Expert Systems With Applications 62 (2016) 358–372

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Hybrid linear matrix factorization for topic-coherent terms clustering

Ping Liang, Sartra Wongthanavasu*

Department of Computer Science, Faculty of Science, Khon Kaen University, Khon Kaen 40002, Thailand

ARTICLE INFO

Article history:
Received 25 February 2016
Revised 20 May 2016
Accepted 12 June 2016
Available online 18 June 2016

Keywords:
Matrix factorization
Dimensional reduction
Term clustering
Karhunen–Loève transformation

ABSTRACT

Topic-coherent term clustering is the foundation of document organization, corpus summarization and document classification. It is especially useful in solving the emerging problem of big data. However, a term clustering method that can cope with high-dimension data with variable length and topics and meanwhile achieve high topic coherence is an ongoing request. It is a challenging problem in research. This paper proposes a hybrid linear matrix factorization method to identify the topic-coherent terms from documents to form a thesaurus for clustering. Starting from an analog Karhunen–Loève transformation from PCA scores fully into FA's factor coefficients space (loadings), the high-dimension of the full set of PCA scores is reduced and topic-coherent terms are classified by the main factors of FA which could be topics. Karhunen–Loève transformation reduces the total mean square error to increase topic coherence. The optimization of the initial transformation is carried out further in a manner of Karhunen–Loève expansion based on stochastic Wiener process. The optimal topic coherent bags of terms are found to build a more topic-coherent model. This approach is experimented on the CISI, MedSH and Tweets dataset in different sizes and number of topics. It achieves outstanding results better than the methods in comparison.

1. Introduction

Text corpora are a large collection of writings in natural text including books, abstracts, emails, news or weather feeds, etc. Text corpora normally are expressed as a sparse and high dimensional representation in variable sizes, including short messages in social media like those in Twitters and Facebook and long text like documents. The topics of text corpora are embedded in the representation and found by categorizing text words into different clusters to form a broad picture of certain knowledge structure. The accuracy of the discovered hidden knowledge structure of text is the key to achieve the goals of automatic document classification, corpus summarization and document organization. As pointed out in Blei (2012), in order to find and construct a thematic knowledge structure for unstructured documents and text, topic model is a solution. And the solution to build up a topic model is to extract topic-coherent terms to form a thesaurus as a topic related model for text documents. Therefore, the critical problem that needs to be addressed is to improve topic coherence, that is, finding topic-coherent words by clustering. It is a challenging research problem.

In general, there are two ways for discovery of latent topic structure by topic coherence – probabilistic methods like Latent Dirichlet Analysis (LDA) and statistical matrix factorization. There are evidences given in Tang, Meng, XuanLong, Mei and Zhang (2014) showing that LDA has its limits when the number of documents is small or the length of documents is too short. When topics are too diverse in large number, LDA cannot be efficient. Non-negative matrix factorization (NNMF) takes a statistical procedure for natural text processing and can achieve the categorization of documents and terms by topic. It indicates the advantages of matrix factorization in topic modeling when considering natural text as "bag of documents and words". Matrix factorization in statistical procedure can not only reduce the dimension of natural text but also expose the latent topic structure in the text in variable length. Apart from the ability of dealing with text in variable length, each way of clustering has the same goal of finding new knowledge by analyzing datasets describing some aspects of the world (Kriegel & Zimek, 2009). Together with Gionis and Mannila (2007) and Udell, Horn, Zadeh and Boyd (2015), these researches suggest the advantages of analysis by combined matrix factorization. Therefore, it is a right way to pursue the combination of different ways of matrix factorization for improving topic coherence.

To achieve the goal above, the normal procedure is in two steps. One is to reduce total mean square error. Two is to optimize. There are many researches in reducing total mean squared

* Corresponding author at: Machine Learning and Intelligent Systems Laboratory, Department of Computer Science, Faculty of Science, Khon Kaen University, Khon Kaen, 40002, Thailand. Tel.: +66 81 9650277, Fax: +6643465377.

E-mail addresses: liang_p@hotmail.com (P. Liang), sartra.wong@gmail.com, wongsar@kku.ac.th (S. Wongthanavasu).

http://dx.doi.org/10.1016/j.eswa.2016.06.022
0957-4174/

error to improve coherence, like by least square regression (Mairal, Bach, Ponce & Sapiro, 2010). The method is applied on a single matrix factorization. The results are optimized by stochastic gradient descent algorithm. And Singh and Gordon (2008) treats multiple matrix factorizations results as a relational database and applies stochastic gradient descent algorithm to optimize the results.

Karhunen–Loève Theorem or Karhunen–Loève Transformation (KLT) has been proved to be a successful method for minimizing total mean squared error (Ghanem & Spanos, 1991). Kittler and Young (1973) focused particularly on the classification potential of KLT and found out that it is particularly useful for pattern recognition when combined with classification procedures based upon discriminant functions obtained by recursive least squares analysis. Keating and Mason (1985) acknowledged the classification feature of KLT and the affect of outliners on the result of KLT classification. In summary, KLT is a method for feature selection and noise reduction.

Moreover, Karhunen–Loève expansion (KL expansion) can optimize the posterior probability distribution in a random field (Dow & Wang, 2015). Chen, Leon,Gibson and Hosseini (2016) embedded KL expansion within the routine of Nondominated Sorting Genetic Algorithm (NSGA-II). The optimization effect has been proved in the application in processing high-dimensional decision variables. However, as early as in 1970, Fukunaga and Koontz (1970) pointed out that before applying the Karhunen–Loève (KL) expansion, it is beneficial to extract "good" features for recognition.

Based on the previous review, it is right to perform KLT as the step one to prepare a “good” foundation. Then KL expansion can proceed for optimization as the step two that is oriented for topic coherence. Therefore, we propose a more effective way of hybrid matrix factorization that can reduce total mean square error to achieve better topic coherence. It is done by Karhunen–Loève Transformation from Principal Component Analysis (PCA) scores to Factor Analysis (FA) loadings initially. The optimization of the results is achieved by a non-parametric KL expansion via a Wiener process to be more deterministic and robust. And the central limit theorem followed by Wiener process guarantees the convergence of KL expansion.

The main contributions of this paper include (1) We propose Karhunen–Loève Transformation from the full PCA scores without truncation to the FA loadings results in a low-rank hybrid matrix. It minimizes the total mean squared error for further analysis as described in Agrawal, Gunopulos, and Raghavan (2005) and Xiu (2010); (2) The columns of the hybrid matrix are fused fully from the PCA components by the FA loadings to form the topics in a reduced number. The rows represent the topic-coherent terms. The process merely merges the full PCA components into the FA main factors. No information of original data are lost; (3) The preceding result is optimized by simulating KL expansion based on a stochastic Wiener process upon it. During the Karhunen–Loève expansion, the most optimal topic-coherent bags of terms are obtained for topic modeling and word categorization. The KL expansion will converge because Wiener process always converges. The method is proved by the document coherence test by UMass (Mimno, Wallach, Talley, Leenders & McCallum, 2011) and UCI (Newman, Bonilla & Buntine, 2011) metric in Section 5; (4) The proposed method can be applied for missing value, reducing outliers and improving clustering robustness. Its total mean square error reduction by KLT and optimization by KL expansion are more deterministic and robust.

The rest of the paper is organized as follows. Section 2 gives the related works. Section 3 introduces our proposed method. Section 4 presents the experiment carried out. Section 5 makes the performance evaluation of the experiment results. And Section 6 discusses the advantage and disadvantage of the proposed method. In Section 7, we conclude the findings and propose the directions for future work.

2. Related works

Matrix factorization in statistical way has many methods, including Principle Component Analysis (PCA), Factor Analysis (FA) and NNMF, etc. Each method has its applications in textual processing as in Zahedi and Sorkhi (2012) and Underhill, McDowell, Marchette and Solka (2007). In Zahedi and Sorkhi (2012), the paper uses PCA and Precision/Recall criteria for text processing. Underhill et al. (2007) compares 5 dimension reduction methods to discuss their advantages and disadvantages in textual processing. In Lan, Waters, and Studer (2014), authors use not only maximum-likelihood to find key concepts in educational domains to complete factor analysis but also further optimize the result in combination of Bayesian solution, called SPARse Factor Analysis (SPARFA). The method limits its usage in text given as student response and the domain of topics also deterministic. In Verma,Jadon and Pujari (2013), a method is proposed to use NNMF and Hadamard product of similarity matrices rather than inner product of tf-idf matrix for short-text clustering. The methods can group documents and words but cannot construct a knowledge structure as a topic model to allow the overlapping of documents and words in different topics. In Crain, Zhou, Yang, and Zha (2012), it combines Latent Semantic Indexing (LSI) and LDA for topic modeling. Its idea is to discuss the possibility of using these methods for large collection of text.

Several approaches have been developed for subspace clustering. In Cui, Fern, and Dy (2010), the paper starts clustering data into k clusters by spectral clustering. Based on the k clusters, an orthogonal subspace is created by the k-means centers of these clusters by SVD to simulate the original data space. The data is then projected into the subspace for dimension reduction. The subspace may not converge. And in Niu, Dy and Jordan (2013), the paper adopts the similar approach but introduces Hilbert-Schmidt Independence Criterion (HSIC) as constraints to achieve a convergence. Both methods can achieve multiple non-redundant clustering. However, as discussed in Niu et al. (2013), the computation complexity is quite high due to the algorithm complexity. In Davidson and Qi (2008), the method does not rely on an existing clustering algorithm. A distance matrix for the original data is used for constructing an alternative distance matrix by remaining the left and right singular vectors and taking the inverse of the singular values of the original distance matrix. The attributes of the original data are preserved while searching for the alternative clusters by minimizing the Kullback–Leibler divergence between the original data and the transformed data.

Based on these previous works, we propose Karhunen–Loève Transformation from the PCA scores of the original data to the main factor loadings of FA to form a new classification of the original data. After the initial projection, the PCA scores of term weight representing the topics are preliminarily classified by the FA factors. In order to optimize the initial clustering, the first projection scores are expanded by a series of Karhunen–Loève expansion based on Wiener orthogonal vector space transformation, which represents the document topics till an optimal topic-coherent term classification is reached.

There are many approaches that collectively analyze the relation between different matrix factorizations, Single Value Decomposition (SVD), NNMF, etc. As in Singh and Gordon (2008) and Bouchard Yin, and Guo (2013), multiple matrix factorization methods are done separately at first and then a relation function like relational database is set up to collect the results in a unified way. Particularly, Singh and Gordon (2008) applies stochastic quasi-newton approximation for optimization of the results. This