Neurocomputing 170 (2015) 240–250

Contents lists available at ScienceDirect

Neurocomputing

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

Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis ☆

Dong Huang [a,d], Jian-Huang 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, PR China
[b] School of Mobile Information Engineering, Sun Yat-sen University, PR China
[c] SYSU-CMU Shunde International Joint Research Institute (JRI), PR China
[d] Guangdong Key Laboratory of Information Security Technology, PR Chinaa

ARTICLE INFO

Article history:
Received 15 December 2013
Received in revised form
15 March 2014
Accepted 4 May 2014
Available online 6 May 2015

Keywords:
Weighted clustering ensemble
Weighted consensus clustering
Weighted evidence accumulation clustering
Graph partitioning with multi-granularity link analysis

ABSTRACT

The clustering ensemble technique aims to combine multiple clusterings into a probably better and more robust clustering and has been receiving an increasing attention in recent years. There are mainly two aspects of limitations in the existing clustering ensemble approaches. Firstly, many approaches lack the ability to weight the base clusterings without access to the original data and can be affected significantly by the low-quality, or even ill clusterings. Secondly, they generally focus on the instance level or cluster level in the ensemble system and fail to integrate multi-granularity cues into a unified model. To address these two limitations, this paper proposes to solve the clustering ensemble problem via crowd agreement estimation and multi-granularity link analysis. We present the normalized crowd agreement index (NCAI) to evaluate the quality of base clusterings in an unsupervised manner and thus weight the base clusterings in accordance with their clustering validity. To explore the relationship between clusters, the source aware connected triple (SACT) similarity is introduced with regard to their common neighbors and the source reliability. Based on NCAI and multi-granularity information collected among base clusterings, clusters, and data instances, we further propose two novel consensus functions, termed weighted evidence accumulation clustering (WEAC) and graph partitioning with multi-granularity link analysis (GP-MGLA) respectively. The experiments are conducted on eight real-world datasets. The experimental results demonstrate the effectiveness and robustness of the proposed methods.

1. Introduction

Data clustering is a fundamental and very challenging problem in data mining and machine learning. The purpose is to partition unlabeled data into homogeneous groups, each referred to as a cluster. Data clustering requires a distance metric for evaluating the similarity between data instances, which, without prior knowledge of cluster shapes, is hard to specify. In the past few decades, a large number of clustering algorithms have been developed [1–9]. However, there is no single clustering method which is able to identify all sorts of cluster shapes and structures in data.

For the same dataset, different methods, or even the same method with different initializations or parameter settings, may lead to very different clustering results. It is extremely difficult to decide which method would be the *proper* one for a given clustering task, not to say how to properly specify the initialization and parameter setting for the chosen method. Each method has its own merits as well as weaknesses. Different clusterings generated by different methods or with varying parameters can provide multiple views of the data. How to combine the information of different clustering results for obtaining a better and more robust clustering remains a very challenging problem [10,11].

In recent years, many clustering ensemble approaches have been developed, which aim to combine multiple clusterings into a probably better and more robust clustering by utilizing various techniques [12–24]. However, in most of the existing methods, there are mainly two aspects of limitations. Firstly, many of the clustering ensemble approaches lack the ability to weight the base clusterings without access to the original data features, which makes them vulnerable to low-quality clusterings and probable to be affected significantly by low-quality clusterings (or even ill clusterings). Secondly, they mainly focus on the instance level or the cluster level in the ensemble system and fail to fuse multi-granularity information into a unified model. In order to address

☆ Requests for the source code should be sent to the first author via email.

* Corresponding author. Tel.: +86 13168313819; fax: +86 2084110175.
E-mail addresses: huangdonghere@gmail.com (D. Huang), stsljh@mail.sysu.edu.cn (J.-H. Lai), changdongwang@hotmail.com (C.-D. Wang).

http://dx.doi.org/10.1016/j.neucom.2014.05.094
0925-2312/

these two limitations, in this paper, we propose a clustering ensemble framework based on crowd agreement estimation and multi-granularity link analysis. By exploring the relationship among the base clusterings, we present a novel clustering validity measure termed normalized crowd agreement index (NCAI), which is able to evaluate the quality of base clusterings in an unsupervised manner and provides information for treating each base clustering accordingly. The source aware connected triple (SACT) similarity is introduced for analyzing the similarity between clusters with regard to their common neighbors and source reliability. Besides the relations between base clusterings and between clusters, we further investigate the linkage between data instances and clusters and incorporate the information from the three levels of granularity in a unified framework. In our previous work [25], we introduced the consensus function termed graph partitioning with multi-granularity link analysis (GP-MGLA). This paper is a major extension of our previous work on clustering ensemble. In this paper, more comprehensive literature and motivation are provided. Besides that, we propose another novel consensus function termed weak evidence accumulation clustering (WEAC), which is developed from the conventional evidence accumulation clustering (EAC) [14] and capable of dealing with ill clusterings by incorporating the clustering validity cue into the ensemble process. Extensive experiments are further conducted on real-world datasets for evaluating the proposed methods against several baseline clustering ensemble methods.

The remainder of this paper is organized as follows. In Section 2, we review the related work of the clustering ensemble technique. In Section 3, we describe the formulation of the clustering ensemble problem. In Section 4, we present the crowd agreement estimation mechanism. The source aware connected triple (SACT) similarity is introduced in Section 5. In Section 6, we propose two novel consensus functions termed weighted evidence accumulation clustering (WEAC) and graph partitioning with multi-granularity link analysis (GP-MGLA) respectively. The experimental results are reported in Section 7. We conclude this paper in Section 8.

2. Related work

Clustering ensemble is also known as clustering combination or clustering aggregation, which aims to combine multiple clusterings, each referred to as a base clustering (or an ensemble member), to obtain a so-called consensus clustering. As illustrated in Fig. 1, the clustering ensemble process involves two steps: the first step is to generate multiple clusterings for a given dataset; and the second step is to construct the consensus clustering from the ensemble of base clusterings using different consensus functions.

Given a dataset, the ensemble of base clusterings can be generated by running different clustering algorithms [21,23,25], running the same algorithm with different initializations and parameters [14,18,20,22], clustering via sub-sampling the data repeatedly [12,13], or clustering via projecting the data onto different subspaces [12,13,15,19]. Compared to generating base clusterings, how to combine multiple base clusterings, i.e., how to design the consensus function, is much more important and challenging in the clustering ensemble problem.

In the past few years, many consensus functions have been developed to fuse information from multiple clusterings [12–24]. These approaches can be classified into mainly three categories, namely, (i) the median partition based methods [26,15,24], (ii) the pair-wise co-occurrence based methods [14,17,20], and (iii) the graph partitioning based methods [12,13,19].

In the median partition based approaches [26,15,24], the clustering ensemble problem is formulated into an optimization problem, aiming to find the partition/clustering that maximizes the similarity between the partition and the base clusterings, over the space of all partitions. The median partition problem is NP-complete [15]. Instead of finding the optimal solution over the huge space of all possible partitions, Cristofor and Simovici [26] used the genetic algorithm to obtain an approximative solution where the clusterings are represented by chromosomes. Topchy et al. [15] cast the median partition problem into a maximum likelihood problem, as a solution to which the consensus clustering is found using the EM algorithm. Franek and Jiang [24] reduced the median partition problem to the Euclidean median problem by clustering embedding in vector spaces and found the median vector by the Weiszfeld algorithm [27]. Then an inverse transformation would be performed to convert the median vector into a clustering, which was taken as the consensus clustering.

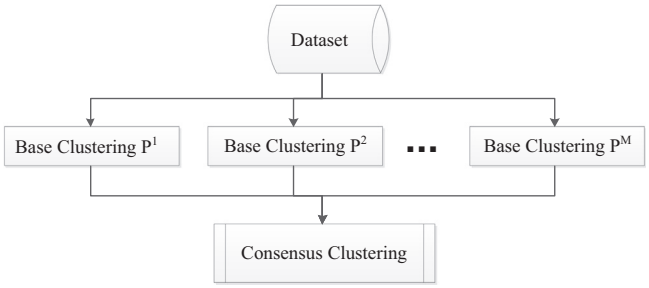

Fig. 1. The clustering ensemble process.

The pair-wise co-occurrence based approaches [14,17,20] construct the similarity between data instances by considering how many times they occur in the same cluster in the ensemble of base clusterings. Fred and Jain [14] introduced the evidence accumulation clustering (EAC) method, which used the co-association matrix to measure the similarity between instances. Then the hierarchical agglomerative clustering algorithms [10], e.g., single-link (SL) and average-link (AL), can be performed on the co-association matrix and thus the consensus clustering is obtained. Li et al. [17] analyzed the co-association matrix and proposed a novel hierarchical clustering algorithm by utilizing the concept of normalized edges to measure the similarity between clusters. Wang et al. [20] generalized the EAC method and proposed the probability accumulation method, which took into consideration the sizes of clusters in the ensemble.

Another category of clustering ensemble is based on graph partitioning [12,13,19]. Strehl and Ghosh [12] modeled the ensemble of clusterings in a hypergraph structure where the clusters are treated as hyperedges. For partitioning the graph and obtaining the consensus clustering, they further proposed three graph partitioning algorithms, namely, the cluster-based similarity partitioning algorithm (CSPA), the hypergraph-partitioning algorithm (HGPA), and the meta-clustering algorithm (MCLA). Fern and Brodley [13] formulated the clustering ensemble into a bipartite graph where both the data instances and clusters are represented as graph nodes. An edge between two nodes exists if and only if one of the nodes is a data instance and the other node is the cluster containing it. The consensus clustering is obtained by partitioning the graph into a certain number of disjoint sets of graph nodes.

Many of the existing clustering ensemble approaches implicitly assume that all the base clusterings contribute equally to the ensemble system and can be affected significantly by low-quality clusterings or even ill clusterings. In recent years, some efforts have been made to weight the base clusterings with regard to the clustering validity. Vega-Pons et al. [28] exploited several property validity indexes (PVIs), namely, variance (VI), connectivity (CI), silhouette width (SI) and Dunn index (DI), to assign a weight to