Journal of Computational and Applied Mathematics 292 (2016) 623–633

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

The role of significance tests in consistent interpretation of nested partitions

Karina Gibert [a,b,c,*], Beatriz Sevilla-Villanueva [a,c], Miquel Sànchez-Marrè [a,c]

[a] Knowledge Engineering and Machine Learning Group (KEMLG), Spain
[b] Department of Statistics and Operation Research, Spain
[c] Universitat Politècnica de Catalunya-BarcelonaTech, Spain

ARTICLE INFO

Article history:
Received 22 October 2014

Keywords:
Clustering
Nested partitions
Statistical tests
Sensitivity of a test
Cluster interpretation
Consistency

ABSTRACT

Cluster interpretation is an important step for a proper understanding of a set of classes, independently of whether they have been automatically discovered or expert-based. An understanding of classes is crucial for the further use of classes as the basis of a decision-making process.

The abundant work on cluster validity found in the literature is mainly focused on the validation of clusters from the structural point of view. However, structural validation does not ensure that the clustering is useful, since meaningfulness is the key to guaranteeing that classes can support further decisions. In previous works, special significance tests taken from the field of multivariate analysis were introduced in an interpretation methodology for automatically assessing relevant variables in particular classes.

In this paper, we present the interpretation of *nested partitions* and the relationships between both interpretations are studied. In particular, the inconsistencies produced in interpretation when a second partition refines the first one with a higher level of granularity are studied, diagnosed, and a modification of the original methodology is provided to guarantee consistency in these cases. The relevant characteristics detected in a parent class must also be inherited in subclasses, or at least in some of them.

The proposal is evaluated using a real data set on baseline health conditions and dietary habits of a sample of the general population.

1. Introduction

When transmitting the results of data mining analysis to the end user, it is important that the results are completely understood [1]. A process of results interpretation provides new knowledge that can be used to support further decision-making. In fact, interpretation is crucial for proper knowledge transfer, especially to experts from other disciplines. In the context of clustering, which is among the most popular data mining methods, interpretation provides a proper understanding of the essence of the classes obtained, even if they have been automatically discovered or expert-based. Thus, a good characterization of the classes requires proper interpretation tools for post-processing the clusters themselves [2,3].

In general, the objective of clustering is to generate a set of different classes that group similar individuals in the same class. Therefore, individuals of the same class can be described by common characteristics, and some of these characteristics are expected to be different from other classes. *Cluster Interpretation* is a post-process of finding the common and distinctive

* Corresponding author at: Universitat Politècnica de Catalunya-BarcelonaTech, Spain.
E-mail addresses: karina.gibert@upc.edu (K. Gibert), bea.sevilla@gmail.com (B. Sevilla-Villanueva), miquel@lsi.upc.edu (M. Sànchez-Marrè).

http://dx.doi.org/10.1016/j.cam.2015.01.031
0377-0427/

characteristics of every class, and creating the corresponding profiles. Most of the work in the literature about analyzing clustering results is focused on cluster validity, and validity indexes are used to provide the structural validity of classes [4]. However, structural validity does not necessarily ensure the usefulness of clustering, as meaningfulness is also key to guaranteeing decision-making support. Usefulness and understandability are part of the characteristics required in a data mining solution according to the seminal paper [5]. However, cluster interpretation is still an open issue from the methodological point of view, and with respect to the automatization of this process, although there are few works on these topics. Some proposals [6] based on the analysis of conditional distributions among classes are already available in this line. In most of the works, concept induction [7], or statistical tests [8,9], are used to identify which variables behave differently in some of the classes.

An existing partition of a set of objects sometimes becomes enriched in a second moment when a refinement and a new nested partition is provided. This might correspond to a partition of higher granularity than the first, or to the incorporation of new information in the clustering process, or to a combination of two or more previous partitions of higher granularity in a Cartesian product [10] that produces nested classes considering criteria used in all the original partitions. The referenced interpretation proposals do not guarantee robust behavior in this case, and might produce contradictory interpretations, as variables that were relevant for characterizing parent classes may disappear from the description of the refined partition.

In this paper, the limitations of the original proposal are analyzed and a new interpretation methodology is proposed to guarantee consistent interpretation with future refinements of a current partition. A mixed methodology for cluster interpretation is proposed that is based on a mixture of interpretation-oriented visualization tools and significance tests. *Class Panel Graphs* [11] are used for the visualization, whereas, *Test-Value* [18] have been imported from the factorial analysis field of multivariate statistics. The use of Test-Value for cluster interpretation was introduced in [9]. In this paper, a contribution is introduced on how Test-Value needs to be generalized to better identify the class characteristics. A modification of the interpretation methodology itself is also introduced to guarantee the consistency of an interpretation with future refinements in a new nested partition. Sensitivity analysis methods have been used for this purpose.

The performance and applicability of this proposal is evaluated on a real data set of baseline characteristics, diet habits, and levels of physical activity of a sample of the general population. The proposed class interpretation process can contribute to identifying standard nutritional patterns in the general population and their association with health conditions and physical activity habits, which is aligned with new preventative 'healthy life-style' policies for a better health condition, especially in the long term and aging. Obtaining a clear interpretation of the nutritional patterns in the population will permit the establishment of dietary guidelines to increase the health of persons and reduce public health costs in the long term.

The structure of the paper as follows: related work is explained in Section 2. The proposed methodology, together with details of several relevant aspects, is then presented in Section 3. This procedure is applied over two nested partitions of the sample regarding nutritional aspects and resulting from a previous clustering. The result is then compared with the expert description in Section 4. Finally, the conclusions of this work are presented in Section 5.

2. Related work

Cluster interpretation is part of the post-processing step in the data mining process. The interpretation of clusters is an important function when presenting the results to experts. In the literature, the characterization or interpretation of the classes is also termed *Cluster Profiling* [12]. Cluster interpretation is usually made by examining the cluster centroids that are built as the average of variables inside each cluster. In [12] it is stated that clusters are distinguishable only if certain variables exhibit significantly different means in some clusters, at least from a data perspective. This significance is often assessed by comparing the clusters with independent t-test samples or ANOVA. In [13] the variables used in the clustering are ranked using the *logWorth* index, based on p-values which are usually assessed with the χ^2-Independence Test. Some efforts also rely on visualizing the variables through the resulting clusters. In [14], the cluster averages of the standardized variables are displayed in a parallel plot. In [15], the conditional probability of the variables against the cluster are represented for the categorical variables. In [6], the *Class Panel Graphs* are introduced as a graphical representation in the form of panels (containing variables in the columns and classes in the rows) and displaying the conditional distributions of the variables against the classes per column. This visualization is interesting for identifying the specific behavior of variables in a certain class and thus, for better understanding the meaning of the classes [16]. *Class Panel Graphs* has been successfully used in previous applications [11,16,17] to present results to experts in support of a class-conceptualization process, and to assess the profiles associated with the classes.

Principal Component Analysis (PCA) is one of the most known techniques in the field of multivariate analysis. A common practice in PCA is to interpret the principal components for eliciting latent variables that are implicitly measured in the data set and associated with factorial components. The contribution of the original variables to the principal components is used for this interpretation and the (*Test-Value*) is introduced by [18] to identify the main contributing variables to for a certain principal component. These tests are based on the comparison of means/percentages within the classes with respect to the global sample indicator. Although Test-Value comes from the multivariate analysis field, it can help in the interpretation of clusters—as will be shown in this work. In a previous work [9], the significances of the variables obtained with the Kruskal–Wallis test, or the χ^2-Independence test, were compared, according to the type of variable, with those obtained using Test-Value, and it was seen that for the clustering context, Test-Value subsumes the results of Kruskal–Wallis and the