Expert Systems With Applications 81 (2017) 268–281

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Variable Global Feature Selection Scheme for automatic classification of text documents

Deepak Agnihotri[a,*], Kesari Verma[a], Priyanka Tripathi[b]

[a] Department of Computer Applications, National Institute of Technology Raipur, C.G.-492010, India
[b] Deptartment of Computer Engineering and Applications, National Institute of Technical Teachers Training and Research Bhopal, M.P.-462002, India

ARTICLE INFO

Article history:
Received 6 January 2017
Revised 16 March 2017
Accepted 24 March 2017
Available online 27 March 2017

Keywords:
Feature selection
Text document classification
Text mining
Text analysis

ABSTRACT

The feature selection is important to speed up the process of Automatic Text Document Classification (ATDC). At present, the most common method for discriminating feature selection is based on Global Filter-based Feature Selection Scheme (GFSS). The GFSS assigns a score to each feature based on its discriminating power and selects the top-N features from the feature set, where N is an empirically determined number. As a result, it may be possible that the features of a few classes are discarded either partially or completely. The Improved Global Feature Selection Scheme (IGFSS) solves this issue by selecting an equal number of representative features from all the classes. However, it suffers in dealing with an unbalanced dataset having large number of classes. The distribution of features in these classes are highly variable. In this case, if an equal number of features are chosen from each class, it may exclude some important features from the class containing a higher number of features. To overcome this problem, we propose a novel Variable Global Feature Selection Scheme (VGFSS) to select a variable number of features from each class based on the distribution of terms in the classes. It ensures that, a minimum number of terms are selected from each class. The numerical results on benchmark datasets show the effectiveness of the proposed algorithm VGFSS over classical information science methods and IGFSS.

1. Introduction

A rapid increase in digital documents due to heavy use of Internet technologies and electronic devices necessitates efficient techniques to efficiently classify the digital documents. The documents are classified based on their contents and a substantial portion of the information on these digital documents is stored as text. The word (named term) is the smallest constituents of text and play a vital role in an Automatic Text Document Classification (ATDC) process (Manoj & Deepak, 2011; Uysal, 2016; Uysal & Gunal, 2012).

In the ATDC process, the first step is the feature extraction from entire corpus, further, preprocessing steps such as tokenization of text contents and removal of unnecessary tokens from the corpus is required. The processed tokens are used to build a vocabulary of the terms for the entire corpus. The vocabulary helps in the construction of a vector space. The occurring frequency of the terms in the documents is represented as a vector and subsequently forms a vector space of all the terms of this vocabulary. In this vector space, each individual term constitutes one dimension and for a typical document collection that requires ATDC, there may be millions of terms. Thus, an automated classification of documents using the resultant high dimensional vector space is challenging and needs an efficient technique to reduce the dimension of the vector space (Agnihotri, Verma, & Tripathi, 2014; 2016b; Manoj & Deepak, 2011; Uysal, 2016). Further, a considerable number of features of the vector space are not relevant to the text of the specific class and can be regarded as random noise features. Therefore, feature selection techniques are used in the second step of the ATDC process to reduce the high dimensionality by eliminating noise features under the premise of guaranteeing the higher performance of the classifiers (Agnihotri, Verma, & Tripathi, 2016b; Uysal & Gunal, 2012).

The feature selection techniques broadly fall into three categories: filters, wrappers, and embedded. The filter methods use an evaluation function to compute the score of a feature, thus depends solely on the properties of the data and independent of any classification algorithm. In contrast, wrapper methods use an inductive algorithm to estimate the value of a given subset. While wrappers and embedded methods require a frequent classifier interaction in their flow, the interaction of classifier is not required in the case of filters during the construction of the final feature set. The requirement of a classifier interaction may increase running

* Corresponding author.
E-mail addresses: dagnihotri.phd2012.mca@nitrr.ac.in (D. Agnihotri), kverma.mca@nitrr.ac.in (K. Verma), ptripathi@nitttrbpl.ac.in (P. Tripathi).

http://dx.doi.org/10.1016/j.eswa.2017.03.057
0957-4174/

time and force the feature selection method to adapt a specific learning model. Thus, filter-based methods are preferred more in comparison to wrappers and embedded methods (Luis, 1999; 2005; Uysal, 2016).

The Global Filter-based Feature Selection Scheme (GFSS) selects the top-N features from the feature set, where N is an empirically determined number. The GFSS assigns a score to each feature, based on their discriminating power and the top-N features are selected by arranging the features in descending order with the help of GFSS score (Guyon & Elisseeff, 2003; Uysal, 2016). The Filter-based methods are further subdivided into One Sided Local Filter-based Feature Selection Scheme (OLFSS) and Global filter-based feature selection scheme (GFSS). In the OLFSS, the local class-based score for each feature is computed and used as a final score. The GFSS follows a global policy and converts the multiple local scores into a global score to compute the final score of the features. The local and global scores can be directly used in feature ranking. The features are arranged in descending order and the top-N features are included in the Final Feature Set (FSS). The Information Gain (IG), Gini Index (GI), Distinguishing Feature Selector (DFS) and Gain Ratio (GR) etc. are known as the methods of GFSS, whereas Mutual Information (MI) and Odds Ratio (OR) as a method of OLFSS (Uysal, 2016). The selected discriminating features of FFS are used by the classifiers in the final step of ATDC process.

Although, the GFSS method improves the performance of the classifiers, but it has some limitations. The top-N features selected using the traditional filter based methods of GFSS, excludes some discriminating features of a few classes either partially or completely. There are two possibilities for the GFSS as further improvements: First, an equal number of features are selected from each class, and second, a variable number of features are selected from each class.

Uysal (2016) extended the work of Forman (2004) and proposed a solution named the Improved Global Feature Selection Scheme (IGFSS) to solve this issue. The IGFSS selects an equal number of representative features from each class in the final feature set. It is best suited with balanced and unbalanced dataset having less number of classes. In the balanced dataset, each class contains an equal number of documents along with sufficient number of feature terms. In case of unbalanced dataset, the distribution of terms in the documents of the class is much variable. However, it suffers in dealing with an unbalanced dataset having large number of classes. The distribution of features in these classes are highly variable. In this case, if an equal number of features are chosen from each class, it may exclude some important features from the class containing a higher number of features. It is due to the scheme of the IGFSS to select an equal number of features from each class. Sometimes this method introduces less important features in the final feature set if the class contains a much smaller number of features.

In order to overcome this limitation, this paper presents a novel Variable Global Feature Selection Scheme (VGFSS) to select the variable number of features from each class based on the variable distribution of terms in the classes and ensures that there are a minimum number of terms present for each class in the final feature set. The selection of variable number is determined mathematically. The two standard classifiers Linear Support Vector Machine (LSVM), and Softmax regression (SOFT MAX) classifiers are applied on five standard text data sets, viz. Webkb, Classic4, Reuters10, Trec2005 and Ohsumed10. The proposed VGFSS algorithm is evaluated using the benchmark Macro_F1 and Micro_F1 measures to demonstrate its effectiveness over state-of-the-art methods, OR+VGFSS and IGFSS algorithms. We have also implemented the VGFSS algorithms by assigning the class label of the features using OR method similar to IGFSS (named OR+VGFSS). The flow of VGFSS is similar to IGFSS, but in an improved way and based on the key points of IGFSS. It provides a generic solution to the filter based feature selection methods with less computational cost than IGFSS. The main contributions of this paper can be summarized as follows;

1. A variable number of features are selected from each class based on the distribution of the terms in those classes. It ensures that there are a minimum number of terms in the final feature set for each class.
2. The VGFSS doesn't compute positive and negative features using Odds Ratio (OR) for deciding the class label of the features. The class label of each feature is decided by the maximum local class-based score of the feature obtained using the standard methods (Forman, 2004). It reduces the overhead involved with IGFSS while deciding the class label of the features.

The rest of the paper is organized as follows. In Section 2, a brief overview of the feature selection methods and related work to this study is outlined. Section 3 introduces the details of the proposed VGFSS method. The Experimental setup, data analysis, classification algorithms, and performance evaluation measures are discussed in Section 4. Section 5 presents the experimental results and discussions. Finally, the paper is concluded in Section 6.

2. Related works

The substantial work has been carried out in the literature to select the most informative terms for ATDC. The Global filter-based feature selection schemes (GFSS) and the class-based One-sided Local Feature Selection Schemes are most widely discussed in the literature. The most popular criteria following these schemes are: Mutual Information (MI), Information Gain (IG), Distinguishing Feature Selector (DFS), Gini Index (GI)(Uysal, 2016), Gain Ratio, and Odds Ratio (OR).

Mutual information (MI) concept (Forman, 2003; Wang, Zhang, Liu, Lv, & Wang, 2014; Yang & Pedersen, 1997) is carried out from information theory to measure the dependencies between random variables and used to measure the information contained by a term t_i. It is strongly influenced by the marginal probabilities of the terms. It assigns higher weight to rare terms than common and sparse terms. Therefore, the term weight are not comparable to the terms with widely differing frequencies. The final score of term t_i is the maximum class-based score as shown in Eq. (1). The brief preliminary notations are shown in the Table 1.

$$MI(t_i) = \max_{j=1}^{j=r}\left[\log\left(\frac{p(t_i, C_j)}{p(t_i) \times p(C_j)}\right)\right] \tag{1}$$

Information Gain (IG) (Agnihotri, Verma, & Tripathi, 2016a; Forman, 2003; Uysal, 2016; Uysal & Gunal, 2012; Wang et al., 2014; Yang & Pedersen, 1997) assigns higher weight to common terms distributed in many categories than rare terms. The IG is also known as average Mutual Information. The computation of IG includes the estimation of the conditional probabilities of a category given a term and entropy. It is the difference between the original information requirement (i.e. based on the proportion of classes) and the new requirement (i.e., obtained after partitioning on term t_i) (see Eq. (2)).

$$\begin{aligned} IG(t_i) = {} & p(t_i) \times \sum_{j=1}^{j=r} p(C_j|t_i) \times \log p(C_j|t_i) \\ & + p(\bar{t}_i) \times \sum_{j=1}^{j=r} p(C_j|\bar{t}_i) \times \log p(C_j|\bar{t}_i) \\ & - \sum_{j=1}^{j=r} p(C_j) \times \log p(C_j) \end{aligned} \tag{2}$$