Knowledge-Based Systems 23 (2010) 389–396

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A robust iterative refinement clustering algorithm with smoothing search space

Yu Zong [a,b,*], Guandong Xu [b], Yanchun Zhang [b], He Jiang [a], Mingchu Li [a]

[a] School of Software, Dalian University of Technology, Dalian 116621, China
[b] School of Science and Engineering, Center for Applied Informatics, Victoria University, Melbourne VIC8001, Australia

ARTICLE INFO

Article history:
Received 26 March 2009
Received in revised form 12 January 2010
Accepted 27 January 2010
Available online 4 February 2010

Keywords:
Clustering
Smoothing search space
Kernel smoothing
Heuristic algorithm

ABSTRACT

Iterative refinement clustering algorithms are widely used in data mining area, but they are sensitive to the initialization. In the past decades, many modified initialization methods have been proposed to reduce the influence of initialization sensitivity problem. The essence of iterative refinement clustering algorithms is the local search method. The big numbers of the local minimum points which are embedded in the search space make the local search problem hard and sensitive to the initialization. The smaller number of local minimum points, the more robust of initialization for a local search algorithm is. In this paper, we propose a Top–Down Clustering algorithm with Smoothing Search Space (TDCS3) to reduce the influence of initialization. The main steps of TDCS3 are to: (1) dynamically reconstruct a series of smoothed search spaces into a hierarchical structure by 'filling' the local minimum points; (2) at the top level of the hierarchical structure, an existing iterative refinement clustering algorithm is run with random initialization to generate the clustering result; (3) eventually from the second level to the bottom level of the hierarchical structure, the same clustering algorithm is run with the initialization derived from the previous clustering result. Experiment results on 3 synthetic and 10 real world data sets have shown that TDCS3 has significant effects on finding better, robust clustering result and reducing the impact of initialization.

1. Introduction

Clustering is a useful approach in data mining processes for identifying patterns and revealing underlying knowledge from large data collections. The application areas of clustering include image segmentation, information retrieval, and document classification, associate rule mining, web usage tracking and transaction analysis. Generally, clustering is defined as the process of partitioning unlabelled data set into meaningful groups (clusters) so that intra-group similarities are maximized and inter-group similarities are minimized at the same time.

In essence, clustering involves the following unsupervised learning process, which can be written as:

Define an 'encoder' function $c(x)$ to map each data object x_i into a particular group $G_k(c(x) = k \Rightarrow x \in G_k\ k = 1,\ldots,K)$, so that a cluster criterion $Q(c) = \sum_{k=1}^{K}\sum_{c(x_i)=k,c(x_j)=k} dist(x_i, x_j)$ is minimized.

As we know, this is a classical combinatorial optimization problem and solving it is exactly NP-hard, even with just two clusters [6]. According to computation complexity theory [22], no complete algorithm can get the overall optimal solutions in a polynomial time, unless $P = NP$. Iterative refinement method, a popular approximate algorithm, is widely adopted by various unsupervised learning algorithms. A general iterative refinement clustering process can be summarized as Algorithm 1 [19].

Algorithm 1. General iterative refinement clustering

```
Initialization: Initialize the parameters of the current cluster model.
Refinement: Repeat until the cluster model converges.
    (1) Generate the cluster membership assignments for all data objects, based on the current model;
    (2) Refine the model parameters based on the current cluster membership assignments.
```

The intuitionistic denotation of iterative refinement clustering algorithm is shown in Fig. 1. The horizontal axis denotes feasible solutions of clustering problem and the vertical axis is the corresponding objective function values of feasible solutions. In this paper, the feasible solution is the results of 'encode' function (or the clustering results) and the objective function value is the values of cluster criterion $Q(c) = \sum_{k=1}^{K}\sum_{c(x_i)=k,c(x_j)=k} dist(x_i, x_j)$. Without loss of generality, we assume that point 3 is selected as the initialization of an iterative refinement clustering algorithm, and by repeating step (1) and (2), the algorithm will converge to point 4, one of the feasible solutions with sub-optimal objective function value.

[*] Corresponding author. Address: Center for Applied Informatics, Victoria University, Melbourne VIC8001, Australia. Tel.: +61 3 9919 9750; fax: +61 3 9919 5060.
E-mail address: Yu.Zong@vu.edu.au (Y. Zong).

0950-7051/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2010.01.012

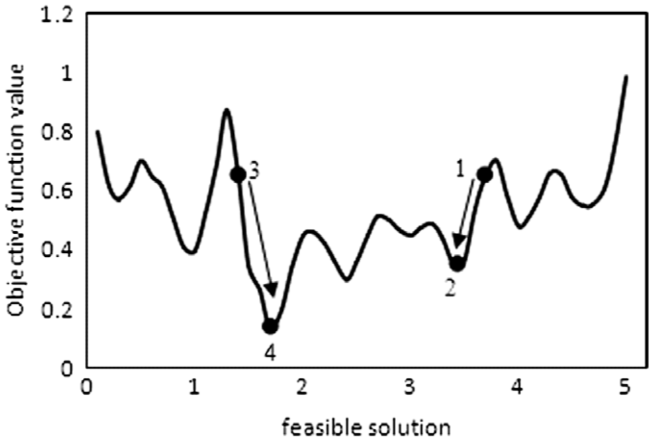

Fig. 1. An example of iterative refinement clustering algorithm.

If point 1 is chosen as the initialization of the same clustering algorithm, it will lead the algorithm converges to point 2, a worse solution with a higher cluster criterion value.

That the initialization model must be correct is an important underlying assumption for iterative refinement clustering algorithm. It can determine the clustering solution [19], that is, different initialization models will produce different clustering results (or different local minimum points as shown in Fig. 1). Since the problem of obtaining a globally optimal initial state has been shown to be NP-hard [9], the study on the initialization methods towards a sub-optimal clustering result hence is more practical, and is of great value for the clustering research. Recently, initialization methods have been categorized into three major families: random sampling methods, distance optimization methods and density estimations [11]. Forgy adopted uniformly random input objects as the seed clusters [8], and MacQueen gave an equivalent way with selecting the first K input objects as the seed clusters [17]. In the FASTCLUS, a K-means variance implemented in SAS [21], the simple cluster seeking (SCS) initialization method is adopted [23]. Katsavounidis et al. proposed a method that utilizes the sorted pairwise distances for initialization [15]. Kaufman and Rousseeuw introduced a method that estimates the density through pairwise distance comparison, and initializes the seed clusters using the input objects from areas with high local density [14]. In Ref. [7], a method which combines local density approximation and random initialization is proposed. Belal et al. find a set of medians extracted from a dimension with maximum and then use the medians as the initialization of K-means [3]. Niu et al. give a novel algorithm called PR (Pointer Ring), which initializes cluster centers based on pointer ring by partition traditional hyper-rectangular units further to hyper-triangle subspaces [18]. The initialization steps of K-means++ algorithm can be described as: choosing an initial center m_1 uniformly at random from data set; and then selecting the next center $m_i = x'$ from data set with probability $dist(x\prime, m)^2/\sum_{x\in D} dist(x, m)$, where $dist(x,m)$ denote the shortest distance from a data object x to the closest center m; iterative until find K centers [1]. The main steps of initialization centers of K-means by kd-tree are: first, the density of a data at various locations are estimated by using kd-tree; and then use a modification of Katsavounidis' algorithm, which incorporates this density information, to choose K seeds for K-means algorithm [20]. And recently, Lu et al. treat the clustering problem as a weighted clustering problem so as to find a better initial cluster center based on the hierarchical approach [16].

The goal of these modified initialization methods, is to reduce the influence of sub-optimal solutions (the local minimum points) bestrewed in the whole search space, as shown in Fig. 1. Although iterative refinement clustering algorithms with these modified initialization methods have some merits in improving the quality of cluster results, they are also have high probability to be attracted by local minimum points. Local search method is the essence of iterative refinement clustering algorithms. Lots of the local minimum points make a local search problem hard and sensitive to the initialization. Those proposed modified initialization methods are only focused on how to select an initialization which can improve the quality of iterative refinement clustering algorithm, but the search space embedded lots of local minimum points is ignored.

Smoothing search space method reconstructs the search space by filling local minimum points, to reduce the influence of local minimum points. In this paper, we first design two smoothing operators to reconstruct the search space by filling the minimum 'traps' (points) based on the relationship between distance metric and cluster criterion. Each smoothing operator has a parameter, smoothing factor, to control the number of minimum 'traps'. And then, we give a top–down clustering algorithm with smoothing search space (TDCS3) to reduce the influence of initialization. The main steps of TDCS3 are to: (1) dynamically reconstruct a series of smoothed search space as a hierarchical structure: the most smoothed search space at the top, and the original search space at the bottom, other smoothed search spaces are distributed between them, by 'filling' the local minimum points; (2) at the top level of the hierarchical structure, an existing iterative refinement clustering algorithm is run with random initialization to generate the cluster result; (3) from the second level to the bottom level of the hierarchical structure, the same clustering algorithm is run with the initialization derived from the cluster result on the previous level. Experiment results on 3 synthetic data sets and 10 real world data sets have shown that TDCS3 has significant effects on finding better, robust cluster result and reducing the influence of initialization.

The contributions of this paper are: (1) we discuss the question why iterative refinement clustering algorithm are sensitive to initialization; (2) we deal with the initialization sensitivity problem by smoothing the search space of iterative refinement clustering algorithms; (3) two smoothing operators are designed based on distance metric; (4) based on the smoothed search spaces, a top–down clustering algorithm is proposed to reduce the influence of initialization. More importantly the existing iterative refinement clustering algorithm can be run in TDCS3 to improve the quality of cluster results.

This paper is organized as follows: in Section 2, we first discuss the local search method and the definition of smoothing search space, and then two smoothing operators are designed based on distance metrics. In Section 3 the top–down clustering algorithm with smoothing search space is proposed, and the strength of the proposed algorithm is also discussed. Then, the experiments on 3 synthetic and 10 real world data sets are conducted and results are presented in Section 4. Finally, we conclude this paper in Section 5.

2. Smoothing search space and smoothing operator

2.1. Local search and smoothing search space

Local search method is the essence of iterative refinement clustering algorithms. During the mid-sixties, local search method was first proposed to cope with the overwhelming computational intractability of NP-hard combinatorial optimization problems. Give a minimization (or maximization) problem with objective function f and feasible region F, a typical local search algorithm requires that, with each solution $x_i \in R^d$, there is associated a