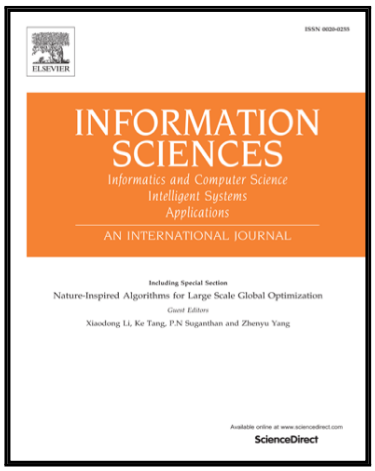

Multigranulation information fusion: a Dempster-Shafer evidence theory-based clustering ensemble method

Feijiang Li, Yuhua Qian, Jieting Wang, Jiye Liang

PII: S0020-0255(16)31156-2
DOI: 10.1016/j.ins.2016.10.008
Reference: INS 12567

To appear in: *Information Sciences*

Received date: 29 August 2015
Revised date: 21 September 2016
Accepted date: 1 October 2016

Please cite this article as: Feijiang Li, Yuhua Qian, Jieting Wang, Jiye Liang, Multigranulation information fusion: a Dempster-Shafer evidence theory-based clustering ensemble method, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.10.008

Multigranulation information fusion: a Dempster-Shafer evidence theory-based clustering ensemble method☆

Feijiang Li[a], Yuhua Qian[a,b,*], Jieting Wang[a], Jiye Liang[b]

[a]School of Computer and Information Technology, Shanxi University, Taiyuan, 030006 Shanxi, China
[b]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan, 030006 Shanxi, China

Abstract

Clustering analysis is a fundamental technique in machine learning, which is also widely used in information granulation. Multiple clustering systems granulate a data set into multiple granular structures. Therefore, clustering ensemble can serve as an important branch of multigranulation information fusion. Many approaches have been proposed to solve the clustering ensemble problem. This paper focuses on the direct approaches which involve two steps: finding cluster correspondence and utilizing a fusion strategy to produce a final result. The existing direct approaches mainly discuss the process of finding cluster correspondence, while the fusing process is simply done by voting. In this paper, we mainly focus on the fusing process and propose a Dempster-Shafer evidence theory-based clustering ensemble algorithm. The advantage of the algorithm is that the information of an object's surrounding cluster structure is taken into consideration by using its neighbors to describe it. First, we find neighbors of each object and generate its label probability outputs in every base partition. Second, these label probability outputs are integrated based on DS theory. Theoretically, our method is superior to other voting methods. Besides, several experiments show that the proposed algorithm is statistically better than seven other clustering ensemble methods.

Keywords: Multigranulation, Information fusion, Clustering ensemble, Dempster-Shafer evidence theory

1. Introduction

Granular computing is efficient in designing intelligent systems [30]. As one of the three key issues (information granulation, organization and causation) in granular computing, information granulation is also regarded as one of the fundamental features of human cognitive ability. The task of information granulation is to granulate a data set into granules to obtain a granular structure. Many strategies of information granulation [13, 24, 28, 29, 42] have been proposed to meet different user demands, in which clustering analysis is a widely used strategy. Clustering analysis is an interesting area in machine learning, whose task is to find the structure of data through dividing a data set into clusters. Good clustering often satisfies two requirements: one is that the objects share high similarity in the same cluster, the other is that the objects share high dissimilarity in different clusters. A clustering algorithm is only suitable for a particular data distribution and a particular data model. However, it is often hard to get data distribution and data model. Using a single clustering system to realize information granulation is not always satisfactory. Therefore, clustering ensemble comes into being. This evolution is similar to the development of multigranulation analysis [31, 32, 33, 34]. Multigranulation analysis offers many new ideas for designing clustering algorithms [23, 36, 48]. In turn, multiple clustering systems explore a data set from different viewpoints, so that the data set can be granulated into multiple granular structures. Thus, integrating multiple clustering systems, the so-called clustering ensemble, is an important branch of multigranulation information fusion.

☆This is an extended version of the paper presented at the International Conference of Machine Learning and Cybernetics (ICMLC) 2015, Guangzhou, China.

*Corresponding author
Email addresses: feijiangli@email.sxu.edu.cn (Feijiang Li), jinchengqyh@126.com (Yuhua Qian), jietingw@163.com (Jieting Wang), ljy@sxu.edu.cn (Jiye Liang)