Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Adaptive Mean-Linkage with Penalty: A new algorithm for cluster analysis

Marcos Dósea [a], Leila Silva [a], Maria A. Silva [b], Sócrates C.H. Cavalcanti [b,*]

[a] *Departamento de Ciência da Computação e Estatística, Universidade Federal de Sergipe, Campus Universitário Prof. José Aloísio de Campos, Rosa Elze, 49100-000, São Cristóvão - SE, Brazil*
[b] *Departamento de Farmácia, Universidade Federal de Sergipe, Campus Universitário Prof. José Aloísio de Campos, Rosa Elze, 49100-000, São Cristóvão - SE, Brazil*

ARTICLE INFO

Article history:
Received 19 July 2006
Received in revised form 18 July 2007
Accepted 29 May 2008
Available online 5 June 2008

Keywords:
Cluster analysis
Chemosystematics
Essential oil composition
Clustering

ABSTRACT

A novel cluster analysis technique, so-called Adaptive Mean-Linkage with Penalty algorithm (AMLP) is proposed. The method is based on a penalty concept applied to the Euclidian distance, which determines the dissimilarity among objects when clustering data. The implementation of this technique is straightforward and provides enhanced classification in our case studies. The proposed clustering procedure was applied to a dataset of compounds from the essential oil of plants acquired for classification purpose. The potentiality of this novel technique to distinguish each plant or group of plants according to the concentration levels of compounds in the essential oil has been validated. A free web tool is available.

1. Introduction

In general, there are two approaches to classify data: supervised and unsupervised. In the former, a set of training data is available and the classifier is designed by exploiting this *a priori* known information. In the later, the training data is not available and the data, represented as vectors, are grouped based on their “similarity”.

Cluster analysis is an unsupervised learning technique that examines the interpoint distances between all samples. The main goal is to extract some sort of organizational entities from datasets. It is a technique used for combining observations into groups or clusters such that each group is homogeneous, i.e. the objects within each group are similar to each other and each group is different from the other groups [1].

Clustering techniques have been classified as hierarchical and non-hierarchical methods. In hierarchical techniques, one can proceed in an agglomerative or divisive way. In an agglomerative process, the principal aim is to join similar objects into clusters and to add objects to clusters already found or to join similar clusters. In divisive strategies one starts with one cluster comprising all objects from which, systematically, the most non-homogeneous objects are stripped, forming themselves into more homogeneous clusters at lower linkage levels. A non-hierarchical method generates a classification by partitioning a dataset, giving a set of non-overlapping groups having no hierarchical relationships between them. A systematic evaluation of all possible partitions is quite impractical, and many different heuristics have been described to allow the identification of good, but possibly sub-optimal, partitions. Our approach is based on hierarchical methods.

In general, when using hierarchical clustering, a two-dimensional plot called a dendrogram is used to represent the achieved results [2]. The dendrogram presents high-dimensional data in a suitable form, simplifying the use of human pattern recognition ability, especially in large, complex datasets.

To generate the dendrogram, cluster analysis methods are based on vectors nearness in a *n*-dimensional space. The simplest similarity measure can be derived from geometry, based on distance measures. The shorter the distance between objects the more similar they are [3]. However, the distance is a dissimilarity measure. A common approach applies Euclidian distance as a dissimilarity measure, but several proximity metrics can be defined [4].

The classical agglomerative Hierarchical Clustering Analysis (HCA) algorithm initially considers every object as a cluster. The Euclidean distance between each pair of clusters is calculated. The two closest objects or clusters are joined, generating a new cluster. This process is repeated until only one cluster remains.

Many clustering methods have been optimized for a particular application. An approach proposed by DeGaetano [5] uses non-hierarchical cluster analysis to reduce bias introduced by both redundant and irrelevant climatic data. Functional cluster analysis was used as a new tool to evaluate perfusion brain imaging in order to identify normal brain, ischemic tissue and large vessels [6]. Fuzzy sets theory, introduced by Zadeh [7] have been applied to cluster analysis

[*] Corresponding author. Tel.: +55 79 3212 6641; fax: +55 79 3243 7457.
E-mail address: socrates@ufs.br (S.C.H. Cavalcanti).

0169-7439/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2008.05.007

aiding on the classification of ill defined clusters [8]. Such method allows overlapping clusters with partial membership of individuals in clusters. An agglomerative hierarchical cluster technique referred to as the Adaptive Mean-Linkage algorithm (AML) was proposed by Magalhaes *et al.* and applied to Quantitative Structure-Activity Relationship (QSAR) [9]. This method introduces the concept of neighboring and cut-off distances, allowing clustering more than two objects in one step.

Generally, the existing implementations differ in the way how distances of clusters are computed, aiming to separate or penalize dissimilar clusters. Biological data is usually ill defined and not handled well by the simpler concepts in classical agglomerative hierarchical clustering. In some applications, such as the chemical classification of plants, the lack of a compound in the essential oil composition of a plant may indicate that the plant should be classified or grouped in a different cluster. In such cases, the object may be penalized with the goal to increase its dissimilarity.

The present work describes a new approach of data clustering based on the AML algorithm, so-called Adaptive Mean-Linkage with Penalty algorithm (AMLP). The method is based on a penalty concept applied to the Euclidian distance, which determines the dissimilarity among objects when clustering the data. Our aim was to develop a clustering method not only to group samples based on the inner variability among objects, but also to penalize objects carrying discrepant variables. Two objects are highly discrepant for a variable if the difference between the values of this variable is greater than an empirical threshold established for the problem, or if at most one of these values is zero. An integer number, whose square root is multiplied by the Euclidean distance, represents the penalty. The precise definition of penalty is given in the mathematical section.

To evaluate the proposed method we have considered a dataset comprising 87 plants. Each plant is represented by a vector of its essential oil composition. The number of compounds in each plant is variable. The information used has been extracted from the literature. Based on this dataset we have considered 100 case studies. Each case study comprises two genera of plants; each one consisted of three to twenty plants.

Furthermore, we have modified the HCA algorithm to include the penalty concept aiming to compare the proposed method. We name this new variant as Agglomerative Hierarchical Clustering Analysis with Penalty (HCAP). The proposed algorithm has been compared with HCA, HCAP and AML algorithms and the achieved results are herein summarized.

2. Mathematical and algorithmic descriptions of AMLP and HCAP algorithms

Clustering algorithms are largely used to group individuals based on similarity criteria. The addressed problem may be stated as follows.

Problem. Let $P_1P_2,\ldots P_n$ be n objects to analyze and $C_1C_2,\ldots C_m$ be m variables to consider in each object. The goal is to group objects whose variable values obey the similarity criterion.

In this section, we describe the AMLP and HCAP algorithms. The only difference between these algorithms compared to AML and HCA algorithms, respectively, is the introduction of the penalty concept (Definition 1) in the dissimilarity criterion. This concept captures the fact that, in some cases, such as the study of variability of plant essential oils, the lack of one variable means the object needs to be classified in a separate group.

Definition 1. Let P_i and P_j be two objects with variables $a_{k1},a_{k2},\ldots a_{km}$, $k=I,j$. Each variable a_{ku}, $1 \le u \le m$, records the value of the variable C_u in the object P_k. The *penalty* g_{ij} is defined as the number of *discrepant* variables of P_i and P_j. A variable is said to be *discrepant* if either it is present in only one of the objects (for a given $u, a_{iu}=0$ and $a_{ju} \neq 0$, or vice-versa) and/or the difference of its value in these objects is greater than a user-established threshold t for the problem ($|a_{iu}-a_{ju}| > t$).

Both algorithms are based on the analysis of a matrix $A(n \times m)$, where each element $a_{ij}, 1 \le i \le n, l \le j \le m$, expresses the value of the variable C_j in the object P_i, as defined previously. In the case of plant essential oil variability, a_{ij} expresses the concentration of the essential oil compound C_j in the plant P_i.

The dissimilarity criterion between two objects P_i and P_j, common to both algorithms, is expressed by:

$$\mathrm{dsim}_{ij} = \sqrt{\left(\sum_{1 \le u \le m} (a_{iu} - a_{ju})^2\right) \times g_{ij}}.$$

Notice that as much *as* P_i and P_j have discrepant variables, higher will be the values of g_{ij} and dsim_{ij}. Thus, plants with several discrepant variables are considered to be more dissimilar.

It is relevant to mention that the choice of t in Definition 1 can have major impact in the quality of the results achieved. This fact motivates the investigation of three variations for discrepant variables. The first variation (V1), applies the penalty for mismatched compounds, by adding one unity to the value of g_{ij} for each mismatched compound. The second variation (V2) compares the concentration of compounds from two plants and applies the penalty when the difference between compounds concentrations is greater then a user-defined value (1% in our methodology). The third variation (V3) represents V1 and V2 applied simultaneously, applying the penalty when the compound is either mismatched or the difference between significant compounds concentrations from two plants is greater than a user-established value.

Both algorithms follow a general pattern, called Generalized Agglomerative Scheme (GAS). The HCAP and AMLP algorithms differ, basically, in two issues: the number of objects merged in each execution of the GAS pattern and the procedure used to merge these objects. To abstract these differences, a function *similarObjects* is used when describing the pattern. This function receives as input the current matrix A and returns the set S:=$\{S_1,S_2,\ldots,S_1\}$, where each $S_k, 1 \le k \le l$ is a set of objects to merge. In HCAP, $|S|=1$ and $|S_1| = 2$, that is, only two objects are merged in each execution of GAS. On the other hand, in AMLP $2 \le |S| \le n/2$, where n is the number of objects available (the current number of rows of A), and $2 \le |S_k| \le n$ Thus, the number of objects to be merged could vary in each execution of GAS; the worst case occurs when AMLP behaves as HCAP.

The GAS algorithm first constructs the matrix A, considering all objects. In each iteration, sets of similar objects to group are defined, each set having at least two elements. A set S_k of similar objects is then represented by a new merged object. Each variable in the merged object is the mean of the values of the corresponding variables of objects in S_k Matrix A is then updated to include the merged object and to delete the objects that have been merged. The procedure is repeated until the matrix becomes a vector (has only one row, representing only one cluster). The GAS algorithm is further summarized, by using an abstract high level language.

Algorithm GAS(A);

{* Input: Matrix A, where each row represents an object to group. The columns represent the value of the variables of each object. Output: a dendrogram representing the procedure of grouping objects. Variables i, j, k, q and z are used as indexes variables. T is a tree that stores the partial hierarchical structure of the objects to construct the dendrogram. Dsimdsim is a square matrix of order n that stores the dissimilarities between objects. *}

```
1. For each pair of rows $i$ and $k$, calculate $\mathrm{dsim}_{ik}$ and stores in a
   dissimilarity matrix dsim.
2. Let $S$=similarObjects($A$, dsim). Consider $S = \{S_1,S_2,\ldots,S_1\}$ and $S_k=\{P_{q1,}$
   $P_{q2,\ldots,}P_{qr}\}$, $1 \le k \le l$, $2 \le r \le n$.
3. For each $S_k, 1 \le k \le l$,
3.1  Remove rows $q_1, q_2,\ldots, q_r$ from $A$ and dsim.
```