Pattern Recognition 43 (2010) 2439–2455

Contents lists available at ScienceDirect

Pattern Recognition

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

An incremental nested partition method for data clustering

Jyrko Correa-Morris [a,*], Dustin L. Espinosa-Isidrón [b], Denis R. Álvarez-Nadiozhin [a]

[a] Mathematic Department, Faculty of Mathematic and Computer Sciences, Havana University, Cuba

[b] Pattern Recognition Department, Advanced Technologies Application Center, Havana, Cuba

ARTICLE INFO

Article history:
Received 2 October 2008
Received in revised form
11 December 2009
Accepted 27 January 2010

Keywords:
Nested partition
Data clustering
Incremental clustering

ABSTRACT

Clustering methods are a powerful tool for discovering patterns in a given data set through an organization of data into subsets of objects that share common features. Motivated by the independent use of some different partitions criteria and the theoretical and empirical analysis of some of its properties, in this paper, we introduce an incremental nested partition method which combines these partitions criteria for finding the inner structure of static and dynamic datasets. For this, we proved that there are relationships of nesting between partitions obtained, respectively, from these partition criteria, and besides that the sensitivity when a new object arrives to the dataset is rigorously studied. Our algorithm exploits all of these mathematical properties for obtaining the hierarchy of clusterings. Moreover, we realize a theoretical and experimental comparative study of our method with classical hierarchical clustering methods such as single-link and complete-link and other more recently introduced methods. The experimental results over databases of UCI repository and the AFP and TDT2 news collections show the usefulness and capability of our method to reveal different levels of information hidden in datasets.

1. Introduction

The development of technology and computing has enabled the processing of large datasets and every day it becomes more necessary to have tools to carry out this task. The exploratory analysis of the data, looking for an underlying structure that allows to manipulate it more efficiently and effectively, is often an obligatory task. In this regard, data clustering is a powerful tool. The clustering approach can be divided into two main groups: non-hierarchical or partitional and hierarchical [1]. The non-hierarchical approach produces only one partition of data whereas the hierarchical approach produces a sequence of nested partition of data. The k-means algorithm [2], expectation maximization algorithm [3], based on graph theory algorithms such as β_0-connected components and β_0-compact sets [4], among others, are examples of non-hierarchical algorithms; whereas single-link algorithm [5,6], complete-link algorithm [7,6] and commute time for grouping [8], are some examples of hierarchical clustering algorithms. The number of clustering algorithms are reported in literature is large. However, neither a clustering algorithm nor a list of clustering algorithms exist that are capable of discovering the subjacent structure in any given data collection. Due to this, and to the little information with which in most occasions we count about the characteristics and generic properties of these methods is that becomes very difficult to choose one of them when we want to classify objects in a real given context. This problematic topic is referred to in literature as: user dilemma [1]. Besides that the clustering results, as several other pattern recognition tasks, can be affected by the data representation [9], the manner in which similarity between the objects is measured [9], assumptions made about the shape and the size of the clusters [10,11], and so on. Due to these reasons, data clustering is an ill-posed problem, and any prior knowledge about the data and the clustering algorithms could be decisive for achieving success in the development of this task [12].

In this paper, we focus on the hierarchical approach. "A hierarchical clustering method is a procedure for transforming a proximity matrix into a sequence of nested partitions" [13]. The hierarchical clustering algorithms [14–16] have a greater importance since they provide several data-views at different levels of abstraction. However, aside from the previously mentioned issues, there are two disadvantages in the majority of the traditional hierarchical methods [17]. Firstly, let us note that in the majority of these methods obtaining a specific hierarchy level is conditioned by all of the previous levels. Secondly, to obtain each hierarchic level, a sole clustering criterion is used. These two aspects, in many cases, reduce the functionality of these hierarchical methods and limit its applications. On [18] the authors declare some of the deficiencies of the hierarchical methods to face regionalization problems, which are linked to the previously mentioned aspects. We particularly have various reasons to claim that these aspects are really disadvantageous for

[*] Corresponding author.
E-mail addresses: jyrkoc@gmail.com (J. Correa-Morris), despinosa@cenatav.co.cu (D.L. Espinosa-Isidrón), nadiozhin@gmail.com (D.R. Álvarez-Nadiozhin).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.01.019

these methods. The first aspect is entirely related with efficiency, in the sense that it relates to computational costs and the algorithm execution time. If, in a given situation, we were interested in obtaining a specific level of hierarchy which can be obtained independently, we can optimize the problem resolution. On the other hand, the second aspect is more related with the efficacy of the method. In order to illustrate the idea we have, we need to first answer the following question: What is the role of the similarity function, and what is the role of the clustering criterion in the process of unsupervised classification? The measurement of similarity is responsible for quantifying the "alikeness" between the objects by looking at the features that the specialist in the area considers as determinant. On its part, the grouping criterion is in charge of utilizing the measurement of similarity to discover certain common properties in sub-collections of objects that are differentiated from the rest and give place to the formation of clusters. As such, if one criterion is utilized to obtain each one of the levels of the hierarchy, then the interpretation of two different levels of the hierarchy is limited to the measurement of similarity. Since only one clustering criterion is utilized, to obtain two levels of the hierarchy one must increase or decrease the thresholds of similarity to consider; whereas, if various criteria are used along with the thresholds of similarity, we have the information that gives us each one of the corresponding levels of criteria. Further, if the relationships between the criteria are known, then we are able to obtain more diverse information of the relations between the different levels. Needless to say, utilizing different levels permits us to better explore the measurement of similarity searching for links between the objects within the same level of the hierarchy; further, it permits us to explore the inner level relations. Because of these reasons, we begin to think of algorithms whose fundamental objective is to obtain a sequence of nested partitions that will also incorporate the functions mentioned above. Although these algorithms are a particular case of hierarchical algorithms, we will refer to them as nested partition algorithms to reassure that different criteria can be utilized to obtain the hierarchy as well as each individual level within the hierarchy can be obtain independently.

Analogous methods have been reported in literature for optimization issues [19–21]. The main goal of *nested partition methods for optimization* problems is to accelerate the search for the global optimum. With this aim, the properties of the target functions and the feasible region are used in order to focus the greatest computation effort on those regions which there are higher possibilities of global optimum is. Those methods have been used in data mining and pattern recognition problems [22–25]. In [22] the nested partition methods were used for variable selection problems, whereas in [23,24], are applications for texture analysis and speaker recognition, respectively. In [25] a study of common aspects between data mining and operation research is done. What is common to all these applications of nested partition methods for optimization in data mining and pattern recognition tasks is that all these problems have to be conceived as explicit optimization problems.

In this article we propose a nested partition algorithm to solve problems of unsupervised classification. Given that we keep in mind to apply it to document analysis, such as news and polls, in which the databases have a large quantity of data and updates occur frequently, we have decided to present an incremental version of this algorithm. Our method is based upon different clustering criteria, which have been previously alluded in literature [4], as well as utilized as the basis for the development of various incremental algorithms [26–28]. The principal predecessor of this work can be found in [17] in which a particular case of the discussed algorithm is exposed, can be considered the root of this methodology. With the intention of clarifying the general properties of each one of these criteria, as well as the relationship that exist within each other, we conducted a study whose results we presented in form of lemmas, propositions, and theorems. Not only does this method formalize the results, but it also allows understanding, in detail, the function of the algorithm and what is behind every step of it. In our opinion, this can help in deciding whether or not it is convenient to use in a determined problem.

In addition to the Introduction, this paper is organized into six sections as follows. In Section 2 some definitions and basic notions are presented. Section 3 is dedicated to the discussion of the main property and relationships of the clustering criteria on which our method is based. The study of the sensitivity of these criteria to the addition of new objects to the dataset is made in Section 4. In Section 5 the algorithm is detailed and its pseudocode is exposed. The experimental results with several dataset of different nature are presented and discussed in Section 6. The last section is devoted to the concluding remarks.

2. Similarity spaces

Suppose that $\mathcal{U}'$ is a set of real objects (data universe) and through a mathematic modeling process is obtained a set $\mathcal{U}$ of object descriptions in terms of a feature set $R=\{f_1,f_2,\ldots,f_n\}$. That is, there is an operator $I:\mathcal{U}'\to\mathcal{U}$ which associates to each object O its description $x(O)$ in terms of features set R; being $x(O)$ the mathematic entity which represents the real object O. This process of mathematic modeling is very important in every task of pattern recognition.

Once the objects are represented, we have to find a manner to measure the similarity between the objects (similarity function). Formally, let $\mathcal{S}\subseteq\mathcal{U}'$ be an object sample set and $\mathcal{X}\subseteq\mathcal{U}$ is the set of its representations, then a function $\Gamma:\mathcal{X}\times\mathcal{X}\to L$ is called a similarity (dissimilarity) function if and only if Γ satisfies the following conditions:

1. L is a field (see [29]);
2. there is a total order relation $\leq$ defined on L which is compatible with the field structure of L (usually $L=\mathbb{R}$ and $\leq$ is the less than relation);
3. $Range(\Gamma)\subseteq L$ has a least element m and a greatest element M;
4. $x=y\Rightarrow\Gamma(x,y)=M(=m)$.

Besides, if $\forall x,y\in\mathcal{X}$, $\Gamma(x,y)=\Gamma(y,x)$ it says that Γ is a symmetric similarity function. The pair $(\mathcal{X},\Gamma)$ is called *similarity* (*dissimilarity*) *space* and $\mathcal{S}$ is called the *support* of $(\mathcal{X},\Gamma)$. In this paper, we only consider similarity spaces $(\mathcal{X},\Gamma)$ such that the similarity function Γ is symmetric.

For each similarity (dissimilarity) function a $\beta_0\in L$ must exist such that if $x,y\in\mathcal{X}$ and $\Gamma(x,y)\geq\beta_0(\leq\beta_0)$, then x and y are very similar objects and reciprocally.

3. Clustering criteria: a nested partition

Given a graph G=(V,E) we mean by path a sequence of vertexes x_1, x_2,..., x_m such that for every i from 1 to $m-1$ it has $(x_i,x_{i+1})\in E$. Henceforth, we use the following notations:

- If G is a directed graph (the edges have an orientation), the arrow $(x,y)\in E$ is denoted by $\overrightarrow{xy}$ and $x\to y$ meaning that exists an arrow from x to y. Moreover, the set of directed paths is denoted by $DP(G)$. An element of $DP(G)$ connecting the elements x and y is denoted by p, and $o(p)$, $d(p)$ denoting the origin and destiny of p, respectively.