Expert Systems with Applications 38 (2011) 12684–12690

Contents lists available at ScienceDirect

Expert Systems with Applications

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

OWA-based linkage method in hierarchical clustering: Application on phylogenetic trees

Efendi Nasıbov [a,b,*], Cagin Kandemır-Cavas [a]

[a] Department of Computer Science, Faculty of Sciences, Dokuz Eylul University, Tinaztepe Campus, 35160 Izmir, Turkey
[b] Institute of Cybernetics, Azerbaijan National Academy of Sciences, 9, F.Agayev str., AZ-1141 Baku, Azerbaijan

ARTICLE INFO

Keywords:
Hierarchical clustering
OWA operator
Protein sequence
Phylogenetic tree

ABSTRACT

The linkage methods are mostly used in hierarchical clustering. In this paper, we integrate Ordered Weighted Averaging (OWA) operator with hierarchical clustering in order to find distances between clusters. In case of using OWA operator in order to find distance between clusters, OWA acts as a generalized case of single linkage, complete linkage, and average linkage methods. In order to illustrate the proposed method, we handle a phylogenetic tree constructed by hierarchical clustering of protein sequences. To illustrate the efficiency of the method, we use 2D-data set. We obtain graphs demonstrating the relationships of the clusters and we calculate the root-mean-square standard deviation (RMSSDT) and *R*-squared (RS) validity indices, respectively, which are frequently used to evaluate results of the hierarchical clustering algorithms.

1. Introduction

Clustering is an unsupervised learning technique that aims at decomposing a given set of elements into clusters based on similarity. The basic goals are to divide dataset in such a way that elements are homogenous within groups and are different between groups.

Since vast amounts of data has rapidly increased in bioinformatics field because of genomic research, one need to use advanced computational tools to analyze and manage the data. Clustering algorithms have been widely applied for managing high-throughput data sets in bioinformatics, including DNA and protein sequence data analysis (Baldacci, Golfarelli, Lumini, & Rizzi, 2006; Chan, Collins, & Kasabov, 2006; Chang & Halgamuge, 2002; Lin & Chien, 2009).

Protein sequences that have evolutionary relationship constitute a family. That is generally reflected by sequence similarity. Therefore, all protein sequences can be organized based on their sequence similarity. Since the aim of protein clustering is to get a biologically meaningful partitioning, a graphical illustration called phylogenetic tree can summarize the relationship between the protein sequences. The methods existed on construction of phylogenetic tree are as follows: Neighbor-joining based (Bruno, Socci, & Halpern, 2000; Zhang & Sun, 2008); maximum parsimony based (Hill, Lundgren, Fredriksson, & Schio, 2005; Sridhar, Lam, Blelloch, Ravi, & Schwartz, 2007) and maximum likelihood based (Hobolth & Yoshida, 2005; Yang, 1997) and distance based (Lian, 2000; Sumner & Jarvis, 2006). A distance based phylogenetic tree is related to hierarchical clustering. The distance between objects can be calculated by linkage methods that the most common and cheap computational methods to divide dataset into clusters; such as single linkage, complete linkage and average linkage. In this study, we analyze the construction of phylogenetic tree based on Ordered Weighted Averaging (OWA) operator, which is most commonly used operator in multicriteria decision-making (Yager, 1988), as a linkage method.

In this paper, the general aspect of hierarchical clustering and OWA operator, in addition integration of OWA in hierarchical clustering is given in Section 2.1. Protein sequence alignment and phylogenetic trees are referred in Sections 2.2 and 2.3, respectively. The validity indices are concerned in Section 2.4. Results and discussion of the study are summarized in Section 3.

2. Methods and materials

2.1. OWA (Ordered Weighted Averaging) operator

Yager (1988) introduced an ordered weighted aggregation (OWA) operator to aggregate distributed information. The OWA operator plays important role in decision making problems (Nasibov & Nasibova, 2005, 2010; Okur, Nasibov, Kilic, & Yavuz, 2009; Yager, 1988). Since aggregating functions is formed for the situation in which all desired criteria are satisfied and the case in which the satisfaction of any of the all desired criteria exist. An

* Corresponding author. Tel.: +90 232 4128600; fax: +90 232 4534265.
E-mail addresses: efendi_nasibov@yahoo.com (E. Nasıbov), cagin.kandemir@deu.edu.tr (C. Kandemır-Cavas).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.055

aggregation which lies in between these two extremes is provided by this operator. Majority of the known averaging operators are special cases of the OWA operator (Yager & Kacprzyk, 1999). OWA differs from classical weighted average in that coefficients are not associated directly with a particular attribute but rather to an ordered position.

Definition. A mapping F: $R^n \rightarrow R$ is called OWA operator of dimension n associated with a weighting vector $\mathbf{W} = (w_1, w_2, \ldots, w_n)^T$ if

$$F(a_1, a_2, \ldots, a_n) = w_1 a_{(1)} + w_2 a_{(2)} + \cdots + w_n a_{(n)} \equiv OWA_W(a_1, a_2, \ldots, a_n), \tag{1}$$

where $a_{(i)}$ is the ith largest element in the collection of $a_1, a_2, \ldots, a_n$. The weighting vector $\mathbf{W}$ satisfies the following constraints:

1. $w_i \in [0, 1], 1 \leqslant i \leqslant n$.
2. $\sum_{i=1}^{n} w_i = 1$.

Let $\mathbf{B} = (a_{(1)}, a_{(2)}, \ldots, a_{(n)})^T$ be the vector consisting of the arguments of F in descending order. The OWA operator F with weight vector $\mathbf{W}$ and an argument tuple $(a_1, a_2, \ldots, a_n)$ can be rewritten as follows:

$$OWA_W(a_1, a_2, \ldots, a_n) = W^T B. \tag{2}$$

There are different approaches to determine weights of the OWA operator (Filev & Yager, 1998; Fuller & Majlender, 2003; Xu, 2005). In the paper Xu (2005), the weight vector $\mathbf{W} = (w_1, w_2, \ldots, w_n)^T$ of the OWA operator is calculated as,

$$w_i = \frac{\frac{1}{\sqrt{2\pi\sigma_n}} e^{-[(i-\mu_n)^2/2\sigma_n^2]}}{\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_n}} e^{-[(j-\mu_n)^2/2\sigma_n^2]}} = \frac{e^{-[((i-\mu_n)^2/2\sigma_n^2)]}}{\sum_{j=1}^{n} e^{-[(j-\mu_n)^2/2\sigma_n^2]}}, \quad i = 1, 2, \ldots, n, \tag{3}$$

where the mean μ_n and standard deviation σ_n used, in the previous formula, are computed, respectively, as follows:

$$\mu_n = \frac{1}{n} \cdot \frac{n(1+n)}{2} = \frac{1+n}{2}, \tag{4}$$

$$\sigma_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (i - \mu_n)^2}. \tag{5}$$

It is obvious that the conditions $w_i \in [0, 1]$ and $\sum_{i=1}^{n} w_i = 1$ will be satisfied. We will use the normal distribution function (3) to determine the weights of OWA operator in our experiments.

2.2. Hierarchical clustering

The hierarchical clustering method generates hierarchical nested partitions of the dataset, using a dendrogram and some termination criterion similarity or dissimilarity matrix is constructed between every pair objects.

Hierarchical clustering algorithm steps can be ordered as follows:

Step 1. Construct n clusters each of them has only one object.

Step 2. While number of clusters is greater than 1, repeat the steps 2a–2g:

> Step 2a. Find the distances between each pair of the objects of the clusters and construct the distance matrix **d** where element d_{ij} is the linkage distance between the clusters C_i and C_j.
> Step 2b. Merge the clusters that are closer to each other (suppose they are C_1 and C_2) into a new cluster C with their elements as $C_1 \cup C_2$.
> Step 2c. Find the distance between the cluster C and the remaining clusters.
> Step 2d. Delete the row and the column of the distance matrix **d** corresponding to the clusters C_1 and C_2.
> Step 2e. Mark the cluster C as C_1 and place a new row with distances between the C_1 and the remaining clusters into the distance matrix **d**.
> Step 2f. Decrease one the number of clusters.

Step 3. Stop.

In hierarchical clustering, the closer two clusters are identified and merged together as a new cluster (Keedwell & Narayanan, 2005). Single linkage, average linkage and complete linkage are the current methods to compute the distance between new constructed cluster and old one. All these mentioned methods take into account the unweighting distance. Many measures have been proposed for calculating the distances; fuzzy distance (Lian, 2000), relative root mean square (Betancourt & Skolnick, 2001), Lempel-Ziv complexity (Otu & Sayood, 2003). However, we use Ordered Weighted Averaging (OWA) operator to identify the distance value of the new merged clusters.

We mentioned above the steps of a hierarchical clustering. Step 2a has performed by computing distances (similarities) between the new cluster and each of the old clusters. It is obvious that the step 2a can be done in different ways, which can be single-linkage, complete linkage, average-linkage and so on.

2.2.1. Single linkage

The distance between two clusters is equal to the shortest distance from any member of one cluster to any member of the other cluster.

$$d_{\min}(C^*, C) = \min_{x \in C^* y \in C} d(x, y). \tag{6}$$

2.2.2. Complete linkage

The distance between two clusters is equal to the greatest distance from any member of one cluster to any member of the other cluster.

$$d_{\max}(C^*, C) = \max_{x \in C^* y \in C} d(x, y). \tag{7}$$

2.2.3. Average linkage

The distance between two clusters is equal to the average of the distance from any member of one cluster to any member of the other cluster.

$$d_{avg}(C^*, C) = \frac{1}{|C^*||C|} \sum_{x \in C^* y \in C} d(x, y). \tag{8}$$

In this study, distance between clusters is calculated with Ordered Weighted Averaging (OWA) operator. Therefore, distance between all pairs (x, y) where $x \in C^*$ and $y \in C$ are calculated as $d(x, y) \equiv d_i$, $i = 1, 2, \ldots, z$, $z = |C^*| \cdot |C|$. Then distance between two clusters are obtained as

$$d_{OWA}(C^*, C) = OWA_W(d_1, d_2, \ldots, d_z) = \sum_{i=1}^{z} w_i d_{(i)}, \tag{9}$$

where the weights w_i, $i = 1, 2, \ldots, z$, of the OWA operator can be given directly or calculated according to the any distribution function.