Expert Systems with Applications 42 (2015) 3105–3114

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering

Kusum Kumari Bharti [*], Pramod Kumar Singh

Computational Intelligence and DataMining Research Lab, ABV-Indian Institute of Information Technology and Management Gwalior, Morena Link Road, Gwalior, Madhya Pradesh, India

ARTICLE INFO

Article history:
Available online 27 November 2014

Keywords:
Text clustering
Feature selection
Feature extraction
Term variance
Document frequency
Principal component analysis

ABSTRACT

High dimensionality of the feature space is one of the major concerns owing to computational complexity and accuracy consideration in the text clustering. Therefore, various dimension reduction methods have been introduced in the literature to select an informative subset (or sublist) of features. As each dimension reduction method uses a different strategy (aspect) to select a subset of features, it results in different feature sublists for the same dataset. Hence, a hybrid approach, which encompasses different aspects of feature relevance altogether for feature subset selection, receives considerable attention. Traditionally, union or intersection is used to merge feature sublists selected with different methods. The union approach selects all features and the intersection approach selects only common features from considered features sublists, which leads to increase the total number of features and loses some important features, respectively. Therefore, to take the advantage of one method and lessen the drawbacks of other, a novel integration approach namely modified union is proposed. This approach applies union on selected top ranked features and applies intersection on remaining features sublists. Hence, it ensures selection of top ranked as well as common features without increasing dimensions in the feature space much. In this study, feature selection methods term variance (TV) and document frequency (DF) are used for features' relevance score computation. Next, a feature extraction method principal component analysis (PCA) is applied to further reduce dimensions in the feature space without losing much information. The effectiveness of the proposed method is tested on three benchmark datasets namely Reuters-21,578, Classic4, and WebKB. The obtained results are compared with TV, DF, and variants of the proposed hybrid dimension reduction method. The experimental studies clearly demonstrate that our proposed method improves clustering accuracy compared to the competitive methods.

1. Introduction

Due to proliferate usage of the Internet, the amount of the digital documents is increasing exponentially. It makes automatic processing of these documents an indispensable need of the current environment. Text clustering is an automatic way of grouping the digital documents in a form of clusters based on their intrinsic characteristics. Due to automatic and proficient processing of the digital documents, text clustering is applied to several application domains such as organization of the results returned by a search engine in response to a user's query (Zamir, Etzioni, Madani, & Karp, 1997), browsing large document collections (Cutting, Karger, Pedersen, & Tukey, 1992), topic detection (Huang, Peng, Niu, & Wang, 2011), and generating a hierarchy of web documents (Koller & Sahami, 1997). Various clustering methods, e.g., *k*-means (MacQueen et al., 1967), expectation–maximization clustering (Dempster, Laird, & Rubin, 1977), and density based clustering (Kriegel, Kröger, Sander, & Zimek, 2011) have been proposed in the past several years to achieve these tasks.

In text clustering, documents are traditionally represented as bag-of-words (Salton & Yang, 1975), where each distinct term present in a document collection is considered as a separate dimension (feature). Hence, a document is represented by a multi-dimensional feature vector where each dimension corresponds to a weighted value of the term within the document collection. This weighted value is computed using term frequency inverse document frequency (*tfidf*). As features originate from distinct terms, a corpus of even moderate-sized documents results in hundreds of thousands of dimensions. One of the most important issue in the text clustering is therefore to deal with high

[*] Corresponding author.
E-mail addresses: kkusum.bharti@gmail.com (K.K. Bharti), pksingh@iiitm.ac.in (P.K. Singh).

http://dx.doi.org/10.1016/j.eswa.2014.11.038
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

dimensionality of the feature space. Immoderate number of features not only increases computational complexity but also deteriorates performance of the clustering method. This problem is increasing day by day with the advancement of the digital document processing. As a consequence, the role of dimension reduction in the text clustering has been shifted from an optional step to a mandatory step. The primary aim of the dimension reduction method is to select a discriminative subset of features from a high dimensional feature space without sacrificing performance of the underlying method. Traditionally, dimension reduction methods are classified as feature extraction (Wang & Paliwal, 2003; Burges, 2005) and feature selection (Blum & Langley, 1997; Liu, Kang, Yu, & Wang, 2005; Saeys, Inza, & Larrañaga, 2007) methods.

The feature extraction methods also known as feature construction methods transform a high dimensional feature space into a distinct low dimensional feature space through a combination or transformation of the original feature space. Principal component analysis (Pearson, 1901), latent semantic indexing (Deerwester, 1988), independent component analysis (Comon, 1994), multidimensional scaling (Kruskal & Wish, 1978), and partial least square (Tenenhaus, Vinzi, Chatelin, & Lauro, 2005) are few examples of feature extraction methods. In this study, we use PCA to reduce dimensions in the feature space.

The filter, wrapper, and embedded methods are three subcategories of the feature selection. Filter methods perform statistical analysis of the feature set to select a discriminative subset of the features. On the other hand, the wrapper (Maldonado & Weber, 2009; Bradley & Mangasarian, 1998) and embedded methods (Miranda, Montoya, & Weber, 2005; Weston, Elisseeff, Schölkopf, & Tipping, 2003) use learning method in order to assess the quality of a given feature set. Though wrapper and embedded methods have an advantage of achieving higher accuracy than filter methods, the disadvantages are that they are computationally more expensive and obtain feature subsets that are biased towards the learning method used. As filter methods consider only intrinsic characteristics of the documents for feature subset selection, they are comparatively fast and general in the sense that the subset obtained is not biased in favor of a specific learning method. Hence, filter methods are widely used to reduce dimensions, especially when dimensions in the feature space are huge. DF (Liu et al., 2005), TV (Liu et al., 2005), term strength (TS) (Yang, 1995), information gain (IG) (Quinlan, 1986), and chi-square (CHI) (Li, Luo, & Chung, 2008), odds Ratio (OR) (Mengle & Goharian, 2009), mutual Information (MI) (Peng, Long, & Ding, 2005), information gain (IG) (Liu et al., 2005), gini index (GI) (Shang et al., 2007), improved Gini index (GINI) (Mengle & Goharian, 2009), distinguishing feature selector (DFS) (Yang, 1995), genetic algorithm (GA) (Wu, Tang, Hor, & Wu, 2011), ant colony optimization (ACO) (Janaki Meena, Chandran, Karthik, & Vijay Samuel, 2012), trace oriented feature analysis (TOFA) (Yan et al., 2011), are few examples of the feature selection methods. A comparative summary of dimension reduction methods is presented in Table 1.

All single dimension reduction methods consider only one aspect of the features for the feature subset selection. Consideration of wider (different) aspects altogether is not possible with a single dimension reduction method. Therefore, recently hybrid methods have received considerable attention for dimension reduction. They integrate different dimension reduction methods considering different aspects of the features into one.

Menga, Lin, and Yu (2011) integrate feature contribution degree (FCD) with LSI to create a discriminative subset of features. They first use a feature selection method namely FCD to select a discriminative features sublist and then construct a new semantic space using the LSI. They demonstrate effectiveness of their method on a spam database categorization. Song and Park (2009) also use LSI to reduce dimensions in the feature space. They demonstrate superiority of their approach genetic algorithm based on a latent semantic model (GAL) over conventional GA applied in VSM on Reuters-21,578 document dataset. Though LSI reduces dimensions in the feature space significantly, reduced feature space still suffers from the irrelevant features.

Akadi, Amine, Ouardighi, and Aboutajdine (2011) propose a two-stage dimension reduction method for gene selection. They integrate maximum relevance minimum redundancy (MRMR) with GA to create an informative gene subset. They initially apply MRMR to filter out the noisy and redundant genes from high dimensional gene space and then utilize GA to select a subset of relevant discriminative features. The authors employ support vector machine (SVM) and naive bayes (NB) classifiers to assess fitness of the selected genes. Their experimental results illustrate that their method is able to select smallest gene subset that achieves the highest classification accuracy to its competitors in leave-one-out-cross-validation (LOO-CV). Zhang, Ding, and Li (2008) and Unler, Murat, and Chinnam (2011) also use MRMR to select a discriminative subset of features. Zhang et al. (2008) integrate MRMR with ReliefF (Kononenko, 1994), which is an extension of Relief (Kira & Rendell, 1992). Unler et al. (2011) integrate MRMR with the discrete PSO to bring efficiency and accuracy of the filter and wrapper methods respectively to select a discriminative subset of features.

Uğuz (2011) uses a hybrid approach to create an informative feature subspace. He introduces a FS-FS method (IG-GA) and a FS-FE method (IG-PCA) to transform a high dimensional feature space into a low dimensional subspace. First, each feature present in the document is ranked based on its discriminative power for classification using FS method IG. In the second stage, a FS method (GA) and a FE method (PCA) are used separately in two distinct experiments to reduce dimensions in the feature space. To assess effectiveness of his proposed methods, the author employs k-nearest neighbour (KNN) and C4.5 decision tree on Reuters-21,578 and Classic3 datasets. The experimental results demonstrate that the all hybrid methods (IG-GA and IG-PCA) are effective in terms of the precision, recall and F1-score. The author integrates filter and wrapper methods. Though his method yields good performance, it generates classifier specific feature subsets, hence leads to overfitting problem. Moreover, the IG-GA considers an interaction with classifier to select a discriminative feature subset, which makes the dimension reduction task computationally expensive.

Micro array data is often asymmetric, redundant, and noisy in nature. Most of these genes are noninformative for classification tasks. To select informative subset of genes, Sahu and Mishra (2012) present a two-stage dimension reduction method. In the first stage, the dataset is grouped using k-means and signal to noise

Table 1
Summary of dimension reduction methods.

Method	Main idea	Strength	Weakness
Feature extraction (FE)	Summarize the dataset by creating linear combinations of the features	Preserves the original, relative distance between objects, covers latent structure	Less effective in case of large number of irrelevant features
Feature selection (FS)	Select a sublist of relevant features based on defined criteria	Robust against irrelevant features	Does not cover latent structure