Pattern Recognition 47 (2014) 758–768

Contents lists available at ScienceDirect

Pattern Recognition

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

Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction

Kyoungok Kim [a], Jaewook Lee [b,*]

[a] Department of Industrial and Management Engineering, POSTECH, 790-784 Pohang, Kyungbuk, South Korea
[b] Department of Industrial Engineering, Seoul National University, 151-744 Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 26 November 2012
Received in revised form
10 July 2013
Accepted 26 July 2013
Available online 11 August 2013

Keywords:
Text visualization
Semi-supervised dimensionality reduction
Laplacian eigenmaps
Sentiment classification

ABSTRACT

Sentiment analysis, which detects the subjectivity or polarity of documents, is one of the fundamental tasks in text data analytics. Recently, the number of documents available online and offline is increasing dramatically, and preprocessed text data have more features. This development makes analysis more complex to be analyzed effectively. This paper proposes a novel semi-supervised Laplacian eigenmap (SS-LE). The SS-LE removes redundant features effectively by decreasing detection errors of sentiments. Moreover, it enables visualization of documents in perceptible low dimensional embedded space to provide a useful tool for text analytics. The proposed method is evaluated using multi-domain review data set in sentiment visualization and classification by comparing other dimensionality reduction methods. SS-LE provides a better similarity measure in the visualization result by separating positive and negative documents properly. Sentiment classification models trained over reduced data by SS-LE show higher accuracy. Overall, experimental results suggest that SS-LE has the potential to be used to visualize documents for the ease of analysis and to train a predictive model in sentiment analysis. SS-LE can also be applied to any other partially annotated text data sets.

1. Introduction

At present, the amount of information from online and offline documents is accumulating rapidly, thereby increasing the need for effective and efficient analysis of text data. Sentiment analysis is an emerging part of textual data analytics for extracting the subjectivity and opinion of specific topics in documents [40]. Sentiment analysis originally focused on reviews of products and extended its application areas to Twitter [54], news [17], and blogs [17,37] because of its potential application. However, sentiment analysis suffers from high complexity of data analysis and degrading prediction models that arise from high-dimensional text data, which is common in other text mining applications.

Text data visualization is one of the widely used tools for analyzing high-dimensional text data. The tool transforms data into two- or three-dimensional space using appropriate dimensionality reduction techniques that closely correspond to the proximities in the original dimension of the data set to figure out the distribution of documents for discovering patterns [60]. In addition, the reduction of redundant features of the data makes data analysis, such as interpreting the features and uncovering latent useful information, easier and saves the cost and time needed to build a predictive model. Several dimensionality reduction methods have been successfully applied to real-world data sets [45,53,46]. However, traditional dimensionality reduction methods are usually unsupervised, that is, researchers only consider input features during the dimension reduction process. If possible, it is desirable to utilize the availability of the sentiment label information of some portion of data since data points with the same labels are located nearby, whereas data points with different labels located faraway in reduced embedded space are necessary for better sentiment visualization. In many real-world problems, such as image and web classification, however, the labeling process of documents for sentiment analysis is very time-consuming and costly, but obtaining unlabeled data samples is relatively easy (hence clustering them [26–30,32,33]). The subjectivity of documents should be obtained by human annotation, which needs several trained people to determine the true sentiment of documents. For this reason, only a small portion of entire documents is usually labeled by human annotation and the other remaining documents are left without labels. Therefore, the application of semi-supervised learning to text data sets with both labeled and unlabeled information, instead of supervised learning requiring labeling information of entire data points, is more realistic.

In this paper, a semi-supervised nonlinear dimensionality reduction method is proposed to adopt label information of a small portion of data in calculating low-dimensional coordinates for better reduction result. Nonlinear dimensionality reduction methods are superior to linear methods in capturing a nonlinear

* Corresponding author. Tel.: +82 2 880 7176.
E-mail addresses: foriness@postech.ac.kr (K. Kim), jaewook@snu.ac.kr (J. Lee)

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.07.022

structure of data [56]. Thus, the proposed method is based on a nonlinear method, Laplacian eigenmaps (LE), and employs a unified objective function calculated from unlabeled and labeled graphs. The unlabeled graph is an original graph in LE. In addition to the original graph, the labeled graph, whose weights are determined by label information, is introduced. The proposed method is evaluated using a multi-domain review data set in sentiment visualization and classification. In reducing features to two or three dimensions, the proposed method can take advantage of label information, which helps in accurately calculating the conceptual similarity between two documents.

The remainder of this paper is organized as follows. The next section reviews the existing literature on sentiment analysis and dimensionality reduction with a brief introduction of LE. The third section explains how a multi-domain review data set is used for experiments. Data preprocessing steps and details about experimental methods in sentiment visualization and classification and experimental results are presented in the fourth section. Finally, the last section presents the conclusions, with the summary and future directions.

2. Literature review

2.1. Sentiment analysis

Sentiment analysis is one of the applications of natural language processing, computational linguistics, and textual analytics to detect and extract subjective information in text documents. Sentiment analysis determines the attitude or emotional information of writers with respect to the main topic treated in each document, by detecting the polarity of documents, and the negative or positive attitude. Applying sentiment analysis enables many companies and analysts to interpret the responses of customers, market trends, and other useful latent information from the huge number of web pages, blogs, and news. Furthermore, analyses enable companies to respond appropriately to customer needs at the right time. Therefore, sentiment analysis has received considerable attention recently from many researchers.

There are two approaches for polarity detection in sentiment analysis that classifies negative and positive sentiment: computational linguistic approach and machine learning approach. The former approach is based on linguistic backgrounds. In this approach, the semantic orientation and polarity of individual words or phrases are first explored. Using this information, the polarity of entire documents is classified automatically by calculating the score based on the occurrences of lexicon words [51,36]. The lexicon used in previous studies is obtained from a variety of lexical sources, such as the General Inquirer lexicon [50] and SentiWordNet [14,1]. Improved versions of algorithms have been developed as well. The main assumption of this approach is that positive words exist with higher probability than negative words in documents with positive sentiment. This approach takes little computation time and cost after the polarity of words or phrases is determined.

Another approach is based on learning classifiers and utilized many classification methods such as Naive Bayes [38,41], support vector machine (SVM) [38,41,39], neural networks [8] and Bayesian network [2] because sentiment analysis for polarity detection determining positive or negative can be viewed as a former binary classification problem dealt in machine learning society. Machine learning methods show better prediction abilities than the former methods partly because they use labeled data samples by domain expert for training a model as mentioned in previous researches. However, when the trained model in a specific domain is transferred to another domain, the classifier often shows poor performance. Many algorithms have been proposed to overcome this domain-transfer problem [6,52]. Recently, a hybrid approach that combines two approaches has also been considered advantageous [43,35].

Sentiment analysis, which especially uses machine learning methods, suffers from the curse of dimensionality which may degrade the performance of learning algorithms because of the high dimensionality of text data sets [15]. An effective dimensionality reduction or, preferably, visualization is needed to facilitate the sentiment analysis of text documents embedded on high-dimensional vector space. The visualization of text documents easily captures some interpretable patterns using the similarity among text documents, given that conceptual similarity from spatial proximity in low-dimensional space can be inferred. The use of graphical representations facilitates the determination of ambiguous documents in deciding their sentiment or outlier documents that are too much off center. In addition, the method enables building better prediction models by reducing a large number of needless features.

Several ways to visualize text documents have been developed. One is latent semantic indexing, which aims to understand the semantic orientation of individual words or phrases and approximate a low-rank matrix [13,16]. Recently, two SVDs were confirmed that produce better low-dimensional representation compared with SVD used in latent semantic indexing when final representations were combined with deep learning [47]. The other approach is the application of dimensionality reduction methods, such as principal component analysis (PCA), multi-dimensional scaling (MDS), and self-organizing maps, to text visualization [21]. PCA and t-distributed stochastic neighbor embedding (t-SNE) combined with domain knowledge, which are used to reduce dimensionality of text data for sentiment and topic visualization, belong to this class [34]. Previous studies using dimensionality reduction methods have the same limitation of unsupervised dimensionality reduction methods that is lack of capacity to utilize label information, because they borrow existing methods as it is or with minor modifications. Furthermore, some works have introduced feature selection to reduce the number of features of text data in [57–59]. However, reduction performance of feature selection is usually worse than that of dimensionality reduction under the same number of features and feature selection is not suitable for data visualization.

2.2. Dimensionality reduction

Real world data in computer vision, bio-informatics, econometrics, and text analytics are usually highly dimensional. However, the performance of prediction models is degraded and takes much time and cost to build models in the original high-dimensional space because of dimensionality. In dealing with high-dimensional data effectively, many studies have applied dimensionality reduction methods to avoid dimensionality and to generate better interpretations.

Dimensionality reduction methods can be categorized into linear and nonlinear types. Linear methods are based on the linearity assumption and are fast and easy to implement. The most popular linear methods are PCA [20] and MDS [25], which are widely used in marketing and business application for data visualization and feature extraction. However, these methods are not appropriate to nonlinear data sets, and many complex data sets, such as images, do not satisfy the assumption of linearity.

Recently, nonlinear types of dimensionality reduction methods aim to overcome disadvantages of linear methods [24,42]. In contrast to linear methods, nonlinear ones are able to deal with nonlinear structures of data. Nonlinear methods have advantages over traditional linear methods because they can capture an