Expert Systems with Applications 41 (2014) 5938–5947

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Text classification using genetic algorithm oriented latent semantic features

Alper Kursat Uysal *, Serkan Gunal

Department of Computer Engineering, Anadolu University, Eskisehir, Turkiye

ARTICLE INFO

Keywords:
Feature selection
Genetic algorithm
Latent semantic indexing
Text classification

ABSTRACT

In this paper, genetic algorithm oriented latent semantic features (GALSF) are proposed to obtain better representation of documents in text classification. The proposed approach consists of feature selection and feature transformation stages. The first stage is carried out using the state-of-the-art filter-based methods. The second stage employs latent semantic indexing (LSI) empowered by genetic algorithm such that a better projection is attained using appropriate singular vectors, which are not limited to the ones corresponding to the largest singular values, unlike standard LSI approach. In this way, the singular vectors with small singular values may also be used for projection whereas the vectors with large singular values may be eliminated as well to obtain better discrimination. Experimental results demonstrate that GALSF outperforms both LSI and filter-based feature selection methods on benchmark datasets for various feature dimensions.

1. Introduction

The goal of text classification, or categorization, is to classify texts of interest into appropriate classes. Along with the increase in the number of electronic documents, text classification has received more attention to be able to organize these documents appropriately. A conventional text classification framework mainly consists of feature extraction, feature selection and classification stages.

Feature extraction stage simply extracts numerical information from raw text documents. For this purpose, most of the studies use bag-of-words technique (Joachims, 1997) to represent a document such that the order of terms within the document is ignored but frequencies of the terms are considered. Hence, each unique term in a document collection constitutes an individual feature. Consequently, a document is represented by a multi-dimensional feature vector, i.e. vector space model (Salton, Wong, & Yang, 1975). In a feature vector, each dimension corresponds to a weighted value (e.g., term frequency (TF), term frequency-inverse document frequency (TF-IDF) (Manning, Raghavan, & Schutze, 2008) of the regarding term within the document collection.

At the end of the feature extraction stage, hundreds or even thousands of features are obtained depending on the size of the document collection. Excessive numbers of features not only increase computational time but also degrade classification accuracy. Therefore, dealing with high dimensionality of the feature space is one of the most critical issues in text classification. Various feature selection methods are usually employed to overcome this issue. Feature selection methods can be divided mainly into three categories: filter, wrapper and embedded (Uysal & Gunal, 2012). Filters evaluate feature relevancies using a scoring scheme that is independent from any classifier (Guyon & Elisseeff, 2003). Filters are computationally fast; but, they usually do not consider feature dependencies. On the other hand, wrappers assess features using a classification and search algorithm (Gunal, Gerek, Ece, & Edizkan, 2009; Kohavi & John, 1997). Wrapper techniques take feature dependencies into consideration, offer interaction between feature subset search and choice of the classifier; however, they are much slower than the filters. Alternatively, embedded feature selection methods integrate feature selection into the training phase of classifier. Hence, these methods are specific to the utilized learning model just like the wrappers (Guyon & Elisseeff, 2003; Saeys, Inza, & Larranaga, 2007). While all these three methods can be applied separately (Guyon & Elisseeff, 2003; Montanes, Quevedo, & Diaz, 2003; Ogura, Amano, & Kondo, 2009; Uysal & Gunal, 2012; Yan, Zheng, Zhu, & Xiao, 2009; Yang & Pedersen, 1997), there also exist several studies combining the filters and wrappers (Gunal, 2012; Uguz, 2011).

As an alternative to feature selection, feature transformation approaches are also used to reduce feature dimension. However,

* Corresponding author.
E-mail addresses: akuysal@anadolu.edu.tr (A.K. Uysal), serkangunal@anadolu.edu.tr (S. Gunal).

http://dx.doi.org/10.1016/j.eswa.2014.03.041
0957-4174/

these approaches project the original feature space into a new lower-dimensional subspace rather than selecting from the original set of features. Although there exist many feature transformation methods, majority of the text classification studies prefer latent semantic indexing (LSI) due to its proven performance (Meng, Lin, & Yu, 2011; Thorleuchter & Van den Poel, 2013; Wang, Xu, Li, & Craswell, 2013; Wang & Yu, 2009; Yang, Sun, Sun, Cao, & Zheng, 2009; Yu, Xu, & Li, 2008; Zhang, Yoshida, & Tang, 2011). The underlying idea in LSI is to obtain the projection directions (i.e., singular vectors, eigenvectors, or principal components) providing the largest variations (i.e., largest singular values, or eigenvalues) based on singular value decomposition (SVD) or principal component analysis (PCA) so that feature dimension is greatly reduced while keeping the discriminative information (Gud & Shatovska, 2009).

While either feature selection or feature transformation methods can be individually used for dimension reduction, combinations of these methods are also possible. Moreover, these combinations may provide even better performance. As an example, a two-stage feature selection strategy consisting of various feature selection methods and LSI is proposed for text classification in (Meng et al., 2011). In this work, feature selection methods are initially applied to obtain a discriminative subset of the original feature set. Then, LSI is used to transform the subset into a further discriminative lower-dimensional set. Experimental results on two spam e-mail datasets demonstrate that this two-stage method performs better against the individual methods. In another example, information gain-based feature selection method and PCA is sequentially applied on multi-class text collections (Uguz, 2011). Yet again, the combination of feature selection and transformation further improves the classification performance.

Considering the feature transformation, there are also several efforts projecting the data in a different way than that of LSI or PCA. For instance, selection of the best subset of principal components among all rather than using those with the largest eigenvalues are found as an efficient method to determine the optimal multivariate regression model in (Barros & Rutledge, 1998). In a recent study, principal component selection based on a genetic algorithm is proposed for production performance estimation in mineral processing (Ding, Zhao, Liu, & Chai, 2014). As another example, a new framework that selects principal components efficiently is constructed in (Zheng, Lai, & Yuen, 2005) for face recognition task, and it is concluded that some smaller principal components are useful whereas some larger ones can be removed as well. Another transformation method, namely common vector approach (CVA), also states that the directions corresponding to the smallest eigenvalues rather than the largest ones may provide more discrimination (Gulmezoglu, Dzhafarov, & Barkana, 2001; Gunal & Edizkan, 2008).

Inspiring from the abovementioned approaches; in this paper, genetic algorithm oriented latent semantic features (GALSF) are proposed for text classification task. The proposed method consists of two stages, namely feature selection and feature transformation. The feature selection stage is carried out using the state-of-the-art filter-based methods. The feature transformation stage employs LSI empowered by genetic algorithm (GA) such that a better projection is attained using appropriate singular vectors, which are not limited to the ones corresponding to the largest singular values, unlike standard LSI approach. In this way, the singular vectors with small singular values may also be used for projection whereas the vectors with large singular values may be eliminated as well to obtain better discrimination. Effectiveness of the proposed method is comparatively evaluated against feature selection, and the combination of feature selection and transformation on two-class and multi-class text collections, namely Enron1, Ohsumed and Reuters-21578. For all collections, GALSF surpasses the other methods in terms of classification performance in almost all cases. Moreover, it is proven that the singular vectors providing better discrimination contain the ones corresponding not only to large but also small singular values rather than the largest singular values alone.

Rest of the paper is organized as follows: feature selection approaches used in the study are briefly described in Section 2. Section 3 explains LSI. Some fundamental concepts about genetic algorithms are provided in Section 4. Section 5 introduces the proposed method. Section 6 presents the experimental study and results. Finally, some concluding remarks are given in Section 7.

2. Feature selection

In this paper, two state-of-art filter methods are employed for the feature selection task. These are namely distinguishing feature selector (DFS) introduced by Uysal and Gunal (2012), and well-known chi square (CHI2) method (Yang & Pedersen, 1997). Mathematical backgrounds of these approaches are provided in the following subsections.

2.1. DFS

DFS selects distinctive features while eliminating uninformative ones considering the following term characteristics (Uysal & Gunal, 2012):

(i) A term frequently occurring in single class and not occurring in the other classes is discriminative.
(ii) A term rarely occurring in single class and not occurring in the other classes is irrelevant.
(iii) A term frequently occurring in all classes is irrelevant, too.
(iv) A term occurring in some of the classes is relatively discriminative.

DFS score of a term in a given text collection is simply computed as

$$\mathrm{DFS}(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\bar{t}|C_i) + P(t|\bar{C}_i) + 1}, \tag{1}$$

where M is the number of classes, $P(C_i|t)$ is the conditional probability of class C_i given presence of term t, $P(\bar{t}|C_i)$ is the conditional probability of absence of term t given class C_i, and $P(t|\bar{C}_i)$ is the conditional probability of term t given all the classes except C_i. Once DFS scores of all terms in the collection are obtained, the terms with the top scores are selected while the others are filtered out.

2.2. CHI2

In statistics, the CHI2 test is used to examine independence of two events (Uysal & Gunal, 2012). For the selection of text features, these two events correspond to occurrence of particular term and class, respectively. CHI2 information can be computed using

$$\mathrm{CHI2}(t, C) = \sum_{t\in\{0,1\}} \sum_{C\in\{0,1\}} \frac{(N_{t,C} - E_{t,C})^2}{E_{t,C}}, \tag{2}$$

where N is the observed frequency and E is the expected frequency for each state of term t and class C (Manning et al., 2008). CHI2 score of a term is calculated for individual classes. This score can be globalized over all classes in two ways. The first way is to compute the weighted average score for all classes while the second one is to choose the maximum score among all classes. In this work, the former approach is used as in

$$\mathrm{CHI2}(t) = \sum_{i=1}^{M} P(C_i).\mathrm{CHI2}\,(t, C_i), \tag{3}$$