Pattern Recognition Letters 110 (2018) 23–29

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A probabilistic model derived term weighting scheme for text classification

Guozhong Feng[a,b,c], Shaoting Li[d], Tieli Sun[a], Bangzuo Zhang[a,*]

[a] Key Laboratory of Intelligent Information Processing of Jilin Universities, School of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China
[b] Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China
[c] Institute of Computational Biology, Northeast Normal University, Changchun 130117, China
[d] School of Statistics, Dongbei University of Finance and Economics, Dalian 116025, China

ARTICLE INFO

Article history:
Received 28 July 2016
Available online 6 March 2018

Keywords:
Latent feature selection indicator
Matching score function
Naive Bayes
Supervised term weighting
Text classification

ABSTRACT

Term weighting is known as a text presentation strategy to assign appropriate value to each term to improve the performance of text classification in the task of transforming the content of textual document into a vector in the term space. Supervised weighting methods using the information on the membership of training documents in predefined classes are naturally expected to provide better results than the unsupervised ones. In this paper, a new weighting scheme is proposed via a matching score function based on a probabilistic model. We introduce a latent variable to indicate whether a term contains text classification information or not, specify conjugate priors and exploit the conjugacy by integrating out the latent indicator and the parameters. Then the non-discriminating terms can be assigned weights close to 0. Experimental results using kNN and SVM classifiers illustrate the effectiveness of the proposed approach on both small and large text data sets.

1. Introduction

Text classification (TC) is the task of automatically classifying unlabeled electronic documents, such as news articles, advertisements, e-mails, call records and so forth, into a predefined set of classes. It is a supervised learning task because it use the information on the membership of training documents in predefined classes to build the classifier by machine learning techniques. Mathematical representation of the text is the primary step. This converts the document content to a high dimension vector, $D = (X_1, X_2, \ldots, X_p)$, where p is the size of the term (feature) set, and X_u is the number of times the term occurs in the document, $1 \le u \le p$. As the number of the documents increases, p increased rapidly, and always larger than ten thousands for a common text collection.

To improve the effectiveness of TC, the first step is to select the features (terms) to reduce the high dimension. Binary document representation is commonly used in feature selection methods, such as information gain, chi-square, etc. $X_u = 1$ represent that the term exists in the document (i.e., the document contains this term) and 0 otherwise.

Another important step is to assign appropriate weights to terms according to their different semantic contribution in a document [23]. Term frequency-inverse document frequency ($tf \times idf$) is a weighting scheme that reflects how important a word is to a document in information retrieval [13], and is widely used in text classification.

In TC, the information of the training documents in predefined classes is very effective and has been widely used for not only classifier building but also feature selection. Recently, research on term weighting methods using this information (named supervised term weighting methods) has gained more attention. The basic concept is that terms that occur differently among documents in different classes are given large weights. Hence, high performance feature selection metrics, information gain (*ig*), chi-square (*chi*), odds ratio (*or*), etc., can be used to replace the collection frequency factor in the weighting scheme [5,6]. Relevance frequency (*rf*), which considers the document frequency ratio, has been shown to provide better results than most information theory or statistical metric based weighting methods [17].

Some other approaches suggested to use probabilistic models, which have proven very effective in information retrieval and TC [14,16]. One approach is to weight terms based on their statistical

* Corresponding author.
E-mail addresses: fenggz264@nenu.edu.cn (G. Feng), zhangbz@nenu.edu.cn (B. Zhang).

https://doi.org/10.1016/j.patrec.2018.03.003
0167-8655/

confidence intervals [24]. It has also been proposed that the ratio and absolute difference of term occurrence can be used to improve the performance [1,7]. Another approach is to use terms weighted on the Kullback–Leibler (KL) divergence measure between pairs of class-conditional term probabilities, and Jensen–Shannon (JS) divergence for multi-class data [20]. Inverse category frequency has also been considered for term weighting schemes, which favors terms occurring in fewer classes rather than fewer documents [21,25].

An obvious drawback of these methods is that the non-discriminating terms cannot be handled properly. In this paper, we will seek some alternative method to overcome this difficulty. As the well performance of the probabilistic models in TC, this study focuses on the construction of a new supervised weighting scheme by building a probabilistic text classification model. To deal with the non-discriminating terms, a latent term selection indicator is introduced. Then, a matching score function is employed to evaluate how the features contribute to select positive samples from negative ones. Finally, we get a new supervised weighting scheme named latent relevance probability under the Bayesian statistic framework. Based on the formulation of the priori and the latent selective model information, non-discriminating terms can be assigned weights close to 0, term weighting and term selection can then organically bond in the proposed method.

The paper is organized as follows. Section 2 reviews the popular traditional unsupervised term weighting method, *idf*, and the state-of-art supervised method, *rf*. Section 3 presents our proposed probabilistic model specification for text classification and a latent term selection indicator is introduced to consider the term relevance uncertainty. A new supervised term weighting scheme is explained in Section 4. Section 5 introduces the experimental methodology, and shows the experiment results. We conclude the paper with a discussion in Section 6.

2. Related work

The most widely used term weighting method in text classification (TC) is $tf \times idf$, which was borrowed from information retrieval [14]. Given a term in a document, *tf* is the term frequency, and *idf* is

$$idf = \log \frac{n}{n_{1\bullet}},$$

where n is the number of documents in the training set, and $n_{1\bullet}$ is the number of documents that contain the term.

In information retrieval, *idf* can be regarded as a simple version of relevance weighting Robertson and Jones [22] when the relevance information is unavailable. In TC, the training documents have known predefined classes. Using this information, Lan et al. [17] proposed relevance frequency. Let n_{11} be the number of positive samples containing the term, and n_{10} be the number of negative samples containing the term. Then,

$$rf = \log\left(2 + \frac{n_{11}}{\max(1, n_{10})}\right). \tag{1}$$

The main concept of relevance weighting is that larger weight is assigned for terms with larger difference between positive and negative classes.

3. Model specification

Previous studies have shown that including document label information in the training set can help improve term weighting in TC. Terms that display more differently among the classes contribute more to the classification. This study used a probabilistic model to measure the difference between term displays, developed latent term (feature) selection to address non-discriminating terms, and derived a new term weighting scheme.

Table 1
Notations used in this paper.

Data	
$\mathcal{X}$	Document collection
$\mathcal{C}$	Corresponding label collection
Basic variables	
D	Document description, $X = (X_1, X_2, \ldots, X_p)$, $X_u = 1$ if the uth word occurs, and 0 otherwise
C	Document label
X	One considered term for convenience
δ	Binary latent selection indicator
Parameters	
θ_1	$\theta_1 = \Pr(X = 1 \mid C = 1)$
θ_0	$\theta_0 = \Pr(X = 1 \mid C = 0)$
η	$\eta = \Pr(X = 1)$
Others	
λ	$\lambda = \Pr(\delta = 1)$ is the priori term selection probability
α	$\alpha = \Pr(\delta = 1 \mid \mathcal{X}, \mathcal{C})$ is the latent selection index

3.1. Notations

Table 1 introduces the notations used throughout this paper.

3.2. Basic model

Without loss of generality, we consider binary TC. Multiple class classification is often transformed into a series of binary cases using the "One versus All" strategy.

For a document, $D = (X_1, X_2, \ldots, X_p)$, and its label, C, let $C = 1$ denote the positive class, and $C = 0$ the negative. Then,

$$\Pr(C = 1 \mid D) = \frac{\Pr(D \mid C = 1)\Pr(C = 1)}{\Pr(D)}. \tag{2}$$

To avoid further expansion of $\Pr(D)$, we use log probability rather than probability. Thus, it satisfies the classification task

$$\log \frac{\Pr(C = 1 \mid D)}{\Pr(C = 0 \mid D)} = \log \frac{\Pr(D \mid C = 1)\Pr(C = 1)}{\Pr(D \mid C = 0)\Pr(C = 0)}. \tag{3}$$

Ignoring the constant (prior class probability ratio), the classification task can be achieved by the matching score function [14],

$$MS(d) = \log \frac{\Pr(D \mid C = 1)}{\Pr(D \mid C = 0)} = \sum_{u=1}^{p} \log \frac{\Pr(X_u \mid C = 1)}{\Pr(X_u \mid C = 0)}, \tag{4}$$

where $X_u \in \{0, 1\}$. The second equality holds based on the NB [A5] assumption that the terms are conditionally independent, given the document label. The matching score is the summation of the log probability ratios, which are used to derive relevance weighting functions [22].

Under the NB assumption, term weights can be considered individually. Therefore, we consider one term, denoted as X for convenience. To avoid negative weights and obtain a symmetric scheme,

$$rp = \left| \log \frac{\Pr(X = 1 \mid C = 1)}{\Pr(X = 1 \mid C = 0)} \right|, \tag{5}$$

where $X = 1$ means that it occurs in the document. Following the *rf* method [17], we call this the relevance probability (*rp*).

When $\Pr(X = 1 \mid C = 1) = \Pr(X = 1 \mid C = 0)$, $rp = 0$, i.e., terms with equal probabilities in positive and negative classes have 0 weight. Large absolute difference leads to a large contribution to the classification task. However, because of data randomness, *rp* cannot be 0 even for these non-discriminating terms. To deal with this problem, we consider the statistical dependence between the terms and the document label in Section 3.3.