Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

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

Ensemble clustering using factor graph

Dong Huang [a,d], Jianhuang Lai [a,*], Chang-Dong Wang [b,c]

[a] School of Information Science and Technology, Sun Yat-sen University, Guangzhou Higher Education Mega Center, Panyu District, Guangzhou, Guangdong 510006, China
[b] School of Mobile Information Engineering, Sun Yat-sen University, China
[c] SYSU-CMU Shunde International Joint Research Institute (JRI), China
[d] College of Mathematics and Informatics, South China Agricultural University, China

ARTICLE INFO

Article history:
Received 24 January 2015
Received in revised form
6 July 2015
Accepted 17 August 2015

Keywords:
Ensemble clustering
Factor graph
Belief propagation
Super-object
Automatic cluster number estimate

ABSTRACT

In this paper, we propose a new ensemble clustering approach termed ensemble clustering using factor graph (ECFG). Compared to the existing approaches, our approach has three main advantages: (1) the cluster number is obtained automatically and need not to be specified in advance; (2) the reliability of each base clustering can be estimated in an unsupervised manner and exploited in the consensus process; (3) our approach is efficient for processing ensembles with large data sizes and large ensemble sizes. In this paper, we introduce the concept of super-object, which serves as a compact and adaptive representation for the ensemble data and significantly facilitates the computation. Through the probabilistic formulation, we cast the ensemble clustering problem into a binary linear programming (BLP) problem. The BLP problem is NP-hard. To solve this optimization problem, we propose an efficient solver based on factor graph. The constrained objective function is represented as a factor graph and the max-product belief propagation is utilized to generate the solution insensitive to initialization and converged to the neighborhood maximum. Extensive experiments are conducted on multiple real-world datasets, which demonstrate the effectiveness and efficiency of our approach against the state-of-the-art approaches.

1. Introduction

The ensemble clustering technique has been receiving increasing attention in recent years due to its ability to combine multiple clusterings into a probably better and more robust clustering [1–10]. Many ensemble clustering approaches have been developed in the past few decades [11]. Despite the great success, there are still three limitations to most of the existing ensemble clustering methods. First, they generally take the cluster number of the final clustering as input and lack the ability to automatically estimate the cluster number. Second, most of them treat each base clustering equally and overlook the different reliability of the base clusterings. Third, many of the existing methods work at the object-level and does not scale well for large ensembles. In this paper, we refer to the ensemble clustering approaches with the ability to automatically estimate the cluster number of the final clustering as the *automatic* approaches, and the approaches without the ability of automatic cluster number estimate as the *non-automatic* approaches. Recently some efforts have been made to address one or two of the three limitations [12–14]. However, to the best of our knowledge, none of the existing ensemble clustering approaches is capable of dealing with all of the three limitations in a unified model.

To overcome the aforementioned three limitations, in this paper, we propose a novel ensemble clustering approach termed ensemble clustering using factor graph (ECFG). We introduce the concept of super-object as a compact and adaptive representation for ensemble data. Instead of using the original data objects, we use the super-objects as the primitive objects, which greatly reduces the problem size and facilitates the overall process. We formulate the ensemble clustering problem into a probabilistic framework. The clustering results are represented by a set of binary decisions. Each binary decision indicates whether the corresponding two super-objects are in the same cluster or not. We assume that each base clustering is associated with a probability of making the *correct* decisions, which can be viewed as the reliability of a base clustering. The consensus clustering and the reliability of each base clustering are estimated iteratively by solving an instance of binary linear programming (BLP) problem. However, the BLP problem is NP-hard. The high computational complexity is the most significant hurdle for it. In this work, we present an efficient solver for this instance of BLP problem based on the factor graph technique [15]. The factor graph is a powerful tool for solving optimization problems and has many successful applications

* Corresponding author. Tel.: +86 13168313819; fax: +86 20 84110175.
E-mail addresses: huangdonghere@gmail.com (D. Huang), stsljh@mail.sysu.edu.cn (J. Lai), changdongwang@hotmail.com (C.-D. Wang).

http://dx.doi.org/10.1016/j.patcog.2015.08.015
0031-3203/

in the field of pattern recognition and machine learning [16–19]. In our work, the constrained objective function is represented by a factor graph. Then the max-product belief propagation [15] is applied, which generates the solution insensitive to initialization and converged to the neighborhood maximum.

The main contributions of this paper are summarized as follows:

1. We introduce the concept of super-object, which serves as a compact and adaptive representation for ensemble data and significantly facilitates the computation of the consensus process.
2. We cast the ensemble clustering problem into a BLP problem and propose an efficient solver for the BLP problem based on factor graph.
3. We propose a novel ensemble clustering approach termed ECFG, which has three advantages: (i) it can automatically estimate the cluster number of the final clustering, (ii) the reliability of each base clustering can be estimated and exploited. (iii) it is efficient w.r.t. both large data sizes and large ensemble sizes.
4. Experimental results on multiple real-world datasets have shown that our approach significantly outperforms the state-of-the-art approaches in terms of both clustering accuracy and efficiency (see Section 4).

The remainder of this paper is organized as follows. We review the related work in Section 2. The proposed ensemble clustering approach termed ECFG is introduced in Section 3. The experimental results are reported in Section 4. We conclude this paper in Section 5.

2. Related work

The purpose of ensemble clustering is to combine multiple base clusterings into a more accurate and robust clustering. With regard to the difference of the input information of the ensemble clustering system, there are two formulations of the ensemble clustering problem. In the first formulation, the ensemble clustering system uses only the information of the multiple clusterings as input and has no access to the original data features [1,5,7,20,21]. In the other formulation, the ensemble system uses both the multiple clusterings and the original features as inputs [2,4]. In this paper, we comply with the former formulation of the ensemble clustering problem with no access to the original data features, which is also the common practice for most ensemble clustering approaches [1,5,7,20,21].

In the past decade, there is a large amount of literature on ensemble clustering [11]. The pair-wise co-occurrence based approaches [1,21,22] construct a co-association matrix by considering how many times two objects occur in the same cluster among the multiple base clusterings and then use the co-association matrix to measure the similarity between objects. Fred and Jain [1] for the first time introduced the concept of co-association matrix and proposed the evidence accumulation clustering (EAC) method. Wang et al. [21] extended the EAC method by taking the sizes of clusters into consideration and presented the probability accumulation method. Yi et al. [5] proposed to deal with the uncertain pairs in the co-association matrix by exploiting the matrix completion technique.

Another main category of ensemble clustering is based on graph partitioning [20,23]. The graph partitioning based approaches typically construct a graph by using the objects and/or clusters as graph nodes and building graph links with regard to the information of the multiple base clusterings [20,23]. Then the consensus clustering is obtained by partitioning the graph into a certain number of segments, each treated as a cluster of the final clustering [20,23]. Strehl and Ghosh [23] formulated the ensemble clustering problem into a graph partitioning problem and introduced three classic algorithms, namely, the cluster-based similarity partitioning algorithm (CSPA), the hypergraph-partitioning algorithm (HGPA), and the meta-clustering algorithm (MCLA). The data objects are treated as graph nodes in CSPA and HGPA, while the clusters are treated as graph nodes in MCLA. Different from the three algorithms in [23], Fern and Brodley [20] constructed a bipartite graph by treating both objects and clusters as nodes and obtained the final clustering by partitioning the bipartite graph into a number of disjoint sets of nodes.

The median partition based approaches are also one of the main categories of ensemble clustering [8,22,24]. In the median partition based methods, the ensemble clustering problem is cast into an optimization problem that aims to maximize the similarity between the consensus clustering (or median clustering) and the base clusterings. The median partition problem is NP-hard [25]. Cristofor and Simovici [22] utilized a genetic algorithm to find an approximate solution to the median partition problem. Singh et al. [24] used an agreement measure based on 2D string encoding and maximized the new agreement measure using 0-1 semidefinite programming (SDP) so as to find the consensus clustering. Franek and Jiang [8] cast the median partition problem into the Euclidean median problem by clustering embedding in vector spaces. The median vector is computed by the Weiszfeld algorithm [26], and then converted back into a clustering, which is taken as the consensus clustering, by an inverse transformation [8].

Although many successful ensemble clustering approaches have been developed, there are still three limitations to most of the existing approaches. First, they generally take the cluster number of the final clustering as input and cannot automatically estimate the cluster number. Second, many of them work at the object-level and are not capable of dealing with the clustering ensembles with very large data sizes and ensemble sizes. Third, they mostly treat each base clustering equally and lack the ability to weight the base clusterings with regard to their reliability. Recently some efforts have been made to address these limitations [9,12,14,27,28]. Mimaroglu and Erdil [27] constructed a similarity graph based on the evidence accumulated from multiple base clusterings and proposed a pivot-based algorithm termed COMUSA to obtain the consensus clustering with the cluster number automatically found. Based on COMUSA, Mimaroglu and Erdil [14] further proposed two scalable ensemble clustering algorithms, termed COMUSACL and COMUSACL-DEW, respectively. Li and Ding [12] introduced a weighted ensemble clustering method based on non-negative matrix factorization. Yu et al. [9] proposed a weighting and selecting scheme for ensemble clustering based on the feature selection technique. Alush and Goldberger [28] addressed the problem of ensemble segmentation, which can be viewed as a special instance of ensemble clustering, and obtained the final segmentation with regard to the reliability of each base segmentation. The method in [28] is not applicable to the general ensemble clustering problem due to two major hurdles: (i) its efficiency heavily relies on the super-pixel map, which is constructed by exploiting the spatial constraints of image data and cannot be used for the general ensemble clustering problem; (ii) the computational complexity of integer linear programming in [28] is prohibitively high for large datasets and is not feasible even for a graph with over 200 nodes. On the whole, each of these methods [9,12,14,27,28] addressed one or two of the aforementioned three limitations. However, none of them is able to tackle all of the three limitations simultaneously. How to overcome all of the three limitations in a unified ensemble clustering framework remains an unaddressed and very challenging problem.

3. Ensemble clustering using factor graph

In this section, we introduce the proposed approach termed ensemble clustering using factor graph (ECFG). The formulation of the ensemble clustering problem is presented in Section 3.1. We

Please cite this article as: D. Huang, et al., Ensemble clustering using factor graph, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.08.015