ARTICLE IN PRESS

Information Sciences xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

Rough clustering utilizing the principle of indifference

Georg Peters *

Munich University of Applied Sciences, Department of Computer Science and Mathematics, Munich, Germany
Australian Catholic University, Australia

ARTICLE INFO

Article history:
Received 10 April 2013
Received in revised form 25 November 2013
Accepted 13 February 2014
Available online xxxx

Keywords:
Rough k-means
Laplace's principle of indifference
Overlapping clusters

ABSTRACT

Clustering is one of the most widely used method in data mining with applications in virtually any domain. Its main objective is to group similar objects into the same cluster, while dissimilar objects should belong to different clusters. In particular k-means clustering, as member of the partitioning clustering family, has obtained great popularity. The classic (hard) k-means assigns an object unambiguously to one and only one cluster. To address uncertainty soft clustering has been introduced using concepts like fuzziness, possibility or roughness. A decade ago Lingras and West introduced a k-means approach based on the interval interpretation of rough sets theory. In the past years their rough k-means has gained increasing attention. In our paper, we propose a refined rough k-means algorithm that utilizes Laplace's principle of indifference to calculate the means. As we will discuss this provides a sounder justification for the impacts of the objects in the approximations in comparison to established rough k-means algorithms. Furthermore, the weighting in the mean function is based on individual objects rather than on aggregated sub-means. In experiments, we compare the refined algorithm to related approaches.

1. Introduction

The objective of clustering is to group similar objects into the same cluster, while dissimilar objects should belong to different clusters. In particular, k-means clustering has attained great popularity [18,31]. The classic (hard) k-means [30] assigns one object unambiguously to one and only one cluster. To address uncertainty, like overlapping clusters, soft clustering has been introduced. Prominent examples are Bezdek's fuzzy c-means [6,7] as a generalization of Dunn's ISODATA [13] and Krishnapuram and Keller's possibilistic c-means [21].

A decade ago Lingras and West [27,28] introduced rough k-means derived from the interval interpretation of rough sets theory. In the past years their algorithm has gained increasing attention. Survey on recent developments of rough clustering can be found in Lingras and Peters [25,26].

A detailed discussion of the relationship between rough clustering, further soft clustering algorithms and classic k-means would go beyond the scope of our paper. See Lingras et al. [29] for insights on the relationship between classic k-means and rough k-means. For a more general discussion of soft clustering extensions and derivatives the reader is referred to Peters et al. [44].

In our paper, we propose a refined rough k-means algorithm that utilizes Laplace's principle of indifference [22] to calculate the means. Furthermore, the weighting in the mean function is based on the objects instead of the sub-means derived from the approximations.

* Address: Munich University of Applied Sciences, Department of Computer Science and Mathematics, Munich, Germany. Tel.: +49 8912653709.
E-mail address: georg.peters@cs.hm.edu

http://dx.doi.org/10.1016/j.ins.2014.02.073
0020-0255/

The proposed algorithm has five distinct advantages in comparison to established rough k-means algorithms. First, it is more robust with respect to the setting of the initial parameters than Lingras and West's original approach. Second, the impact of an object in a lower approximation on a mean function is higher than the impact of a boundary object. Third, the impact of a boundary object is shared competitively among the clusters. Fourth, the mean function does not have any parameters that need to be set. Fifth, the weighting in the mean function is derived from Laplace's principle of indifference; that is, it stands on firm common ground.

The principle of indifference is often abbreviated by PI; this made us to take the phonetically similar Greek letter π to label the proposed algorithm as *π rough k-means* (*πRKM*).

The remainder is organized as follows. In the next section we discuss original rough k-means and its refinements that are relevant to our paper. Then, in Section 3, we propose π rough k-means and discuss its properties. In Section 4 we conduct experiments on artificial and real data. Section 5 concludes the paper.

2. Foundations of rough k-means clustering

2.1. Rough k-means algorithms in the context of rough sets theory

2.1.1. Rough sets theory

Pawlak [39] proposed rough sets in the beginning of the eighties of the last century. Since then rough sets theory has gained increasing attention and has established itself as a core part of granular computing (see [40] for a recent survey on granular computing). The basic idea of rough sets is to describe a set by two approximations, its lower approximation and its boundary. The lower approximation and the boundary together form the upper approximation of a set. While objects in the lower approximation surely belong to the set, objects in its boundary are possible members only – they may or may not belong to the set. Reasons for an unclear membership of a boundary object include missing or contradicting information. Hence, rough sets theory is a powerful concept to describe a kind of uncertainty that is characteristic for many real life situations. For introductions to rough sets theory the reader is referred to, e.g., Grzymala-Busse [15] or Yao and Slezak [54].

2.1.2. Rough clustering

Rough clustering as introduced by Lingras and West [27,28] is derived from the interval interpretation of rough sets in contrast to the original set-based rough sets theory. See Yao [53] for a discussion on these two views on rough sets. Rough clustering utilizes three basic properties of original rough sets theory [38]. They are:

- If and only if an object is member of a lower approximation of a cluster it is not a member of any other cluster.
- The lower approximation of a cluster is a subset of its upper approximation.
- If and only if an object does not belong to any lower approximation it is member of at least two upper approximations.

In Fig. 1 all possible regions for a three clusters configuration are depicted. The boxes show the upper approximations of the clusters C_1, C_2 and C_3.[1] We define $B_{x_n \in R_i}$ as set of the upper approximations an object x_n $(n = 1, \ldots, N)$ in region R_i $(i = 1, \ldots, 7)$ is member of. $|B_{x_n \in R_i}|$ is the cardinality of $B_{x_n \in R_i}$. For example, an object in region R_6 is member of the upper approximations of the clusters C_2 and C_3. Hence, we get $B_{x_n \in R_6} = \{C_2, C_3\}$ and $|B_{x_n \in R_6}| = 2$. Obviously, for objects in lower approximations $|B_{x_n \in R_i}| = 1$ holds, while for boundary objects $|B_{x_n \in R_i}| \geqslant 2$.

Generally, for three clusters seven regions can be distinguished (see Fig. 1). For illustrative purposes we take a closer look at cluster C_1 only. It consists of the regions R_1, R_2, R_3, and R_4.

- Region R_1 is exclusively covered by the upper approximation of C_1. Hence, this region represents the lower approximation of C_1. The lower approximation is also regarded as positive region of C_1, since its members surely belong to the cluster. We get $B_{x_n \in R_1} = \{C_1\}$ and $|B_{x_n \in R_1}| = 1$.
- Region R_2 is defined by the overlapping upper approximations of C_1 and C_2. Hence, the membership of an object x_n in R_2 is unclear, since it could belong to C_1 or C_2. The region is called boundary region and consists of the boundaries of C_1 and $C_2 : B_{x_n \in R_2} = \{C_1, C_2\}$ and $|B_{x_n \in R_2}| = 2$. In R_3 all three upper approximations overlap. We get $B_{x_n \in R_3} = \{C_1, C_2, C_3\}$ and $|B_{x_n \in R_3}| = 3$. In R_4 the upper approximations of C_1 and C_3 overlap ($B_{x_n \in R_4} = \{C_1, C_3\}$ and $|B_{x_n \in R_4}| = 2$).
- The regions R_5, R_6 and R_7 are not covered by the upper approximation of C_1. Hence, objects in these regions surely do not belong to C_1. So, in contrast to the positive region (lower approximation), this region is referred to as negative region of cluster C_1.

In our paper, the set of objects belonging to the lower approximation of C_k is denoted as $\underline{C_k}$ and the upper approximation as $\overline{C_k}$. The boundary of C_k is indicated by a hat diacritic: $\widehat{C_k}$. This hat diacritic is motivated by the unclear status of the

[1] The arrangement of the clusters in rectangle shapes has been chosen for illustrative reasons only. For visual clarity the boxes representing the lower approximations are displayed with offsets.

Please cite this article in press as: G. Peters, Rough clustering utilizing the principle of indifference, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.02.073