Computers and Electrical Engineering 68 (2018) 603–615

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

A comparative study of clustering ensemble algorithms☆

Xiuge Wu[a], Tinghuai Ma[*,a,b], Jie Cao[c], Yuan Tian[d], Alia Alabdulkarim[d]

[a] School of Computer & Software, Nanjing University of information science & Technology, Jiangsu, Nanjing 210044, China
[b] CICAEET, Jiangsu Engineering Center of Network Monitoring, Nanjing University of information science & Technology, Nanjing 210044, China
[c] School of Economics & Management, Nanjing University of Information Science & Technology, Nanjing 210044, China
[d] Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh 11362, Saudi Arabia

ARTICLE INFO

Keywords:
Clustering ensemble
Generative mechanism
Consensus function
Ensemble member
Diversity
Ensemble size

ABSTRACT

Since clustering ensemble was proposed, it has rapidly attracted much attention. This paper makes an overview of recent research on clustering ensemble about generative mechanism, selective clustering ensemble, consensus function and application. Twelve clustering ensemble algorithms are described and compared to choose a basic one. The experiment shows that using k-means with different initializations as generative mechanism and average-linkage agglomerative clustering as consensus function is the best one. As ensemble size increases, the performance of clustering ensemble improves. The basic clustering ensemble algorithm with suitable ensemble size is compared with clustering algorithms and the experiment shows that clustering ensemble is better than clustering. The influence of diversity on clustering ensemble is instructive to selecting members. The experiment shows that selecting members in high quality and big diversity for low-dimensional data sets, and selecting members in high quality and median diversity for high-dimensional data sets are better than traditional clustering ensemble.

1. Introduction

With rapid progress of clustering technology, clustering analysis plays an important role in various fields, such as pattern recognition, image processing, business intelligence, document clustering, market research, data analysis and customer recommendation. It is difficult to find one clustering algorithm that can be applied to all data sets, so various clustering algorithms are improved and different clustering algorithms are proposed. For this problem, authors in [1] proposed the concept of clustering ensemble in 2003. Specifically, the definition of clustering ensemble is as follows: there is a dataset $X = \{x_1, x_2, ..., x_n\}$ that has n data, Then M clustering algorithms are used to cluster X and generate M partitions. The ensemble member set $P = \{P_1, P_2, ..., P_M\}$ is formed with these partitions and $P_m(m = 1, 2, ..., M)$ is the clustering partition obtained by the mth clustering algorithm. Subsequently, consensus function Γ will combine these ensemble members and get the final partition P^*. The intuitive illustration of clustering ensemble is shown in Fig. 1.

Clustering ensemble combines different clustering partitions about dataset into a final one. The result of clustering ensemble is superior to single clustering algorithm. Single clustering algorithm has its own weakness, so it leads to one algorithm being only suitable for a specific dataset. Clustering ensemble combines these clustering algorithms to avoid the shortcoming of single clustering algorithm. It fits more datasets than clustering and it is also robust against noise and outliers [2].

In order to make a comprehensive research on clustering ensemble and choose a basic algorithm for our further researches, we

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. J. D. Peter.

* Corresponding author.
E-mail address: thma@nuist.edu.cn (T. Ma).

https://doi.org/10.1016/j.compeleceng.2018.05.005
Received 11 August 2017; Received in revised form 3 May 2018; Accepted 4 May 2018

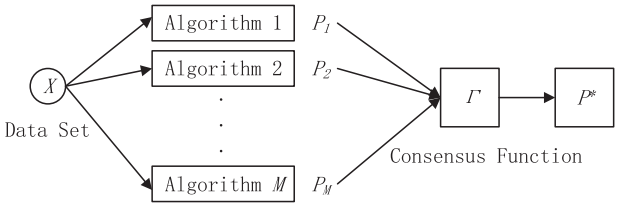

Fig. 1. Framework of clustering ensemble.

introduce twelve clustering ensemble algorithms. The twelve algorithms are composed of three generative mechanisms and four consensus functions. Then, the comparative experiment of these algorithms finds the best one as the basic algorithm for further researches. The basic algorithm generates ensemble members using k-means with different initializations and combines members using average-linkage agglomerative clustering. Next, the influence of ensemble size on clustering ensemble is analysed to find an appropriate ensemble size. Then the basic algorithm with suitable ensemble size is compared with standard clustering algorithms. In addition, the relation of diversity and performance of clustering ensemble is explored to guide the selection of ensemble members. Finally, the selective clustering ensemble based on quality and diversity is compared with traditional clustering ensemble.

The rest of this paper is organized as follows. Section 2 reviews the recent research on clustering ensemble. Three generative mechanisms and four consensus functions are described in Section 3. Section 4 compares twelve clustering ensemble algorithms on six datasets, analyzes the influence of ensemble size and diversity, and compares clustering ensemble with standard clustering algorithms and selective clustering ensemble. This paper is concluded in Section 5 with discussion about future works of clustering ensemble.

2. Literature review on clustering ensemble

There are two main phases in clustering ensemble. The first stage is producing ensemble members while the second stage is combining these ensemble members to get the final partition. As indicated in Fig. 2, the left side shows different generative mechanisms and the right side displays different consensus functions. By selecting different clustering algorithms, setting different initializations for same clustering algorithm, using sampling data or using feature subsets, we can produce different ensemble members. Whereas the consensus functions include voting approach, hierarchical clustering, graph method, information theory and mixture model.

Accordingly, the recent research mainly focus on four aspects. (1) Generative mechanism: the approach to get the ensemble members [3–12]. (2) Selective clustering ensemble: selecting effective ensemble members before consensus function [13–18]. (3) Consensus function: the method of combining ensemble members [19–23]. (4) Application: the practical applications of clustering ensemble [24–27].

2.1. Generative mechanism

Adopting different clustering algorithms to generate ensemble members is one of the common generative mechanisms. Authors in [1] propose to apply different clustering algorithms on the same dataset. In [3], authors use self-organizing maps and k-means, the two well-known clustering algorithms in neural network and statistical field, to generate ensemble members.

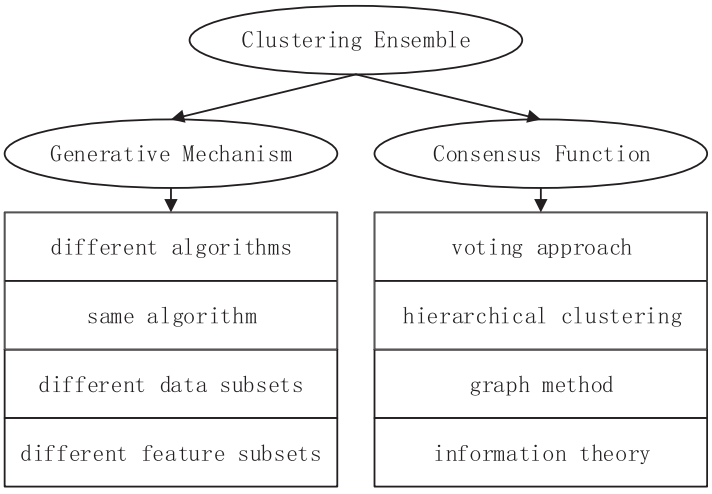

Fig. 2. Clustering ensemble algorithms classification from generative mechanism and consensus function.