Information Sciences 220 (2013) 269–291

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Efficient stochastic algorithms for document clustering

Rana Forsati [a,*], Mehrdad Mahdavi [b], Mehrnoush Shamsfard [a], Mohammad Reza Meybodi [c,d]

[a] Faculty of Electrical and Computer Engineering, Shahid Beheshti University, G. C., Tehran, Iran
[b] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[c] Department of Computer Engineering and IT, Amirkabir University of Technology, Tehran, Iran
[d] Institute for Studies in Theoretical Physics and Mathematics (IPM), School of Computer Science, Tehran, Iran

ARTICLE INFO

Article history:
Received 16 January 2009
Received in revised form 18 June 2012
Accepted 20 July 2012
Available online 31 July 2012

Keywords:
Document clustering
Stochastic optimization
Harmony search
K-means
Hybridization

ABSTRACT

Clustering has become an increasingly important and highly complicated research area for targeting useful and relevant information in modern application domains such as the World Wide Web. Recent studies have shown that the most commonly used partitioning-based clustering algorithm, the *K*-means algorithm, is more suitable for large datasets. However, the *K*-means algorithm may generate a local optimal clustering. In this paper, we present novel document clustering algorithms based on the Harmony Search (HS) optimization method. By modeling clustering as an optimization problem, we first propose a pure HS based clustering algorithm that finds near-optimal clusters within a reasonable time. Then, harmony clustering is integrated with the *K*-means algorithm in three ways to achieve better clustering by combining the explorative power of HS with the refining power of the *K*-means. Contrary to the localized searching property of *K*-means algorithm, the proposed algorithms perform a globalized search in the entire solution space. Additionally, the proposed algorithms improve *K*-means by making it less dependent on the initial parameters such as randomly chosen initial cluster centers, therefore, making it more stable. The behavior of the proposed algorithm is theoretically analyzed by modeling its population variance as a Markov chain. We also conduct an empirical study to determine the impacts of various parameters on the quality of clusters and convergence behavior of the algorithms. In the experiments, we apply the proposed algorithms along with *K*-means and a Genetic Algorithm (GA) based clustering algorithm on five different document datasets. Experimental results reveal that the proposed algorithms can find better clusters and the quality of clusters is comparable based on F-measure, Entropy, Purity, and Average Distance of Documents to the Cluster Centroid (ADDC).

1. Introduction

The continued growth of the Internet has made available an ever-growing collection of full-text digital documents and new opportunities to obtain useful information from them [3,16,45]. At the same time, acquiring useful information from such immense quantities of documents presents new challenges which has led to increasing interest in research areas such as information retrieval, information filtering and text clustering. Clustering is one of the crucial unsupervised techniques for dealing with massive amounts of heterogeneous information on the web [9,25,41], with applications in organizing information, improving search engines results, enhancing web crawling, and information retrieval or filtering. Clustering is the

* Corresponding author.
E-mail addresses: rana.forsati@gmail.com (R. Forsati), mahdavi@ce.sharif.edu (M. Mahdavi), m-shams@sbu.ac.ir (M. Shamsfard), mmeybodi@aut.ac.ir (M. Reza Meybodi).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.07.025

process of grouping a set of data objects into a set of meaningful partitions, called clusters, such that data objects within the same cluster are highly similar in comparison with one another and are very highly dissimilar to objects in other clusters.

Some of the most conventional clustering algorithms can be broadly classified into two main categories, hierarchical and partitioning algorithms [23,26]. Hierarchical clustering algorithms [22,28,38,52] create a hierarchical decomposition of the given dataset which forms dendrograma tree by splitting the dataset recursively into smaller subsets, representing the documents in a multi-level structure [14,21]. The hierarchical algorithms can be further divided into either agglomerative or divisive algorithms [51]. In agglomerative algorithms, each document is initially assigned to a different cluster. The algorithm then repeatedly merges pairs of clusters until a certain stopping criterion is met [51]. Conversely, divisive algorithms repeatedly divide the whole documents into a certain number of clusters, increasing the number of clusters at each step. Partition clustering, the second major category of algorithms, is the most practical approach for clustering large data sets [6,7]. They cluster the data in a single level rather than a hierarchical structure such as a dendrogram. Partitioning methods try to divide a collection of documents into a set of groups, so as to maximize a pre-defined objective value.

It is worth mentioning that although the hierarchical clustering methods are often said to have better quality, they generally do not provide the reallocation of documents, which could have been poorly classified in the early stages of the clustering [26]. Moreover, the time complexity of hierarchical methods is quadratic in the number of data objects [49]. Recently, it has been shown that the partitioning methods are more advantageous in applications involving large datasets due to their relatively low computational complexity [8,29,36,49,55]. The time complexity of partitioning techniques are almost linear, which makes them appealing for large scale clustering. The best known method in partitioning clustering is *K*-means algorithm [34].

Although *K*-means algorithm is straightforward, easy to implement, and works fast in most situations, it suffers from some major drawbacks that make it unsuitable for many applications. The first disadvantage is that the number of clusters *K* must be specified in advance. In addition, since the summary statistic that is maintained for each cluster by *K*-means algorithm is simply the mean of samples assigned to that cluster, the individual members of the cluster can have a high variance and hence the mean may not be a good representative for the cluster members. Further, as the number of clusters grows into the thousands, *K*-means clustering becomes untenable, approaching $O(m^2)$ comparisons where m is the number of documents. However, for relatively few clusters and a reduced set of pre-selected features, *K*-means performs well [50]. Another major drawback of the *K*-means algorithm is its sensitivity to initialization. Lastly, the *K*-means algorithm converges to local optima, potentially leading to clusters that are not globally optimal.

To alleviate the limitations of traditional partition based clustering methods discussed above, particularly the *K*-means algorithm, different techniques have been introduced in recent years. One of these techniques involves the use of optimization methods that optimize a pre-defined clustering objective function. Specifically, optimization based methods define a global objective function over the quality of clustering algorithm and traverse the search space trying to optimize its value. Any general purpose optimization method can serve as the basis of this approach such as Genetic Algorithms (GAs) [10,26,40], Ant Colony Optimization [43,46] and Particle Swarm Optimization [11,12,53], which have been used for web page and image clustering. Since stochastic optimization approaches are good at avoiding convergence to a locally optimal solution, these approaches could be used to find a global near-optimal solution [35,30,48]. However the stochastic approaches take a long time to converge to a globally optimal partition.

Harmony Search (HS) [18,32] is a new meta-heuristic optimization method imitating the music improvisation process where musicians improvise the pitches of their instruments searching for a perfect state of harmony. HS has been very successful in a wide variety of optimization problems [17–19,32], presenting several advantages over traditional optimization techniques such as: (a) HS algorithm imposes fewer mathematical requirements and does not require initial value settings for decision variables, (b) as the HS algorithm uses stochastic random searches, derivative information is also unnecessary, and (c) the HS algorithm generates a new vector, after considering all of the existing vectors, whereas methods such as GA only consider the two parent vectors. These three features increase the flexibility of the HS algorithm.

The behavior of the *K*-means algorithm is mostly influenced by the number of specified clusters and the random choice of initial cluster centers. In this study we concentrate on tackling the latter issue, trying to develop efficient algorithms generating results which are less dependent on the chosen initial cluster centers, and hence are more stabilized. The first algorithm, called Harmony Search CLUSTtering (HSCLUST), is good at finding promising areas of the search space but not as good as *K*-means at fine-tuning within those areas. To improve the basic algorithm, we propose different hybrid algorithms using both *K*-means and HSCLUST, that differ on the stage in which we carry out the *K*-means algorithm. The hybrid methods improve the *K*-means algorithm by making it less dependent on the initial parameters such as randomly chosen initial cluster centers, and hence, are more stable. These methods combine the power of the HSCLUST with the speed of *K*-means. By combining these two algorithms into a hybrid algorithm, we hope to create an algorithm that outperforms either of its constituent parts. The advantage of these algorithms over *K*-means is that the influence of the improperly chosen initial cluster centers will be diminished by enabling the algorithm to explore the entire decision space over a number of iterations and simultaneously increasing its fine-tuning capability around the final decision. Therefore, it will be more stabilized and less dependent on the initial parameters such as randomly chosen initial cluster centers, while it is more likely to find the global solution rather than a local one. To demonstrate the effectiveness and speed of HSCLUST and hybrid algorithms, we have applied these algorithms to various standard datasets and achieved very good results compared to *K*-means and a GA based clustering algorithm [4]. The evaluation of the experimental results shows considerable improvements and demonstrate the robustness of the proposed algorithms.