Pattern Recognition Letters 38 (2014) 99–106

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Weighted ensemble of algorithms for complex data clustering ☆

Vladimir Berikov *

Sobolev Institute of mathematics, 4 Koptyug pr., Novosibirsk 630090, Russia
Novosibirsk State University, 2 Universitetsky pr., Novosibirsk 630090, Russia

ARTICLE INFO

Article history:
Received 26 August 2012
Available online 3 December 2013

Keywords:
Clustering
Classification
Weighted clustering ensemble
Latent variable model
Classification error bound

ABSTRACT

This paper considers a problem of clustering complex data composed from various structures. A collection of different algorithms is used for the analysis. The main idea is based on the assumption that each algorithm is "specialized" (as a rule, gives more accurate partition results) on particular types of structures. The degree of algorithm's "competence" is determined by usage of weights attributed to each pair of observations. Optimal weights are specified by the analysis of partial ensemble solutions with use of the proposed model of clustering ensemble. The efficiency of the suggested approach is demonstrated with Monte-Carlo modeling.

1. Introduction

The purpose of cluster analysis is to form a small number of distinct groups which include similar objects. Grouping results can be either "hard" or "fuzzy" ("fuzzy" clustering is not considered in this paper).

To date, there exist a large number of various clustering algorithms (Jain, 2010). These algorithms are based on different interpretations of notions "distance" and "similarity", use specific procedures of finding optimal variant of grouping and utilize various additional information on application area.

In last decades, an approach based on a collective decision is actively used in cluster analysis (see overview of recent work in Ghosh and Acharya (2011) and Vega-Pons and Ruiz-Shulcloper (2011)). The main advantages of ensemble clustering are as follows:

- under proper conditions, this approach allows increasing the stability of clustering process (reduces the dependence from algorithm parameters) and improve clustering quality;
- it enables different algorithms to collaborate when searching the consensus partition, considering a problem from "multiple views";
- it makes possible to effectively solve clustering tasks with mixed numerical and categorical features, missed values and presence of noise.

Theoretical studies of clustering ensemble methods were performed in a number of works (see, for example, Topchy et al. (2004), Wang et al. (2011) and Hadjitodorov et al. (2006)). It is worth mentioning that the problem of theoretical substantiation of methods is one of the most important in cluster analysis.

The following methodologies are frequently used to obtain ensemble clustering solutions (Ghosh and Acharya, 2011; Vega-Pons and Ruiz-Shulcloper, 2011):

- maximization of likelihood function in the distribution mixture model framework;
- finding maximum degree of consensus between clustering partitions (with use of such characteristics as Normalized Mutual Information, Adjusted Rand index etc.);
- calculation of ensemble pairwise similarity (dissimilarity) matrix (this approach is considering in the presented work);
- usage of graph-theoretic methods;
- analysis of bootstrap samples.

One of the arguments for the ensemble approach is the absence of a universal clustering algorithm: each method has a specific area of its implementation. Some algorithms give more accurate results on data described by spherical patterns in multidimensional feature space; other methods are intended for searching strip-like clusters or groups of other complicated form (under algorithm's accuracy one can understand the degree of matching between found clusters and the true ones, following data generation mechanism; due to unsupervised nature of clustering, algorithm's accuracy can be estimated only in test mode).

When data has complex nature (see an example in Fig. 1), the reasonable way is to apply not a single algorithm, but an organized collection of highly specialized algorithms.

☆ This paper has been recommended for acceptance by M.A. Girolami.

* Address: Sobolev Institute of mathematics, 4 Koptyug pr., Novosibirsk 630090, Russia. Tel.: +7 3833634681; fax: +7 3833332598.
E-mail address: berikov@math.nsc.ru

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.11.012

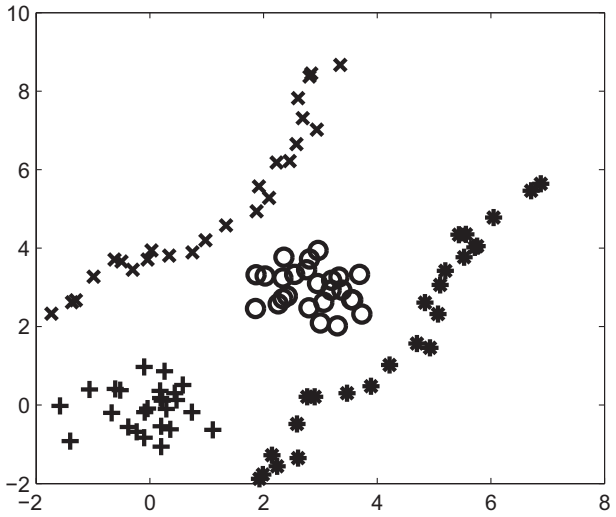

Fig. 1. An example of data structure: two distorted spherical clusters and two strip-like clusters in two-dimensional feature space.

A serious problem lies in possible ambiguous interpretation of obtained clustering solutions. Methods based on different approaches can produce incompatible variants of grouping. In this work, it is supposed that base algorithms of the ensemble are not mutually discordant; figuratively speaking, they "supplement" each other; each one compensates "weak points" of other algorithms. So it is required to determine an objective measure of algorithm's contribution into the collective decision (the "competence" of algorithm).

In the approach suggested in the presented paper, a vector of algorithms' weights is assigned to every pair of objects. This methodology allows one to determine algorithm's competence in dependence from types of structures to which the pair belongs.

The model of clustering ensemble previously introduced in Berikov (2011) is used for evaluating the competence of algorithms. The underlying idea follows L. Breiman's random forest approach (Breiman, 2001). The model is based on a latent variable framework that allows finding an upper bound for probability of error in ascribing a pair of objects whether to the same cluster or to different clusters. With use of the model, the observed characteristics of the ensemble are associated with directly unobserved classification error. In Berikov (2011), it was theoretically concluded that the probability of error decreases with an increase in ensemble's homogeneity (i.e., with reducing variance and increasing correlation between ensemble solutions). In the presented paper, this model is modified by introducing algorithms' weights. Optimal weights are estimated for every pair of objects to obtain the lowest error bound.

The rest of the paper is organized as follows. Section 2 briefly discusses the literatures about weighted clustering ensembles. Section 3 gives basic definitions and introduces a model of pairwise weighted ensemble clustering. The main theoretical result that allows determining optimal weights is considered in Section 4. The proposed PWEC algorithm is presented in Section 5. Numerical experiments with Monte Carlo modeling are described in Section 6. The last section summaries the work, discusses current limitations and outlines future directions for research.

2. Weighted clustering ensembles

Weighted clustering ensembles is a quite recent research topic in data mining area. From the point of view of consensus partitioning approach, the problem is specified as follows:

$$\text{find} \quad \mathbb{P}^* = arg\max_{\mathbb{P}\in\mathcal{P}} \sum_{l=1}^{L} w_l \delta(\mathbb{P}, \mathbb{P}_l),$$

where $\mathcal{P}$ is the set of all partitions of data sample $s = \{o_1, \ldots, o_N\}$, comprised of N objects, $\{\mathbb{P}_1, \ldots, \mathbb{P}_L\}$ are given (base) variants of partitioning, δ is similarity measure between two partitions, $w_l \geqslant 0$ is a weight of lth partition, $l = 1, \ldots, L$; $\sum_l w_l = 1$.

The earlier works on ensemble clustering assumed equal contribution of each variant to the consensus partition ($w_l \equiv 1/L$). This approach was theoretically validated in Topchy et al. (2004). Using central limit theorem, it was proved that the probability of finding true consensus partition increases with an increase in ensemble size, provided that: (a) each base algorithm possesses better quality than a trivial algorithm of random partitioning; (b) the algorithms run independently.

In real-life clustering tasks, ensemble size is always limited, and theoretical assumptions can be violated. However, it would be desirable to obtain the highest possible clustering quality in these situation. With that end in view, a number of methods aimed to speed up the convergence to optimal clustering partition were suggested by different authors. These methods are based on the assessment of significance values to ensemble solutions: the variant with higher importance obtains greater weight in the resultant partition. The assignment of weights is implemented in different ways.

A reasonable way is to set the weights of partitions proportionally to the obtained values of specified cluster validity index. In this method, the partition variants are generated with bootstrap subsamples of the given sample (Frossyniotis et al., 2004).

Another method determines a criterion of ensemble quality by introducing a set of weights associated with feature projections of clusters (Al-razgan and Domeniconi, 2009). The weights are defined inversely proportionally to the scatter of observations along the coordinate axes. This method aims to reduce a negative impact of low information (noise) features on the consensus solution.

The weights of partitioning variants can be specified proportionally to the impact of each partition on the overall measure of ensemble diversity (Gullo et al., 2009). The authors present a number of alternative definitions of this measure, using different strategies for determining the distance between clustering partitions.

In the method called Partition Relevance Analysis (Vega-Pons et al., 2008), the quality of each base algorithms is evaluated with a number of different cluster validity indices. For each index, the entropy measure of its variation on the set of obtained clustering partitions is determined. High weights are assigned to ensemble variants with large degree of agreement, proportionally to the sum of the entropy measures. Variants with large scattering of validity indices are considered noise and obtain small weights.

One of the difficulties in this approach is a labeling correspondence problem: as numberings of clusters do not matter, any permutation of labels is possible. Finding optimal consensus partition is rather time-consuming problem, so approximate algorithms are applied as a rule.

Another direction in clustering ensembles literature employs a notion of co-association (CA) matrix S, whose elements are zero-one values representing object pairs; each value indicates whether the pair belongs to the same cluster (1) or to different clusters (0). The averaged CA matrix is defined as

$$\mathbf{S} = \sum_l w_l S^{(l)},$$

where $S^{(l)}$ is CA matrix for lth variant of partition; $w_l \geqslant 0$ is a weight of lth variant, $l = 1, \ldots, L$; $\sum_l w_l = 1$.

Following this direction, the averaged matrix $\mathbf{S}$ is treated as pairwise similarity matrix and used as input data to produce the final clustering partition.