Knowledge-Based Systems 91 (2016) 179–188

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Hierarchical cluster ensemble model based on knowledge granulation

Jie Hu [a], Tianrui Li [a,*], Hongjun Wang [a], Hamido Fujita [b]

[a] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[b] Faculty of Software and Information Science, Iwate Prefectural University, 020-0693, Iwate, Japan

ARTICLE INFO

Article history:
Received 26 July 2015
Revised 14 September 2015
Accepted 3 October 2015
Available online 16 October 2015

Keywords:
Cluster ensemble
Granular computing
Rough sets

ABSTRACT

Cluster ensemble has been shown to be very effective in unsupervised classification learning by generating a large pool of different clustering solutions and then combining them into a final decision. However, the task of it becomes more difficult due to the inherent complexities among base cluster results, such as uncertainty, vagueness and overlapping. Granular computing is one of the fastest growing information-processing paradigms in the domain of computational intelligence and human-centric systems. As the core part of granular computing, the rough set theory dealing with inexact, uncertain, or vague information, has been widely applied in machine learning and knowledge discovery related areas in recent years. From these perspectives, in this paper, a hierarchical cluster ensemble model based on knowledge granulation is proposed with the attempt to provide a new way to deal with the cluster ensemble problem together with ensemble learning application of the knowledge granulation. A novel rough distance is introduced to measure the dissimilarity between base partitions and the notion of knowledge granulation is improved to measure the agglomeration degree of a given granule. Furthermore, a novel objective function for cluster ensembles is defined and the corresponding inferences are made. A hierarchical cluster ensemble algorithm based on knowledge granulation is designed. Experimental results on real-world data sets demonstrate the effectiveness for better cluster ensemble of the proposed method.

1. Introduction

Clustering is an important unsupervised classification technique, which has been extensively researched in different fields such as statistics, pattern recognition, machine learning, and data mining [1–3]. Following several clustering criteria and different methods of similarity measurement, the underlying structure of a data set can be revealed, e.g., the similar objects can be grouped into the same cluster, while dissimilar objects can be assigned to different clusters [4].

Actually, each clustering algorithm has its own strengths and weaknesses, and there is no single clustering algorithm capable of delivering sound solutions for all data sets. With the objective of improving the robustness, consistency, novelty and stability of single clustering algorithm's results, cluster ensemble (cluster fusion, or consensus clustering) has emerged as a tool for leveraging the consensus across multiple clustering results and combining them into an optimal solution. It has gained increasing attention of researchers in recent years [4–8].

Generally, cluster ensemble method involves two major steps: Generation and Consensus Function. In the first step, a set of diverse partitions of objects will be produced using a generative mechanism, such as by homogenous algorithm with different parameters (or initializations) [9,10] or heterogeneous algorithms [5], etc. The consensus function is the main step in any cluster ensemble algorithm, by which a new partition is acquired by integrating all partitions obtained in the generation step. There are numerous consensus function approaches, which can be classified into two main types: methods based on objects co-occurrence and methods based on median partition [4]. In the median partition based consensus function approach, the resulting partition is acquired by finding an optimization partition which maximizes the similarity (or minimizes the dissimilarity) with all partitions in the cluster ensemble. Although a great number of cluster ensemble methods have been proposed over the past years, there are relatively few techniques in handling uncertain, vague and overlapping information in the cluster ensemble process.

Granular computing (GrC) [11,12], emerged as one of the fastest growing information-processing paradigms in the domain of computational intelligence and human-centric systems, has been successfully applied in many fields. As a core part of GrC, the rough set theory (RST) [13] forms the granules through the equivalence relation defined on objects of the universe and approximately express information granulation by using a pair of non-numerical operators, i.e., lower and upper approximation operators [14]. Nowadays, RST has

* Corresponding author. Tel.: +86 28 66367458.
E-mail addresses: jiehu@swjtu.edu.cn (J. Hu), trli@swjtu.edu.cn, trli30@gmail.com (T. Li), wanghongjun@swjtu.edu.cn (H. Wang), HFujita-799@acm.org (H. Fujita).

http://dx.doi.org/10.1016/j.knosys.2015.10.006
0950-7051/

been widely applied in machine learning and knowledge discovery related areas.

Essentially, the clusters in the cluster ensemble can be viewed as a synonym of information granules [15]. From this perspective, in this paper, a knowledge granulation model for hierarchical cluster ensemble (HCEKG) is proposed, which is attempt to provide a new way to tackle the cluster ensemble problem together with new fields of applications for knowledge granulation. We mainly focus on its model establishment and consensus function as it provides more predictable methods for semi-supervised learning for multidimensional data as knowledge granulation model.

The rest of this paper is organized as follows. In Section 2, the background and related studies of GrC and rough clustering are reviewed. Section 3 briefly introduces some basic concepts of knowledge, knowledge granulation and rough sets. In Section 4, the cluster ensemble problem is formalized in the framework of GrC and the cluster ensemble problem is defined as a maximum density granular selection problem. A hierarchical cluster ensemble model based on knowledge granulation is proposed in Section 5 and the corresponding algorithm is also developed in Section 5. Experimental results and the corresponding analysis are presented in Section 6. The paper ends with conclusions and further research topics in Section 7.

2. Related work

2.1. Information granulation and knowledge granularity

GrC involves representing, constructing, as well as manipulating of information granules for complex problem solving [14,16]. The basic notions and principles of granular computing have appeared in several reasoning formalisms, such as in interval analysis, RST, cluster analysis, machine learning, databases, and many others [17]. Since the cluster can be graphically explained by the concept of information granule, more and more scholars make the clustering analysis by using the methods of GrC in the past ten years [18–27].

The key issues of GrC are granules and granulation [28]. A granule may be class, subset, object, or cluster of a universe generated by distinguishability, similarity, and functionality, while granulation involves the operations on granules, that is, construction and decomposition on granules [28,29]. From the viewpoint of information granularity, we will find that the clustering is actually calculated under a uniform granularity and relates to group individual objects into clusters. Partitions and coverings are two simple and commonly used granulations of the universe [30]. In a partition, the subsets of the universe are mutually disjoint, while for a covering, the subsets of the universe maybe overlap. Each element of a partition or covering is treated as a granule in GrC, which can also be further divided through partitioning or covering. At the stage of generation in cluster ensemble, the diversity of partitions or coverings of universe can be obtained via various methods.

In the context of rough sets, knowledge is linked closely with the variety of classification patterns of the universe [31]. The notion of derived knowledge granularity, instantiated from information granularity, can be applied to measure the classification ability of a certain knowledge to a universe. Actually, there are extensive studies on the measurement of knowledge granularity which can be roughly classified into two classes, namely, information-theoretic measures and interaction based measures. For a survey and a class of measures, please refer to [32]. In the information-theoretic measures, Shannon entropy and Hartley entropy were used to design measures of knowledge granularity [33–42]. For example, Qian et al. [39] introduced the concepts of combination entropy, combination granulation, conditional combination entropy and the mutual information as well as their several useful properties. In [40], Liang et al. introduced an axiomatic definition of knowledge granulation for an information system, and presented the revised conventional definitions of accuracy, roughness and approximation accuracy. In [41], by uniformly considering the complete and incomplete knowledge structures in knowledge bases, Qian et al. proposed a novel axiom definition of knowledge granulation in knowledge bases and introduced the notion of knowledge distance for measuring the distance between two knowledge structures in the same knowledge base. Under the ordered information systems, Xu et al. [42] introduced the concepts of knowledge granulation, knowledge entropy and knowledge uncertainty measurement, and also gave the definition of rough entropy of rough sets. In the interaction based measures, the measures of knowledge granulation were designed based on counting the cardinality of an equivalence class under a knowledge [36,37,39,43,44].

2.2. Rough set for clustering or cluster ensemble

RST [13], as a formal model to define and process the information granules, generates the information granules by the equivalence relation in advance, and emphasizes the vagueness description of a given concept (set) X in terms of a pair of approximation operators: lower and upper approximations of X. It has been widely applied in clustering analysis. For a survey of fuzzy and rough approaches and their extensions and derivatives for soft clustering, please refer to [45].

Considering the vague or imprecise boundaries in the web mining clusters, Lingras et al. [46] modified the conventional k-means algorithm to include the effects of lower as well as upper bounds, and developed an unsupervised rough set clustering, e.g., rough k-means. Chen et al. [47] proposed a rough set based clustering method with the aid of Shannon's entropy theory to refine the clustering results by assigning relative weights to the set of attributes according to the mutual entropy values. Mitra et al. proposed a novel rough-fuzzy collaborative clustering by incorporating with the fuzzy sets and rough sets [48]. Peters presented an improved rough k-means clustering method as in [46] by analyzing its objective function, numerical stability, as well as the stability of the clusters and others [49]. Kumar et al. [50] designed a rough agglomerative hierarchical clustering algorithm for sequential data, where generating the initial clusters by using the a similarity upper approximation and merging the acquired initial clusters to get the subsequent clusters by the definition of constrained-similarity upper approximation. Parmar et al. [51] presented a RST based algorithm for clustering categorical data, called Min-Min-Roughness (MMR). MMR was defined as the minimum of the min-roughness of the attributes, which was used to determine the splitting attribute. Compared with MMR, Herawan et al. [52] proposed a clustering attribute selecting method based on RST by replacing the MMR with maximum dependency attributes (MDA) when measuring importance of attributes. In [53], Yanto et al. proposed an alternative technique of the MMR algorithm into variable precision rough set model based clustering method. Gao et al. developed a novel cluster ensemble algorithm in [54], which mainly focused on decomposing the attributes of categorical data into a number of rough subspaces and then used these subspaces to generate partitions. In [55], Janusz et al. applied RST in attribute clustering and selection. Li et al. [56] developed a decision-theoretic rough sets model based c-means clustering approach, which utilized a loss function to capture the loss information of the neighbors. Considering the effect of the imbalanced spatial distribution within a cluster, Zhang et al. [57] defined a hybrid imbalanced measure of distance and density for the rough c-means clustering, and designed the corresponding algorithm. Li et al. [58] proposed a hierarchical clustering algorithm for categorical data by defining the Total Mean Distribution Precision (TMDP) which was used to measure the significance of attributes and combining with the concept of granularity. By combining decision-theoretic rough set model with clustering, Yu et al. [59] proposed a hierarchical clustering algorithm, which can acquire the appropriate number of clusters automatically. Aimed at detecting the outlier of categorical data, Suri et al. [60] proposed a novel