Expert Systems With Applications 43 (2016) 82–92

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An improved global feature selection scheme for text classification

Alper Kursat Uysal*

Department of Computer Engineering, Anadolu University, Eskisehir, Turkiye

ARTICLE INFO

Keywords:
Global feature selection
Filter
Text classification
Pattern recognition

ABSTRACT

Feature selection is known as a good solution to the high dimensionality of the feature space and mostly preferred feature selection methods for text classification are filter-based ones. In a common filter-based feature selection scheme, unique scores are assigned to features depending on their discriminative power and these features are sorted in descending order according to the scores. Then, the last step is to add top-*N* features to the feature set where *N* is generally an empirically determined number. In this paper, an improved global feature selection scheme (IGFSS) where the last step in a common feature selection scheme is modified in order to obtain a more representative feature set is proposed. Although feature set constructed by a common feature selection scheme successfully represents some of the classes, a number of classes may not be even represented. Consequently, IGFSS aims to improve the classification performance of global feature selection methods by creating a feature set representing all classes almost equally. For this purpose, a local feature selection method is used in IGFSS to label features according to their discriminative power on classes and these labels are used while producing the feature sets. Experimental results on well-known benchmark datasets with various classifiers indicate that IGFSS improves the performance of classification in terms of two widely-known metrics namely Micro-F1 and Macro-F1.

1. Introduction

Rapid developments of internet technologies lead an increase on the amount of electronic documents worldwide. Consequently, hierarchical organization of these documents becomes a necessity. This situation enhances the importance of text classification whose goal is to classify texts into appropriate classes according to their contents. Text classification is applied to numerous domains such as topic detection (Rill, Reinel, Scheidt, & Zicari, 2014), spam e-mail filtering (Gunal, Ergin, Gulmezoglu, & Gerek, 2006; Idris & Selamat, 2014), SMS spam filtering (Uysal, Gunal, Ergin, & Gunal, 2013), author identification (Zhang, Wu, Niu, & Ding, 2014), web page classification (Saraç & Özel, 2014), and sentiment analysis (Medhat, Hassan, & Korashy, 2014). Text classification tasks can be realized with schemes having different settings. A fundamental text classification scheme, as in many different pattern recognition problems, consists of feature extraction and classification stages. Due to the nature of the problem, feature extraction mechanism needs to extract numerical information from raw text documents. Then, any classifier can be used to finalize the text classification process by predicting the label of documents. However, preprocessing (Uysal & Gunal, 2014) and feature selection (Uysal et al., 2013) are known as very important stages besides feature extraction and classification. Researchers in this field are still studying on enhancing the performance of text classification by incorporating various preprocessing (Dara, Dowling, Travers, Cooper, & Chapman, 2008; Uysal & Gunal, 2014), feature extraction (Vicient, Sánchez, & Moreno, 2013), feature selection (Uysal & Gunal, 2012; Wang, Liu, Feng, & Zhu, 2015), and classification (B. Yang, Zhang, & Li, 2011) methods.

Although there exist some recent studies about improving the feature extraction with the contribution of Wikipedia or similar resources, bag-of-words approach (Joachims, 1997) is the commonly used technique for feature extraction stage. In this approach, the orders of terms are neglected and text documents are represented with weighted frequencies (i.e., TF-IDF (Manning, Raghavan, & Schutze, 2008)) of the unique terms in the collection. As each unique term is used in the construction of the feature set, even a collection including small number of documents may be expressed with thousands of features. Excessive numbers of features may have negative effects on both classification accuracy and computational time. Therefore, most of the researchers concern with the feature selection stage in order to overcome these kinds of negative effects.

Feature selection techniques are generally categorized as filters, wrappers, and embedded methods. While wrappers and embedded methods require a frequent classifier interaction in their flow, filters do not need any classifier interaction during the construction of the feature set. Requirement of a classifier interaction may increase running time and make the feature selection method adapted

* Tel.: +90 2223213550.
E-mail address: akuysal@anadolu.edu.tr

http://dx.doi.org/10.1016/j.eswa.2015.08.050
0957-4174/

to a specific learning model. Due to these reasons, filter-based methods are preferred more compared to wrappers and embedded methods.

Filter-based methods can be divided into two categories referred as global and local depending on whether they assign a unique score or multiple class-based scores for any feature (Taşcı & Güngör, 2013). In the case of local feature selection methods, a globalization policy is necessary to convert the multiple local scores into a unique global score (Uysal & Gunal, 2012). On the other hand, in the case of global feature selection methods, the scores can be directly used for feature ranking. The features are ranked in descending order and top-N features are included in the feature set (Guyon & Elisseeff, 2003) where N is usually an empirically determined number. Some examples to global feature selection methods for text classification are document frequency (Yang & Pedersen, 1997), information gain (Lee & Lee, 2006), improved Gini index (Shang et al., 2007), and distinguishing feature selector (Uysal & Gunal, 2012). Another categorization about characteristics of filter-based feature selection methods is whether they are one-sided or two-sided (Ogura, Amano, & Kondo, 2011). In one-sided metrics, while features indicating membership to classes have a score greater than or equal to 0, features indicating non-membership to classes have a score smaller than 0. As features are ranked in descending order and the features having highest scores are included in the feature set, the negative features are not used in case there is no candidate positive feature. However, scores of two-sided methods are greater than or equal to 0. They implicitly combine positive and negative features which indicate the membership and non-membership to any class, respectively. In this case, considering one-against-all strategy in feature selection, positive features attain higher scores than negative ones. Thus, the negative features are rarely added to the feature set in two-sided metrics. Some examples to one-sided feature selection metrics for text classification are odds ratio (Zheng, Wu, & Srihari, 2004) and correlation coefficient (Ogura et al., 2011). In addition to the proposal of new metrics, feature selection studies for text classification proceed with improvement of current feature selection methods and developing ensemble approaches which combine various methods.

In the literature, there exist some studies dealing with integration of negative features in the feature set especially to handle the problems resulting from class imbalances. In previous studies, a local feature selection method which explicitly combines positive and negative features is proposed (Zheng & Srihari, 2003; Zheng et al., 2004). Experimental results on a single dataset show the efficiency of the proposed approach on imbalanced datasets. In a more recent study, the ability of selecting suitable negative features for some local feature selection methods is investigated on imbalanced datasets (Ogura, Amano, & Kondo, 2010). In another study, one-sided and two-sided feature selection metrics are compared for imbalanced text classification (Ogura et al., 2011). In one of the previous studies, a feature selection technique that automatically detects appropriate number of features containing both positive and negative features is proposed (Pietramala, Policicchio, & Rullo, 2012). The performance of the proposed approach which selects dynamic amount of features is compared with the performance of feature sets with some predetermined feature dimensions. The experiments show that the proposed approach succeeds in most of the experiments. Also, a comparison is carried out on two-sided feature selection metrics for text classification and an adaptive feature selection framework is proposed (Taşcı & Güngör, 2013). It is concluded that selecting different number of features for each class improves the performance of classification on imbalanced datasets. Apart from these, there exist some previous text classification studies dealing with combining the power of various feature selection methods. In a study, information gain method is separately combined with genetic algorithm and principal component analysis (Uguz, 2011), respectively. It is reported that both of these combination methods attains better performance than the individual performance of information gain. In a more recent study, several filter methods are combined with genetic algorithm (Gunal, 2012). The results indicate that this combination outperform the individual performances of the filter methods. In this study, contribution ratio of various feature selection metrics into the final feature set is also investigated. Besides, there exist some recent studies proposing solutions to determination of ideal number of features used for representation of documents automatically. As an example, a method that attempts to represent each document in the training set with at least one feature is proposed (Pinheiro, Cavalcanti, Correa, & Ren, 2012). It is stated that this approach obtains equivalent or better results than classical filter-based feature selection methods that attempts to determine the ideal number of features in a trial and error methodology. As another example to this kind of approaches, in a more recent study, representation of documents with more than one feature is proposed in order to improve the performance of classification (Pinheiro, Cavalcanti, & Ren, 2015). It is concluded that this approach performs better than or equal to the former one that each document is represented with only one feature. In addition, an improved feature selection scheme aiming to improve filter-based feature selection methods is proposed (J. Yang, Qu, & Liu, 2014). The main idea behind this study is to consider the imbalance factor of the training sets in the globalization process of class-based feature selection scores. It is reported that this improved scheme can significantly improve the performance of feature selection methods.

In spite of numerous approaches in the literature, feature selection for text classification is still an ongoing research topic. In this study, being inspired from some of the abovementioned studies, a new method namely improved global feature selection scheme (IGFSS), is proposed. IGFSS is a new approach which has some similarities with the characteristics of other approaches in the literature. These similarities can be listed as being a hybrid approach combining the power of two feature selection methods, benefiting from the power of negative features, and proposing a generic solution for all of the filter-based global feature selection methods. IGFSS aims to improve the classification performance of global feature selection methods by creating a feature set representing all classes nearly equally. For this purpose, a one-sided local feature selection method is integrated to the feature selection process besides a global feature selection method. Initially, the one-sided local feature selection method assigns a class label to each feature with a positive or negative membership degree. So, positive and negative features mentioned in the previous works are used as a part of the new method. Odds ratio was employed as one-sided local feature selection method during experiments. Instead of adding top-N features having highest global feature selection scores to the feature set, equal number of features representing each class equally with a certain membership and non-membership degree were included in the final feature set. In the experiments, an empirically determined negative feature ratio was used to represent each class with nearly same number of negative features. The experiments were carried out for different classification algorithms, datasets, and success measures. So, effectiveness of IGFSS was observed under different conditions. Results of the experimental analysis revealed that IGFSS offers better performance than the individual performance of global feature selection methods for all cases. In order to analyze classification performances, two common metrics for text classification was employed in the experiments.

Rest of the paper is organized as follows: feature selection methods used in this study are briefly described in Section 2. Section 3 introduces the details of IGFSS method. In Section 4, the classifiers used in the experiments are explained in details. Section 5 presents the experimental study and results which are related to accuracy, for each dataset, classifier, and success measure. Finally, some concluding remarks are given in Section 6.