Engineering Applications of Artificial Intelligence 70 (2018) 25–37

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A novel multivariate filter method for feature selection in text classification problems

Mahdieh Labani [a], Parham Moradi [a,*], Fardin Ahmadizar [b], Mahdi Jalili [c]

[a] Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran
[b] Department of Industrial Engineering, University of Kurdistan, Sanandaj, Iran
[c] School of Engineering, RMIT University, Melbourne, Australia

ARTICLE INFO

Keywords:
Text classification
Feature selection
Dimensionality reduction
Filter approach
Multivariate analysis

ABSTRACT

With increasing number of documents in digital format, automatic text categorization has become a crucial task in pattern recognition problems. To ease the classification task, feature selection methods have been introduced to reduce the dimensionality of the feature space, and thus improve the classification performance. In this paper a novel filter method for feature selection, called Multivariate Relative Discrimination Criterion (MRDC), is proposed for text classification. The proposed method focuses on the reduction of redundant features using minimal-redundancy and maximal-relevancy concepts. To this end, the proposed method takes into account document frequencies for each term, while estimating their usefulness. The proposed method not only selects the features with maximum relevancy, but also the redundancy between them is takes into account using a correlation metric. MRDC does not employ any learning algorithm to evaluate the usefulness of the selected features, and thus it can be categorized as a filter method. In order to assess the effectiveness of the proposed method, several experiments are performed on three real-world datasets. The obtained results are compared to the state-of-the-art filter methods. The reported results show that in most cases MRDC results in better classification performance than others.

1. Introduction

With increasing growth of the Internet and information technologies, the massive volume of electronic text documents are given through web pages, the news feeds, electronic emails and digital libraries. To handle such massive information, text categorization has become a key technology to discover and classify text documents. Text categorization is defined as a task of automatically classifying unlabeled documents into predefined categories (Adeva and Atxa, 2007). It has been successfully developed in many applications such as topic detection (Zeng and Zhang, 2007), spam e-mail filtering (Guzella and Caminhas, 2009), SMS spam filtering (Idris and Selamat, 2014), author identification (Zhang et al., 2014), Bioinformatics (Saeys et al., 2007, Tabakhi et al., 2015), web page classification (Özel, 2011), document classification (Jiang et al., 2016) and sentiment analysis (Medhat et al., 2014). In the process of text categorization, documents are generally modeled as a vector space, in which each word is considered as a feature. In the vector model of a document, the value of a feature can be its corresponding word's frequency or term frequency-inverse document frequency (tf-idf). One of the most important issues in the text categorization is to deal with high dimensionality of the feature space. Excessive number of features not only increases the computational time, but also degrades the classification accuracy (Shang et al., 2013). Feature selection and extraction are two main approaches for reducing the dimensionality of the text feature space (Bharti and Singh, 2015). The feature extraction refers to the process of generating a small set of new features by combining or transforming the original ones (Agarwal and Mittal, 2014), while in the feature selection the dimension of the space is reduced by selecting the most prominent features (Saleh and El-Sonbaty, 2007).

Feature selection methods can be classified into four categories: filter, wrapper, embedded and hybrid approaches. Filter methods perform a statistical analysis over the features space to select a discriminative subset of features. In the wrapper approach, various subsets of features are first identified, and then evaluated using classifiers (Agarwal and Mittal, 2014). The hybrid approach takes advantages of both filter and wrapper approaches, and in the embedded approach the feature selection process is embedded into the training phase of the classification

* Corresponding author.
E-mail addresses: m.labani@eng.uok.ac.ir (M. Labani), p.moradi@uok.ac.ir (P. Moradi), f.ahmadizar@uok.ac.ir (F. Ahmadizar), mahdi.jalili@rmit.edu.au (M. Jalili).

https://doi.org/10.1016/j.engappai.2017.12.014
Received 31 May 2016; Received in revised form 23 November 2016; Accepted 28 December 2017
0952-1976/

process (Chouaib et al., 2008). Due to the use of a learning model in the selection process of wrapper, hybrid and embedded approaches, they have an advantage of achieving higher accuracy compared to the filter approach, while being more computationally expensive. The filter approach is often fast and its results are not biased to the choice of classifiers, and thus are widely used to reduce the dimensionality, especially for large-scale feature spaces (Günal, 2012).

In general, the filter approach can be classified into univariate and multivariate methods. In the univariate methods, a specific criterion is used to evaluate the relevance of features independently. Although, these methods can effectively identify irrelevant features, they are unable to remove redundant ones. In other words, univariate filter methods only evaluate features individually, and completely ignore the redundancy between them. On the other hand, multivariate methods consider correlation between features in their process, and thus can handle both irrelevant and redundant features. Although the performance of multivariate methods is better than the univariate ones, they are computationally inefficient.

Relative discrimination criterion (RDC) is an effective univariate filter criterion, which has been recently proposed to reduce the dimensionality of text data (Rehman et al., 2015). In this method, it is assumed that the terms that frequently occur in a specific class compared to others have much higher discriminative properties, and thus are assigned high scores. Although RDC is an effective method for identifying relevant features, the correlation between features is ignored in its evaluation process, and thus it cannot identify redundant features. There often remarkable number of correlated features in the text data identifying which leads to enhance the quality of text classifiers. On the other hand, the aim of feature selection is to select a compact feature subset with maximal discriminative capability, which requires having a high relevance to class labels and low redundancy within the selected feature subset. To reach this goal, in this paper, a novel multivariate feature selection method, called Multivariate Relative Discrimination Criterion (MRDC), is proposed to consider both relevancy and redundancy concepts in its evaluation process. To this end, the proposed method first computes the relevancy of each feature using RDC measure, and then Pearson correlation is used to compute correlation values between features. This results in avoiding higher correlated features. Several experiments are performed on three real-world datasets including Reuters-21,578, 20-Newsgroups and WebKB to evaluate the performance of the proposed method. The reported results reveals that MRDC performs much better compared to state-of-the-art filter methods.

The rest of the paper is organized as follows. Section 2 gives a brief review of previous works. Section 3 presents the details of proposed MRDC method and experimental results are reported and discussed in Section 4. Finally, Section 5 concludes the paper.

2. Literature review and background

2.1. Overview of feature selection methods for text classification

Text categorization is to assign documents to one or more classes. Manual text classification is time-consuming, especially for large-scale dataset; therefore automatic text classification methods have been increasingly used in various applications (Perikos and Hatzilygeroudis, 2016). A text document is a collection of words arranged according to their corresponding language grammatical rules. Although arrangement of words is necessary for constructing meaningful sentences, a text document is usually represented as a "bag of words" for text classifiers, where the order of the words is not considered in the classification process (Badawi and Altinçay, 2014). Therefore, a document d_j is represented as a vector $d_j = \{tw_{1j}, tw_{2j}, \ldots, tw_{vj}\}$ where tw_{ij} shows the weight of ith term from a vocabulary of words $T = \{t_1, t_2, \ldots, t_v\}$. A general method to weight the terms in documents is $tf.idf$, where $tf(t,d)$ and $idf(t,d)$ are the term frequency and the inverse document frequency of term t in document d, respectively (Erenel and Altinçay, 2012). The term frequency tf is the term count normalized by the document size, while idf is defined as $log(N/df)$, where the document frequency df is the number of documents containing a specific term.

Text classification tasks often involve thousands of features, and the classification is indeed a high dimensional problem. Although there are tens of thousands of words in a typical text collection, most of them contain little or no information to predict the text label. The relevance of a feature indicates that the feature is always necessary for predicting the class label, and feature redundancy is usually defined in terms of some kind of correlations in the features. The goal of feature selection is to select a highly-relevant subset with minimum redundancy. To this end, dimensionality reduction approaches (such as feature extraction or selection) is curtail not only to improve the classifier's prediction performance, but also to reduce storage requirements.

Feature extraction methods can be used to reduce the size of feature vector by transforming a higher dimensional feature space to a lower dimension. There are a number of feature extraction methods to reduce the dimensionality of text documents (Agarwal and Mittal, 2014). For example, in Li and Park (2007) a Singular Value Decomposition (SVD) based method was used to learn and represent relations among large numbers of words and natural text documents including these words. This method did not take into account the semantic relationship between the terms, resulting in rather poor performance. In Kolenda et al. (2000) an Independent Component Analysis (ICA) was employed to find k components that effectively contain maximum variability of the original data. This method transforms the original high dimensional data into lower dimensional components that are maximally independent from each other. In another research, Linear Discriminant Analysis (LDA) was used to transform the original high dimensional text data into a lower dimension (Wang and Qian, 2008).

Compared to feature extraction methods, there are varieties of text feature selection methods in the literature, each being filter, wrapper, hybrid or embedded methods. Filter methods require a statistical analysis on a feature set without utilizing any learning algorithm, and are the prime choice in many cases due to much lower computational complexity than others. Filter methods can be implemented as univariate or multivariate fashions (Hu et al., 2015). Many univariate methods have been proposed in the literature. Examples include Document frequency (DF) (Liu et al., 2005), Term variance (TV) (Liu et al., 2005), Term strength (TS) (Yang, 1995), Information gain (IG) (Liu et al., 2005), Chi-square (CHI) (Li et al., 2008), Odds ratio (OR) (Mengle and Goharian, 2009), Gini index (GI) (Shang et al., 2007), Improved Gini index (GINI) (Mengle and Goharian, 2009), distinguishing feature selector (DFS) (Yang, 1995), bi-normal separation (BNS) (Forman, 2003), mutual information (Xu et al., 2007) and relative discrimination criterion (RDC) (Rehman et al., 2015). In DF the number of documents containing a specific term is considered in its evaluation process. In TV it is assumed that features with higher variance values contain valuable information (Liu et al., 2005). TS measures a term's importance based on how commonly the term is likely to appear in similar documents (Yang, 1995). The OR method evaluates the ratio of odds occurring in positive classes to its odds in negative classes (Mengle and Goharian, 2009). In DFS the contributions of terms to the class discrimination is first estimated using a probabilistic approach, and then certain importance scores are assigned to them (Yang, 1995). In BNS the occurrence of a given term in each document is modeled as a normal distribution and its corresponding area under the curve that exceeds a threshold value is considered as the importance of that term (Forman, 2003). In DP the degree of deviation from the Poisson distribution is used to evaluate the importance of the terms (Ogura et al., 2009). On the other hand, to avoid the effects of unbalanced classes, the GINI method considers the term's condition probability and combines the posterior probability and conditional probability in its evaluation process (Shang et al., 2007). All these methods are univariate methods that do not consider the dependency between features, and thus are unable to remove redundant features. While, in multivariate methods the dependencies between