Simulation Modelling Practice and Theory 54 (2015) 49–63

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

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

A novel approach for initializing the spherical *K*-means clustering algorithm

Rehab Duwairi [a,*], Mohammed Abu-Rahmeh [b]

[a] Department of Computer Information Systems, Jordan University of Science and Technology, Irbid, Jordan
[b] Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan

ARTICLE INFO

Article history:
Received 4 December 2014
Received in revised form 19 March 2015
Accepted 20 March 2015
Available online 6 April 2015

Keywords:
Spherical *K*-means clustering
K-means initialization
Intra-cluster similarity
Cluster compactness

ABSTRACT

In this paper, a novel approach for initializing the spherical *K*-means algorithm is proposed. It is based on calculating well distributed seeds across the input space. Also, a new measure for calculating vectors' directional variance is formulated, to be used as a measure of clusters' compactness. The proposed initialization scheme is compared with the classical *K*-means – where initial seeds are specified randomly or arbitrarily – on two datasets. The assessment was based on three measures: an objective function that measures intra cluster similarity, cluster compactness and time to converge. The proposed algorithm (called initialized *K*-means) outperforms the classical (random) *K*-means when intra cluster similarity and cluster compactness were considered for several values of *k* (number of clusters). As far as convergence time is concerned, the initialized *K*-means converges faster than the random *K*-means for small number of clusters. For a large number of clusters the time necessary to calculate the initial clusters' seeds start to outweigh the convergence criterion in time. The exact number of clusters at which the proposed algorithm starts to change behavior is data dependent (=11 for dataset1 and = 15 for dataset2).

1. Introduction

The *k*-means algorithm is an iterative partitioning algorithm that starts with a set of *n* points in R^w, and ends up with a grouping of these points in clusters, by determining a set of *K* points, which are considered as centers of the resulting clusters. The problem is to find the optimal set of such points that minimizes the mean of squared distances from each data instance to its nearest center [30], this objective function is variance based and known as the squared error distortion [27].

Other minimization objective functions include the sum of distances, as in the Euclidean *K*-medians problem [4,33], or to minimize the maximum distance between any point and its nearest center, which is adopted in the Geometric *K*-center problem [1]. Other objective functions describe the goodness of clusters, and these functions are subject to maximization; as in the case of document clustering [17,18], where the function measures the sum of intra cosine similarities within clusters.

The simplicity of the algorithm made it a very attractive choice for clustering, and its known for its efficiency, the algorithm described above is known as Lloyd's algorithm [21]. Obviously; it's similar to the fitting routines; starting with an initial state, then optimizing the parameters subsequently. It has been used in many applications such text categorizations [3], consensus clustering [42,45] and the segmentation of images [5].

* Corresponding author. Fax: +962 2 7201077.
E-mail address: rehab@just.edu.jo (R. Duwairi).

http://dx.doi.org/10.1016/j.simpat.2015.03.007
1569-190X/

Even though the algorithm is simple and efficient, it is not exempt to drawbacks, such as the selection of bad initial centers (slower convergence) [48], which motivated our modifications to the algorithm, and the hill-climbing problem that results in local optimum solutions [7], where local optimality corresponds to the centroidal voronoi problem [19,21,30]. Other problems include the determination of the optimal number of clusters and sensitivity to outliers [12,26,27].

Generally, the application of the *K*-means algorithm requires determining a proper proximity measure, an assessment metric of the clusters' quality, the number of clusters (k), the values of initial means, and a convergence condition. Usually the algorithm terminates when the centroids become stable, but the proximity measure and the objective function are dependent on the type of data being clustered. Initializing the algorithm proved to be a sufficient approach for overcoming the problem of getting stuck at bad local optimum solutions [22,24,25,31].

Clustering algorithms have targeted numerous data types such as textual data [3,20,28,37], images [15] and communication data [44]. Text clustering requires preprocessing the documents, so they are represented via constructing a vector space; a well known procedure to facilitate information retrieving process. A set of terms among all the terms that occur in the corpora are selected to comprise the axis, and the documents are vectors in the space, each of which has term weights as its components [6,11], distinct weights indicate the significance of the corresponding term to the document. Beside the huge size of document collections present, the performance requirements are confronted by the large number of dimensions that occur in such contexts, which is described as the curse of dimensionality, thus simple techniques are desirable.

As text documents are converted into numeric data vectors, clustering can take place by applying the *K*-means, but a number of considerations have to be made, for instance, choosing the Euclidean distance as a proximity measure is not appropriate to cluster document vectors [40], since the location of a pair of points cannot decide their relevance subjectively, rather it is the angle between any two documents that defines their similarity [6,11,39], therefore, it is the notion of direction that must be the foremost rule that guides the process.

A variant of the *K*-means that uses the cosine similarity is known as the spherical *K*-means, this algorithm can be applied to document vectors or any type of directional data. In addition to representing documents in a vector space, the obtained vectors can be normalized to be of unit length in the space, resulting in a set of points that occur on the surface of the unit sphere about the origin, after which the algorithm was named [17,18]. The logical interpretation of the produced clusters that a single cluster is expected to contain documents that belong to semantically related subjects. In a typical data set of documents, large number of terms may not occur in a single document, and the document vector would contain large number of zeros, causing the documents to be sparse [13,17]. Spherical *K*-means has shown its capability of taking advantage of the sparsity of documents.

The authors of this work present a new technique to enhance the performance of the spherical *K*-means to cluster sets of documents efficiently, and propose a new assessment metric that measures the clusters' compactness. First; we will introduce a procedure to initialize the algorithm by finding the seeds by which the algorithm starts. The initialization process relies on perturbing the space systematically, such that these initial points are vectors distributed among the document vectors as evenly as possible, following Anderberg's observation [41]. Starting with such seeds gives the opportunity for the algorithm to find better clusters, and faster convergence conditions.

Secondly; the centroids (mean vectors) of the spherical *K*-means tend to be orthonormal at the time of convergence [18], so we are interested in investigating the compactness of the resulting clusters by presenting a new measure for the vectors' dispersion about the centroids in the directional sense. The proposed measure can be represented as a special case of the moment generating function $G^m(t) = E(e^{t(\cos_i\theta - |c_i|)})$. The formula is obtained by setting $t = 0$ and $m = 2$. Further discussion and interpretation of the formula will be provided in Section 4.

The proposed algorithm and quality function were extensively tested on two different datasets. The first dataset consists of 21,826 documents acquired from [52]. The second dataset is the 20 News Group collection and consists of approximately 20,000 documents [53]. The assessment was based on comparing the proposed algorithm with the standard randomly initialed *K*-means algorithm on three metrics: the quality of generated clusters (intra cluster similarity), the cluster compactness and the time necessary of achieving the convergence function. The proposed algorithm straightforwardly outperforms the randomly initialized *K*-means algorithm when cluster quality and compactness are considered. However, the time necessary to achieve convergence was lesser in the case of the proposed algorithm for small number of clusters but started to increase (even become larger than the random *K*-means) for large number of clusters. The details of the experimentations and the results obtained are explained in Section 5.

The rest of this paper is organized as follows: Section 2 presents documents clustering using spherical *K*-means. Section 3, by comparison, describes existing initialization schemes for the *K*-means algorithm. Section 4 explains the proposed initialization technique and the proposed cluster quality measure. Section 5 summarizes the properties of the datasets used in this research and describes the experiments and the results that were obtained. Finally, Section 6 presents the conclusions of this work.

2. Document clustering using the spherical *K*-means

The *K*-means algorithm can be applied to cluster text documents as well. Normally the proximity measure will be the cosine similarity, when combined with the *K*-means, the resulting algorithm is known as the spherical *K*-means. This variant is isomorphic to the standard version, but differs in the associated accessories such as the specifically designated objective