Information Sciences 181 (2011) 2293–2302

Contents lists available at ScienceDirect

Information Sciences

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

Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization

Xiaodi Huang [a,d], Xiaodong Zheng [b,c], Wei Yuan [b,c], Fei Wang [b,c], Shanfeng Zhu [b,c,d,]*

[a] School of Computing and Mathematics, Charles Sturt University, Albury, NSW 2640, Australia
[b] The School of Computer Science, Fudan University, Shanghai 200433, China
[c] Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China
[d] State Key Lab of Software Engineering, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Article history:
Received 18 July 2009
Received in revised form 5 December 2010
Accepted 16 January 2011
Available online 31 January 2011

Keywords:
Biomedical document clustering
Non-negative matrix factorization
Ensemble clustering

ABSTRACT

Searching and mining biomedical literature databases are common ways of generating scientific hypotheses by biomedical researchers. Clustering can assist researchers to form hypotheses by seeking valuable information from grouped documents effectively. Although a large number of clustering algorithms are available, this paper attempts to answer the question as to which algorithm is best suited to accurately cluster biomedical documents. Non-negative matrix factorization (NMF) has been widely applied to clustering general text documents. However, the clustering results are sensitive to the initial values of the parameters of NMF. In order to overcome this drawback, we present the ensemble NMF for clustering biomedical documents in this paper. The performance of ensemble NMF was evaluated on numerous datasets generated from the TREC Genomics track dataset. With respect to most datasets, the experimental results have demonstrated that the ensemble NMF significantly outperforms classical clustering algorithms of bisecting *K*-means, and hierarchical clustering. We compared four different methods for constructing an ensemble NMF. For clustering biomedical documents, this research is the first to compare ensemble NMF with typical classical clustering algorithms, and validates ensemble NMF constructed from different graph-based ensemble algorithms. This is also the first work on ensemble NMF with Hybrid Bipartite Graph Formulation for clustering biomedical documents.

1. Introduction

MEDLINE is the US National Library of Medicine's premier biomedical literature database [30]. Indexing 18 million biomedical documents, MEDLINE has accumulated scientific findings in the biomedical field for more than 40 years. Biomedical researchers regard MEDLINE as the main source for generating scientific hypothesis and discovering new knowledge [15]. With thousands of new citations being added into MEDLINE each day, it is obvious that researchers cannot browse all relevant literature in the database. In order to alleviate this problem, similar biomedical documents are grouped using document clustering techniques [5,33]. In this way, the major findings reported in the literature can be easily digested.

In general, a clustering algorithm needs to address two underlying issues: in which way elements are grouped and what criteria are used for governing such groupings. According to the ways of grouping elements, clustering algorithms are categorized into two types: partitional (flat) clustering and hierarchical clustering [16]. Elements in a partitional clustering are grouped into a number of flat clusters without examining their explicit relationships. Hierarchical clustering, however, produces a hierarchy of clusters in which the different numbers of clusters can be obtained by examining groups at different

* Corresponding author at: The School of Computer Science, Fudan University, Shanghai 200433, China. Tel.: +86 21 65643786.
E-mail address: zhushanfeng@gmail.com (S. Zhu).

0020-0255/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2011.01.029

levels in a tree. In terms of criteria used for finding the clusters, clustering methods can be categorized into various types, such as *K*-means, mixture model-based methods, and graph-based methods [1,6,14]. *K*-means [12] minimizes within-cluster variability and maximizes between-cluster variability. Assuming that data objects in the same cluster are generated from the same distribution, mixture-based clustering methods [6] estimate the parameters of these distributions by maximizing the likelihood objective function. The well-known algorithm, expectation maximization [3] estimates the values of hidden parameters from incomplete data. In addition, a number of clustering algorithms derived from graph partition has been proposed [24]. They aim to minimize the cuts between different subsets of vertices in a graph. Briefly, biomedical researchers face at least two challenges: the huge amount of biomedical literature, and a large number of available clustering algorithms. One of the promising ways for overcoming these challenges is to find a suitable algorithm for clustering biomedical documents. The desirable candidate cluster algorithms should be capable of finding clusters that are an accurate and meaningful summary of the clustered documents.

As one of the dimension reduction methods (such as principal component analysis), non-negative matrix factorization (NMF) is one type of flat clustering method. By imposing constraints of non-negativity elements in both basis and weight matrixes on factorization, NMF has distinct features of preserving the local structure of original data. In the context of image processing, it was first proposed by Lee and Seung [18], and later found its wide applications in other areas, such as information retrieval. Xu et al. applied NMF to clustering general text documents using TDT2 and Reuters document corpora [31]. They found that the NMF-based method is superior to the latent semantic indexing method, and the spectral clustering method. Furthermore, the resulting latent semantic space of clustering documents produced by NMF is explanative intuitively. Specifically, each axis in the space represents the basis topic of a particular cluster, whilst each document in a collection is viewed as the additive combination of these different basis topics. A document is grouped into the cluster where it has the largest projection value. NMF has been applied to many areas in computational biology, ranging from gene expression analysis [17], protein sequence recognition [13], class comparison and prediction [9], cross-platform and cross-species characterization [27], function characterization of genes [22], biological network analysis [29], to biomedical informatics [2]. As an increasingly important tool in computational biology for analysis and interpretation [4], NMF has recently gained more and more attention in biomedical fields.

In this study, we focus on clustering biomedical documents by extending our previous work [36], although many researchers have already studied this topic. Using hierarchical clustering techniques (group-wise average and single pass clustering), Lee et al. clustered 15,405 articles cited in the OMIM database, and an additional 56 articles cited in four biological review articles [20]. For clustering MEDLINE documents on different kinds of diseases, Yoo and Hu [32] compared various approaches, such as *K*-means, bisecting *K*-means, and suffix tree clustering, as well as three hierarchical methods (single-link, complete-link, and average-link). They found that partitional clustering techniques outperform hierarchical clustering techniques significantly in their experiments [32].

Due to the sensitivity to the initial values of their parameters, some clustering algorithms may not converge to the same solution on each run. For obtaining a consensus solution, ensemble clustering has been employed to combine multiple clustering results on a given dataset [7,8,26]. Many methods for constructing a clustering ensemble are graph-based, where a vertex represents a data object and an edge indicates the similarity between two data objects. The Cluster-based Similarity Partitioning Algorithm (CSPA) [26] constructs a similarity matrix between data points. CSPA achieves a moderate performance, however requires significant computation. As one of the ensemble clustering approaches, the Hyper Graph Partitioning Algorithm (HGPA) [26] partitions a hyper-graph that represents clusters. The objective of HGPA is to roughly partition a hyper-graph into the same size of desirable, unconnected components which cut the minimum number of hyper-edges. Unlike CSPA that considers only pairwise similarities, HGPA can reveal many types of relationships between data objects by grouping and collapsing related hyper-edges. Its computation is also cheaper than CSPA. The Meta-Clustering Algorithm (MCLA) assigns each object to its most associated meta-cluster. By collapsing a group of related hyper-edges into a single hyper-edge, MCLA performs better than HPGA with a lower computational cost. A graph model called Hybrid Bipartite Graph Formulation (HBGF) was proposed recently by Fern and Brodley [8], which simultaneously treats both data objects and clusters as the vertices in a bipartite graph. The strength of HBGF is the use of a bipartite graph which retains all the information in base clusters. In addition, the problem of graph partition is solved efficiently.

As mentioned before, biologists face the problem of choosing one from a large number of algorithms for clustering biomedical documents. Our major contribution in this paper is to answer the question as to which algorithm is the most effective in clustering biomedical documents through comprehensive experiments. Precisely, the performance of biomedical document clustering is enhanced by using the ensemble NMF. To the best of our knowledge, this is the first time that NMF has been used to cluster biomedical text documents. In order to speed up the convergence of NMF, we make use of a new algorithm for updating the parameter values which is based on the projected gradient method [21]. Considering the stochastic characteristic of NMF, we use ensemble clustering to achieve a consensus solution from a number of runs of NMF with different initial conditions. Ensemble clustering enables NMF to be less sensitive to these initial conditions, thereby producing robust clustering results. We also experimented with NMF by constructing four ensemble methods including CSPA, HGPA, MCLA, and HBGF. This enabled us to obtain the best results for ensemble NMF. Finally, the performance of ensemble NMF in this research has been tested on a large number of datasets, which were generated from the TREC genomics 2004 track. Not only do we experimentally compare ensemble NMF with the classical clustering algorithms such as bisecting *K*-means, *K*-means, and hierarchical clustering, but also explore four different ensemble methods for constructing the ensemble NMF. In particular, this is the first work on ensemble NMF with Hybrid Bipartite Graph Formulation (HBGF)