Information Sciences 257 (2014) 14–31

Contents lists available at ScienceDirect

Information Sciences

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

# Iterative meta-clustering through granular hierarchy of supermarket customers and products

Pawan Lingras [a,*], Ahmed Elagamy [a], Asma Ammar [b], Zied Elouedi [b]

[a] Department of Mathematics and Computing Science, Saint Mary's University, Halifax, Nova Scotia B3H 3C3, Canada
[b] LARODEC, Institut Supérieur de Gestion de Tunis, Université de Tunis, 41 Avenue de la Liberté, 2000 Le Bardo, Tunisia

ARTICLE INFO

*Article history:*
Received 31 March 2013
Received in revised form 10 August 2013
Accepted 3 September 2013
Available online 13 September 2013

*Keywords:*
Meta-clustering
Granular computing
Iterative clustering
K-means

ABSTRACT

This paper proposes a novel iterative meta-clustering technique that uses clustering results from one set of objects to dynamically change the representation of another set of objects. The proposal evolves two clustering schemes in parallel influencing each other through indirect recursion. The proposal is based on the emerging area of granular computing, where each object is represented as an information granule and an information granule can hierarchically include other information granules. The paper describes the theoretical and algorithmic formulation of the iterative meta-clustering algorithm followed by its implementation. The proposal is demonstrated with the help of a retail store dataset consisting of transactions involving customers and products. A customer granule is represented by static information obtained from the database and dynamic information obtained from clustering of products bought by the customer. Similarly, the product granule augments the static representation from the database with clustering profiles of customers who buy these products. The algorithm is tested for a synthetic dataset to explore various nuances of the proposal, followed by an extensive experimentation with a real-world retail dataset.



## 1. Introduction

Granular computing is an emerging area of research that provides an ability to create innovative representations of objects that can facilitate the development of new algorithms [1–4]. In granular computing an object is represented as an information granule. For example, in a retail store a customer can be represented by an information granule consisting of the customer's spending habits, profitability, and loyalty to the store. Traditionally, data mining process begins with representation of objects based on raw data from the dataset. Spending and profits for the customer can be easily retrieved from the database of transactions. Loyalty can be represented using the number of visits by the customer to the store. A product can be similarly represented by an information granule consisting of the amount of revenues and profits received from the product as well as the popularity of the product. Again, revenue and profits for the product can be retrieved from the transactions. Popularity of a product can be measured by the number of visits that result in purchase of the product. The metrics described above can be directly retrieved from a transaction database and can be used in data mining activities such as clustering, classification, association mining, and prediction. This paper focuses on clustering. Clustering is an unsupervised learning process that groups similar objects. It is used at various stages in data mining from preliminary exploration of a new dataset,

* Corresponding author. Tel.: +1 902 420 5798; fax: +1 902 420 5035.
*E-mail addresses:* pawan@cs.smu.ca (P. Lingras), elagamy.ahmed@gmail.com (A. Elagamy), asma.ammar@voila.fr (A. Ammar), zied.elouedi@gmx.fr (Z. Elouedi).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.09.018

identification of outliers, as well as sophisticated analysis for decision making. Clustering has been used in a wide variety of applications from engineering [5,6], web mining [7–10], to retail data mining [11].

More refined data mining activities use results from previous data mining activities to improve the quality of results. The examples of such secondary data mining techniques include ensemble of classifiers [12] or stacked regression [13], where the results of previous classification/prediction are combined to produce more accurate results. The combination of the results can be based on a predetermined formulation or could use further machine learning techniques such as the rough set based ensemble proposed by Saha et al. [14]. Recently, Lingras and Rathinavel [15] proposed a novel integrated secondary data mining approach to clustering in a network environment that recursively uses clustering results from previous iteration in clustering of mobile phone users. This paper proposes a similar meta-clustering technique for a granular hierarchy, where clustering results from one set of granules affect clustering results of another set of granules. We choose *k*-means [16,17] as the underlying clustering algorithm. The *k*-means algorithm offers a multitude of advantages for quantitative datasets. It is fast, robust, and is known to provide reasonable clustering results in a wide variety of applications.

We will use a retail dataset to illustrate our proposed iterative meta-clustering in a granular hierarchy. Clustering is used in retail industry for unsupervised identification of customer profiles, which makes it possible for an organization to identify previously unknown characterizations of groups of customers based on their revenue potential, profitability, and loyalty. Similar profiling can be applied to products in the store to categorize them into different categories based on their revenue potential, profitability, and popularity. This paper proposes the use of granular hierarchy to enhance the representation of information granules. For example, we can add the profiles of customers obtained from clustering of customers who buy a product in the product granule. Similarly, we can add the profiles of products obtained from clustering of products bought by a customer in the customer granule. This creates an indirectly recursive definition of information granules. The meta-clustering algorithm proposed in this paper will iterate through granular hierarchy to resolve the indirect recursive granule representations. The resulting meta-clusters of customers will also describe the profiles of products (created as part of the integrated meta-clustering) bought by the customers. On the other hand, the meta-clusters of products will also contain information about the profiles of customers who buy the products from a given cluster.

The rest of the paper is organized as follows. Section 2 provides a review of clustering algorithms and the evaluation of resulting cluster quality. A discussion on the emerging area of granular clustering can be found in Section 3. The mathematical and algorithmic foundations of the proposed iterative meta-clustering for a granular hierarchy is presented in Section 4. Section 5 describes experiments with a synthetic dataset to highlight salient features of the proposal. The granular meta-clustering algorithm is then applied to a real data world dataset in Section 6. A summary and conclusions are reported in Section 8.

## 2. Review of *k*-means and cluster quality indices

This section reviews the conventional clustering with the help of a popular algorithm called *k*-means [17]. Let $X = \{\vec{x}_1, \ldots, \vec{x}_n\}$ be a finite set of objects. Assuming that the objects are represented by *m*-dimensional vectors. A clustering scheme groups *n* objects into *k* clusters $C = \{\vec{c}_1, \ldots, \vec{c}_k\}$. Here, *C* is the set of clusters. And each of the clusters $\vec{c}_i$ is represented by an *m*-dimensional vector, which is the centroid or mean vector for that cluster. Each cluster centroid $\vec{c}_i$ is also associated with a set of objects assigned to the *i*th cluster. We will use $\vec{c}_i$ for both the centroid vector or set representation of *i*th cluster depending on the context.

*k*-means clustering is one of the most popular statistical clustering techniques [16,17]. The objective is to assign *n* objects to *k* clusters. The process begins by randomly choosing *k* objects as the centroids of the *k* clusters. The objects are assigned to one of the *k* clusters based on the minimum value of the distance $\|\vec{x} - \vec{c}_i\|$ between the object vector $\vec{x}$ and the cluster vector $\vec{c}_i$. The distance $\|\vec{x} - \vec{c}_i\|$ can be the standard Euclidean distance.

After the assignment of all the objects to various clusters, the new centroid vectors of the clusters are calculated as:

$$\vec{c}_i = \frac{\sum_{\vec{x}_l \in \vec{c}_i} \vec{x}_l}{|\vec{c}_i|}, \quad \textit{where } 1 \leqslant i \leqslant k.$$

Here $|\vec{c}_i|$ is cardinality of cluster $\vec{c}_i$. The process stops when the centroids of clusters stabilize, i.e. the centroid vectors from the previous iteration are identical to those generated in the current iteration.

Quality of clustering is an important issue in application of clustering techniques to real world data. A good measure of cluster quality will help in deciding various parameters used in clustering algorithms. One such parameter that is common to most clustering algorithms is the number of clusters. Several cluster validity indices have been proposed to evaluate cluster quality obtained by different clustering algorithms. An excellent summary of various validity measures can be found in Halkidi et al. [18]. Many of the cluster validity measures are functions of the sum of within-cluster scatter to between-cluster separation. The scatter within the *i*th cluster, denoted by $S_i$ is defined as follows:

$$S_i = \left( \frac{1}{|\vec{c}_i|} \sum_{\vec{x} \in \vec{c}_i} \|\vec{x} - \vec{c}_i\| \right)^{1/2} \tag{1}$$