Chapter 41
Ensemble Clustering Data Mining and Databases

Slawomir T. Wierzchon
Polish Academy of Sciences, Poland

ABSTRACT

Standard clustering algorithms employ fixed assumptions about data structure. For instance, the k-means algorithm is applicable for spherical and linearly separable data clouds. When the data come from multidimensional normal distribution, so-called EM algorithm can be applied. But in practice, the assumptions underlying given set of observations are too complex to fit into a single assumption. We can split these assumptions into manageable hypothesis justifying the use of particular clustering algorithms. Then we must aggregate partial results into a meaningful description of our data. The consensus clustering does this task. In this chapter, the authors clarify the idea of consensus clustering, and they present conceptual frames for such a compound analysis. Next, the basic approaches to implement consensus procedure are given. Finally, some new directions in this field are mentioned.

INTRODUCTION

Clustering is an exploratory activity relying upon dividing a given collection X of objects, or entities, into a set of categories, called groups or clusters, in such a way that any two objects placed in the same group have more in common than any two objects assigned to different groups. Consensus clustering has been proposed to overcome some drawbacks of individual clustering algorithms. Usually we assume that the clusters are disjoint subsets of X such that the objects belonging to a single cluster are sufficiently similar to each other (i.e. the clusters should be homogeneous), while objects from different clusters should be sufficiently diversified (i.e. clusters should be well separated). Splitting given collection into disjoint clusters is termed hard clustering. Otherwise we say about soft clustering, i.e. – depending on the formalism used – probabilistic or fuzzy clustering.

The most popular clustering algorithm is the *k*-means algorithm producing hard partitions – consult (Jain, 2010) for historical background and deeper discussion of its current improvements and variations. Soft version of the algorithm, called fuzzy *c*-means, was proposed by Bezdek (1981). This author used

letter c to name the number of clusters, hence the name of the algorithm. Both the algorithms minimize the squared-error criteria. They are computationally efficient and do not require the user to specify many parameters. However, there are three main disadvantages of both the algorithms. First, they require that the entities must be represented as points in n-dimensional Euclidean space. To alleviate this assumption Hathaway, Davenport and Bezdek (1989) introduced relational version of fuzzy c-means algorithm: instead of the distance between the points representing the objects, a similarity measure between all pair of objects was used. In case of hard partitions the k-medoids algorithm was proposed: here a dissimilarity measure between pairs of objects replaces the distance measure – see e.g. Section 14.3.10 in (Hastie, Tibshirani, & Friedman, 2009). The second disadvantage results from the way in which objects are assigned to the clusters. Namely, in case of hard k-means each object is located to the cluster with nearest centroid (empirical mean of the cluster). Thus resulting clusters are spherical (more precisely, they are Voronoi regions). A similar assignment rule is used in the fuzzy c-means algorithm. Third disadvantage is that, the clusters should be of approximately similar cardinality and of similar shape. In case of unbalanced clusters erroneous results are frequently obtained. Similarly, if one cluster is located within a ball of small radius and the second – within an ellipsoid with one axis much greater than others, we can obtain erroneous results. Examples of "easy" and "difficult" data are depicted in Figure 1. Left panel presents compact, well separated, convex and linearly separated clusters, while non-convex clusters that are not linearly separated are shown in right panel.

To avoid these disadvantages, the ideas of ensemble methods used by machine learning community to improve results of classification methods, have been adapted to the requirements of clustering. In general, the ensemble methods use multiple models to obtain better predictive performance than could be obtained from any of the constituent models. A nice overview of these methods used in machine learning can be found e.g. in (Zhou, 2012). When transposed to the field of unsupervised learning (i.e. clustering) this idea translates to collecting multiple partitions of the same data. By combining these partitions, it is possible to obtain a good data partition even when the clusters are not compact and/or not well separated (Jain, 2010). Consensus clustering seems to be especially recommendable to analyze huge datasets. As noted in (Hore, Hall, & Goldgof, 2009): "The advantage of these approaches is that they provide a final partition of data that is comparable to the best existing approaches, yet scale to extremely large datasets. They can be 100,000 times faster while using much less memory."

Irregular, of complex shape and structure, clusters is only one aspect of the problem. The other is strictly pragmatic. In some applications we are simply "knowledge consumers", i.e. we use a knowledge created by others. In the context of clustering such knowledge is represented by a set of partitions, and consensus clustering is used to integrate these partitions into consistent form. Strehl and Ghosh (2003) propose the term "Knowledge reuse" to label such an activity. Further, the knowledge acquired in such a way may be prepared using different points of view, different needs or different criteria, and it may be generated by large number of sources. Thus these authors distinguish between feature distributed clustering (FDC) and object distributed clustering (ODC). In first case it is assumed that all the data are available, but each time they are clustered using only a subset of features, or attributes, characterizing each piece of data. In the second case a fixed set of attributes is used but the collection of data vastly exceeds the size of a typical single memory. So, different partitions are obtained by using only pieces of the whole collection. Again consensus clustering allows consolidate these different clusterings into consistent partition (Hore, Hall, & Goldgof, 2009). A nice illustration of the FDC principle is e.g. the study by Helsen and Green (1991) who applied cluster analysis to define market segments for a new computer system. The dispersions in the opinions collected from 319 users resulted in different parti-

Related Content

Accessible and Inclusive Content and Applications

Tom Brunetand P. G. Ramachandran (2017). *Mobile Application Development, Usability, and Security (pp. 54-67).*

www.irma-international.org/chapter/accessible-and-inclusive-content-and-applications/169676

Towards A Virtual Machine Migration Algorithm Based On Multi-Objective Optimization

Xiang Chen, Jun-rong Tangand Yong Zhang (2017). *International Journal of Mobile Computing and Multimedia Communications (pp. 79-89).*

www.irma-international.org/article/towards-a-virtual-machine-migration-algorithm-based-on-multi-objective-optimization/188625

Experimental Validation of Distributed Cooperative Control of Multiple Mobile Robots via Local Information Exchange

Gregory A. Bock, Ryan T. Hendrickson, Jared Allen Lamkin, Brittany Dhall, Jing Wangand In Soo Ahn (2017). *International Journal of Handheld Computing Research (pp. 19-40).*

www.irma-international.org/article/experimental-validation-of-distributed-cooperative-control-of-multiple-mobile-robots-via-local-information-exchange/187199

Mobile Agent Based Network Defense System in Enterprise Network

Yu Cai (2011). *International Journal of Handheld Computing Research (pp. 41-54).*

www.irma-international.org/article/mobile-agent-based-network-defense/51573

A Novel Energy Saving Approach through Mobile Collaborative Computing Systems

Xiaoxin Wu, Huan Chen, Yaoda Liuand Wenwu Zhu (2012). *Emergent Trends in Personal, Mobile, and Handheld Computing Technologies (pp. 90-104).*

www.irma-international.org/chapter/novel-energy-saving-approach-through/65334