Pattern Recognition 63 (2017) 427–436

Contents lists available at ScienceDirect

Pattern Recognition

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

Ensemble clustering based on weighted co-association matrices: Error bound and convergence properties

Vladimir Berikov[a,b,*], Igor Pestunov[b,c]

[a] *Sobolev Institute of Mathematics SB RAS, Novosibirsk 630090, Russia*
[b] *Novosibirsk State University, Novosibirsk 630090, Russia*
[c] *Institute of Computational Technologies SB RAS, Novosibirsk 630090, Russia*

ARTICLE INFO

Keywords:
Weighted clustering ensemble
Co-association matrix
Latent variable model
Cluster validity index
Ensemble size
Error bound
Hyperspectral image segmentation

ABSTRACT

We consider an approach to ensemble clustering based on weighted co-association matrices, where the weights are determined with some evaluation functions. Using a latent variable model of clustering ensemble, it is proved that, under certain assumptions, the clustering quality is improved with an increase in the ensemble size and the expectation of evaluation function. Analytical dependencies between the ensemble size and quality estimates are derived. Theoretical results are supported with numerical examples using Monte-Carlo modeling and segmentation of a real hyperspectral image under presence of noise channels.

1. Introduction

Cluster analysis is one of the important problems in data mining. Suppose we are given a data set $A = \{a_1, \ldots, a_N\}$ consisting of N objects. The description of this set can have two major forms: the table of feature observations, and the pairwise distance matrix whose elements are distances between object pairs. We consider the case when the information about the objects is presented as a data table $\mathbf{X} = (x_{i,m})$, where $x_{i,m} = X_m(a_i)$ is a value of feature $X_m \in \mathbb{R}$ for object a_i ($m \in \{1, \ldots, d\}$, $i \in \{1, \ldots, N\}$), d is feature space dimensionality.

It is required to find a partition $P = \{C_1, \ldots, C_K\}$ of the set A on a relatively small number of homogeneous subsets (groups, clusters). A certain function dependent on the scatter of observations within groups and the distances between clusters is usually understood as a criterion of homogeneity. The number of clusters K could be chosen beforehand or not defined (in the latter case it is necessary to find the optimal number of groups). In the given work, we assume that the number of clusters is fixed.

There exist a large number of clustering methods. These methods are characterized by different ways of understanding the notion of homogeneity, diverse procedures for searching the optimum partition and various problem-dependent restrictions (see an overview of existing methods in [1–3]). Since the brute-force methods are impractical, approximate iterative algorithms are usually used for finding the optimal clustering partition. In each step, these algorithms modify current partition in order to locally improve the clustering quality. This process is guided by certain user-specified parameters.

In last decades, an approach based on the collective decision making is actively developing in cluster analysis [4,5]. The interest to this research area has particularly grown in the light of achievements in the ensemble methods for classifying and forecasting [6,7].

It is known that clustering algorithms are nor universal; each one may have its own specific implementation area: some algorithms work better in situations when clusters are described by spherical regions; other algorithms are intended for strip-like clusters, etc. When data has complex structure, the reasonable strategy is using not a single algorithm, but a collection of different algorithms “compensating” each other's weaknesses [8]. The collective approach allows one to increase the quality of clustering in the situations when it is not clear, which of the algorithm parameters are most appropriate for resolving a particular problem. In this case, one may consider several partition variants (obtained for different parameters) and then make a final conclusion on their basis.

A number of approaches for finding ensemble clustering decision were suggested in the literature. In the consensus-style procedures, it is required to reach the optimal degree of consistency with the results of individual algorithms. Let us consider L variants $P_1, \ldots, P_L$ of partitioning defined on the set A. For the consensus partition P^*, we have

$$P^* = \arg\max_{P \in \mathbf{P}} \sum_{l=1}^{L} w_l \varphi(P, P_l),$$

[*] Corresponding author at: Sobolev Institute of Mathematics SB RAS, Novosibirsk 630090, Russia.
E-mail addresses: berikov@math.nsc.ru (V. Berikov), pestunov@ict.sbras.ru (I. Pestunov).

http://dx.doi.org/10.1016/j.patcog.2016.10.017
Received 29 June 2016; Received in revised form 20 September 2016; Accepted 15 October 2016
Available online 17 October 2016
0031-3203/ © 2016 Elsevier Ltd. All rights reserved.

where **P** is the set of all partitions of A, φ is a measure of similarity between two partitions, $w_l \geq 0$ is a "weight" of lth partition, $\sum_l w_l = 1$. The weight w_l, assigned to a clustering variant, allows one taking into account its "importance". The weight can depend on the estimation of partition quality with some cluster validity measure. Approximate iterative procedures are usually used for searching the optimal consensus partition.

Evidence accumulation approach [9] is based on the notion of averaged co-association matrix. The matrix defines how often object pairs fall into different (or the same) clusters over all clustering variants. Let H_l be the co-association matrix for lth variant of partitioning, where $H_l = (h_l(i, j))$; its element $h_l(i, j) = 0$ when objects a_i and a_j $(i \neq j)$ join the same group in lth variant; otherwise $h_l(i, j) = 1$. Averaged co-association matrix obtained over all variants of partitioning is defined as follows:

$$\mathbf{H} = \sum_l w_l H_l.$$

The elements of averaged matrix can be considered as the analogs of pairwise distances between objects: the higher the value of an element, the more frequently the pair was assigned to different clusters; i.e. the more dissimilar are the objects in this sense. To obtain the final partition, one can apply algorithms which make use of pairwise distance matrices, such as hierarchical agglomerative algorithm or spectral clustering.

Besides the above mentioned approaches for ensemble clustering, there exist other methods, such as analysis of distribution model mixtures, graph-theoretic algorithms, bootstrap samples analysis [4,5,10]. Consensus clustering approach was theoretically substantiated in [11] for the case of equal algorithms' weights. Applying the central limit theorem, the authors proved that an increase in the ensemble size decreases the probability of disagreement between consensus partition and the "true" partition (under the assumption that each base algorithm has better quality than a trivial algorithm of partitioning at random).

In real clustering problems, the ensemble size is always finite and the assumptions lying at the basis of limit theorems can be violated. Nevertheless, it is desirable to insure the best achievable quality of the ensemble. To this objective, various methods based on the evaluation of the contribution made by different variants into the overall solution were suggested. In these methods, partitions with higher grade receive greater weights in the final decision. The attributing of weights can be performed in different ways [12,13].

In *ensemble selection* methods, the resulting partition is obtained using a subset of base partitions. The candidate partitions are ranked according to some criterion; the given number of best variants are selected in the ensemble. This procedure is equivalent to attributing zero weights w_l to low-grade partitions and constant weights to the selected ones. Experiments show that the optimal performance of the ensemble is reached then the selection procedure takes into account both quality and diversity of base partitions [14,15]. Here the partition quality is determined with the information on cluster labels assigned to data objects. A different formulation [16] of quality of variants under selection relies on cluster validity indices specifying compactness-remoteness characteristics of clusters in the feature space (see, e.g., [17] for an overview of existing indices). The weights can be attributed with a collection of different cluster validity indices [18].

Another way of assigning weights revolves around the notion of *refined co-association matrix* with elements in the continuous interval. Each lth variant of clustering determines its own refined co-association matrix; the ensemble partition is found using the equally averaged matrix. In probability accumulation method [19], a weight of the pair of objects in the same cluster depends on its size. The authors of [20] assign weights according to the distances from each object in the pair to cluster centroids. A weight characterizing similarity between two objects can be evaluated taking into account their neighboring objects [21] or a path connected them [22].

It may be noticed that a procedure of assigning weights makes use of two levels of information: local (data point) level and global (cluster) level. At the local level, the properties of clustering ensemble in relation to an object pair are taken into account. At the global level, general characteristics of clusters in the partition (we call them *evaluation functions*) make a contribution to the weight.

In the current work, we follow general framework of ensemble clustering based on weighted co-association matrices. There are two schemes one could use to design base elements of the ensemble. First of all, a single algorithm may create data partitions by varying its parameters ("homogenous ensemble"). Another scheme involves a number of completely different clustering algorithms, each one operating on its own domain of parameters ("heterogeneous ensemble"). In our earlier work [23], a probabilistic model of clustering ensemble was suggested under the first scheme. The model was modified in [24] for the analysis of ensembles organized in accordance with the second scheme. It was shown that the largest weights should be attributed to most "stable" algorithms. However, both works consider only variations of cluster labels and disregard potentially useful information on other characteristics of base partitions (cluster validity indices, diversity measures, etc).

The aim of this paper is to theoretically investigate the model of clustering ensemble based on co-association matrices with elements dependent on local and global levels of information and weights assigned proportionally to arbitrary evaluation function. We consider a single algorithm that constructs base partitions using parameters taken at random. In particular, our interest is in studying the convergence properties of the solutions under increasing ensemble size. Is growing ensemble size really causes the improvement of clustering quality and under which conditions? What is the benefit of using weighted voting against simple averaging with equal weights? Another problem is how the characteristics (i.e., evaluation functions) of partial clusterings influence the quality of the overall decision.

The rest of the paper is organized as follows. In the second section we describe the scheme of clustering ensemble based on weighted co-association matrices. The third section describes the model of clustering ensemble. Using the concept of ensemble margin, we find an upper bound for misclassification error in attributing pairs of objects to clusters. Applying the obtained bound, we formulate some theoretical properties of the ensemble related to its convergence. In the forth section we consider the results of numerical experiments with the model aimed at its practical confirmation. Using validity indices as evaluation functions, we present the results of experiments using Monte-Carlo simulations and clustering of a real hyperspectral image. The conclusion summarizes the work.

2. Clustering ensemble

Let us consider the following algorithmic construction. We shall suppose that a clustering algorithm μ is enabled to work a number of times with different parameter settings (in general, under different working conditions such as initial centroids coordinates, subsets of features and number of clusters). In each lth run, it generates a clustering variant composed of K_l groups, $l \in \{1, \ldots, L\}$, where L is total number of runs. Each variant is appraised using some evaluation function γ; thus a collection of values $\gamma_1, \ldots, \gamma_L$ is obtained. It is allowable to suppose that the values are appropriately standardized so that (a) $0 < \gamma_l \leq 1$, $l \in \{1, \ldots, L\}$, and (b) the better are the found variants according to some criterion, the larger are the function values.

Because the numberings of clusters do not matter, it is convenient to define an equivalence relation, i.e. to determine whether the algorithm μ assigns each pair of objects to the same cluster or to the separate clusters. For a pair of different objects a_i and a_j, we define the value $h(i, j)$,