Pattern Recognition Letters 32 (2011) 2163–2170

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Weighted association based methods for the combination of heterogeneous partitions

Sandro Vega-Pons *, José Ruiz-Shulcloper, Alejandro Guerra-Gandón

Advanced Technologies Application Center (CENATAV), Havana, Cuba

ARTICLE INFO

Article history:
Available online 19 May 2011

Keywords:
Clustering ensemble
Similarity measure
Evidence accumulation
Object representation
Kernel function

ABSTRACT

Co-association matrix has been a useful tool in many clustering ensemble techniques as a similarity measure between objects. In this paper, we introduce the weighted-association matrix, which is more expressive than the traditional co-association as a similarity measure, in the sense that it integrates information from the set of partitions in the clustering ensemble as well as from the original data of object representations. The weighted-association matrix is the core of the two main contributions of this paper: a natural extension of the well-known evidence accumulation cluster ensemble method by using the weighted-association matrix and a kernel based clustering ensemble method that uses a new data representation. These methods are compared with simple clustering algorithms as well as with other clustering ensemble algorithms on several datasets. The obtained results ratify the accuracy of the proposed algorithms.

1. Introduction

Cluster analysis is an essential technique in a considerable number of practical problems in pattern recognition as well as in data mining. Clustering algorithms find a direct application in problems related with several engineering fields, computer, medical and social sciences among others (Everitt et al., 2001; Jain et al., 1999). Nowadays, there is a large number of clustering algorithms reported in the literature. Among the most widely used we can mention k-Means, expectation maximization, kernel based algorithms, hierarchical clustering algorithms like single-link and complete-link, the fuzzy *c*-Means, etc. (see Xu and Wunsch (2005)). However, as it is known, there is no clustering method capable of correctly finding the *underlying structure* for all datasets. Indeed, when there is no knowledge about the particular properties that we want to find or that we consider *good* in the data, it is not possible to speak in terms of the *underlying structure* of the data. There could be different data partitions such that, each one brings some useful information about the data organization in the problem at hand. This is because each clustering algorithm imposes a particular organization to the data. Thus, in these cases, an interesting problem is how to unify these data partitions in order to obtain a *consensus* one. In fact, clustering ensemble algorithms combine different partitions of the same dataset aiming at finding a consensus result.

Traditionally, given a set of objects, a clustering ensemble method consists of two principal steps: *generation*, where a set of partitions of these objects is obtained, and *consensus*, where all the generated partitions are combined to obtain the *consensus partition*. In the last years, many clustering ensemble methods have been proposed (Ghaemi et al., 2009; Vega-Pons and Ruiz-Shulcloper, 2011) e.g. evidence accumulation based methods (Fred and Jain, 2005), (hyper)graph partitioning based methods (Strehl and Ghosh, 2002), information theory based methods (Topchy et al., 2005), and kernel based methods (Vega-Pons et al., 2010). In the vast majority, the original set of objects is used to obtain the set of partitions, but only the set of partitions is used to obtain the consensus result. In other words, the peculiarities, the properties of the original set of objects are not explicitly used in the combination process.

Intuitively, we can say that if the original objects are available in the consensus step, they represent an extra information that may be useful for improving the quality of the consensus partition (see Fig. 1). However, if we try to work with the original set of objects in the consensus step, we have to take into account that, in the generation step, different representations of the objects and/or different (dis)similarity measures between objects could be used. Therefore, we have to deal with the question: which object representation and/or which (dis)similarity measure should be used in any data processing in the consensus step?

A possible answer to the above questions can be found from the definition of a procedure that immediately after the generation step summarizes the information in the clustering ensemble, i.e., somehow grouping in a unified concept the information about the possible different object representations, (di)similarity

* Corresponding author. Tel.: +537 271 4787; fax: +537 273 0045.
E-mail addresses: svega@cenatav.co.cu (S. Vega-Pons), jshulcloper@cenatav.co.cu (J. Ruiz-Shulcloper), aguerra@cenatav.co.cu (A. Guerra-Gandón).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.05.006

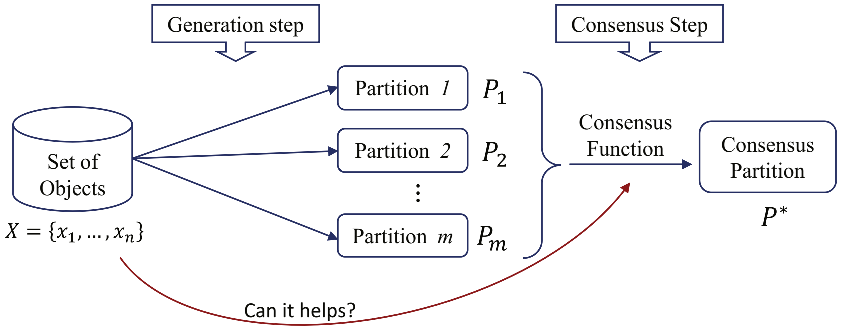

Fig. 1. Diagram of the clustering ensemble process. In the generation step a set of partitions is obtained and in the consensus step, the partitions are combined for obtaining the consensus result. A possible question is: could the original set of objects be somehow useful for the consensus step?.

measures used to obtain each partition, and the assignation of the objects to the different clusters in all partitions. After that, this unified concept could be used to obtain the consensus partition.

If we analyze the evidence accumulation based clustering ensemble methods (Fred and Jain, 2005) we can find a very similar idea. Evidence accumulation methods try to unify the information in the clustering ensemble into a new similarity matrix: the co-association matrix. This way, the co-association matrix is built from the set of partitions in the clustering ensemble, and the consensus partition is obtained from the co-association matrix. In this sense, co-association matrix seems to be the ideal tool for accumulating all information in the clustering ensemble. However, as we can see in more details in Section 2, there is valuable information that the traditional co-association matrix cannot extract from the clustering ensemble.

Thus, in this paper, we show how the use of the original set of objects after the generation step, can improve the consensus quality. Besides, based on the evidence accumulation philosophy, we present an effective way of using the original set of objects to improve the consensus process, thus giving a solution to the question presented above. As a result, we obtain two new clustering ensemble methods. Specifically, we introduce in Section 2 the *weighted-association matrix*, which is more expressive than the traditional co-association as a similarity measure between objects taking into account the information in the clustering ensemble. This new matrix will be the starting point of both clustering ensemble methods proposed in this paper. In Section 3, we present a clustering ensemble method based on the idea of evidence accumulation that uses the weighted-association matrix. In Section 4, we use the weighted-association matrix to obtain a new representation of the original objects. Due to this new data representation, we can work with the original set of objects in the consensus process no matter the generation mechanism applied. Besides, this new representation will allow the use of the mathematical tools for vectorial spaces that do not have to be available for the original data representation, e.g. when working with categorical or mixed[1] data. The clustering ensemble algorithm that uses this result is also presented in Section 4. Several experimental results comparing the proposed clustering ensemble algorithms with simple clustering algorithms as well as with other clustering ensemble algorithms over different databases are discussed in Section 5. Finally, Section 6 concludes this research.

Some initial ideas of the methods proposed here were introduced in our previous work (Vega-Pons and Ruiz-Shulcloper, 2009). However, in this paper, the general analysis of the problem is more complete, new results are presented and a better experimental study is carried out.

In this paper, we will use the concepts: similarity and dissimilarity measures, for which there is not an universally agreement. We assume a general definition of these concepts in the following way. Given a set S, a similarity (dissimilarity) measure is a bounded function $\Gamma : S \times S \to \mathbb{R}_+ (\pi : S \times S \to \mathbb{R}_+)$, such that $\Gamma(s,s) = M(\pi(s,s) = m)$ for all $s \in S$ where $M(m)$ is the maximum (minimum) value of the function. Moreover, the similarity (dissimilarity) concept is associated with the following intuitive idea: the higher the values of the similarity (dissimilarity) measure, the higher (lower) the likeness between the compared elements. Similarly, the lower the values of the similarity (dissimilarity) measure, the lower (higher) the likeness between the compared elements. If other properties are added to these measures, very common concepts can be obtained, e.g., a distance or metric is a dissimilarity measure π that for all s_1, s_2, $s_3 \in S$ satisfies:

- $(\pi(s_1,s_2) = 0) \Leftrightarrow (s_1 = s_2)$
- $\pi(s_1,s_2) = \pi(s_2,s_1)$
- $\pi(s_1,s_2) \leqslant \pi(s_1,s_3) + \pi(s_3,s_1)$

On the other hand, a similarity measure Γ is a *positive definite kernel*[2] (Scholkopf and Smola, 2002), if the following properties hold:

- For all s_1, $s_2 \in S$, $\Gamma(s_1,s_2) = \Gamma(s_2,s_1)$
- For all $t \in \mathbb{N}$, for all $s_1,\ldots,s_t \in S$ and for all sequence of real numbers $\alpha_1,\ldots,\alpha_t \in \mathbb{R}$:

$$\sum_{i=1}^{t}\sum_{j=1}^{t}\alpha_i\alpha_j\Gamma(s_i,s_j) \geqslant 0$$

We use the following notation in this document. Let $X = \{x_1,x_2,\ldots,x_n\}$ be a set of objects, where each x_i is a tuple of some $f-$dimensional space $\mathbb{G}^f$, for all $i = 1,\ldots,n$. $\mathbb{P} = \{P_1, P_2, \ldots, P_m\}$ is a clustering ensemble, where each $P_j = \{C_1^j, C_2^j, \ldots, C_{d_j}^j\}$ is a partition of the set of objects X with d_j clusters, for all $j = 1,\ldots,m$. $\mathbb{P}_X$ is the set of all possible partitions of X and the consensus partition is represented by P^*.

In this paper, we use (dis)similarity measures between objects, i.e., defined over the set X, and (dis)similarity measures between partitions, i.e., defined over $\mathbb{P}_X$. In particular, in different moments, we used kernel functions defined over X and $\mathbb{P}_X$.

2. Weighted association matrix

Fred et al. Fred and Jain (2005) proposed the evidence accumulation approach for clustering ensemble. The main idea behind this approach is to use the partitions in the clustering ensemble $\mathbb{P}$ to

[1] Composed by numerical and non-numerical attributes.

[2] For simplicity, we will refer to these functions as *kernels*.