel Amigo, Carolina Santos, Chapter 2.1 - Preprocessing of hyperspectral and multispectral images, Editor(s): José Manuel Amigo, Data Handling in Science and Technology, Elsevier, Volume 32, 2019, Pages 37-53, ISSN 0922-3487, ISBN 9780444639776, https://doi.org/10.1016/B978-0-444-63977-6.00003-1.

[12] Randall B. Smith, ©MicroImages, Inc., 1999-2012, Tutorial: Introduction to Hyperspectral Imaging - MicroImages, Inc., https://www.yumpu.com/s/qydZUfFaS7aBxUP8

[13] Transon, Julie, Raphaël D'Andrimont, Alexandre Maugnard, and Pierre Defourny. 2018. "Survey of Hyperspectral Earth Observation Applications from Space in the Sentinel-2 Context" *Remote Sensing* 10, no. 2: 157. https://doi.org/10.3390/rs10020157

[14] Shabbir, Sidrah & Ahmad, Muhammad. (2021). Hyperspectral Image Classification -- Traditional to Deep Models: A Survey for Future Prospects. https://arxiv.org/pdf/2101.06116.pdf

[15] Bilius, Laura Bianca, and Stefan Gheorghe Pentiuc. 2020. "Unsupervised Clustering for Hyperspectral Images" Symmetry 12, no. 2: 277. https://doi.org/10.3390/sym1202027