Pattern Recognition 60 (2016) 62–71

Contents lists available at ScienceDirect

Pattern Recognition

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

An efficient parameter estimation method for generalized Dirichlet priors in naïve Bayesian classifiers with multinomial models

Tzu-Tsung Wong*, Chao-Rui Liu

Institute of Information Management, National Cheng Kung University 1, Ta-Sheuh Road, Tainan City 701, Taiwan, ROC

ARTICLE INFO

Article history:
Received 13 May 2015
Received in revised form
4 March 2016
Accepted 29 April 2016
Available online 19 May 2016

Keywords:
Covariance matrix
Document classification
Generalized Dirichlet distribution
Multinomial model
Naïve Bayesian classifier

ABSTRACT

Generalized Dirichlet priors have been shown to be an effective way for improving the performance of naïve Bayesian classifiers with multinomial models, called multinomial naïve Bayesian classifiers, in document classification. For the sake of computational efficiency, a previous study divided distinct words into groups, and proposed a searching mechanism to determine the values of the parameters in a generalized Dirichlet prior group by group. That searching approach increases the computational cost of the multinomial naïve Bayesian classifier. In this paper, the covariance matrices for word groups are first calculated from available documents. A parameter estimation method and four strategies for choosing the value of a parameter corresponding to a word group are then proposed to solve for the parameters of the noninformative generalized Dirichlet priors for distinct words. The experimental results on two document sets show that the best strategy is to choose the largest value calculated from the statistics in a row, and that our parameter estimation method can efficiently solve for the parameters of generalized Dirichlet priors to significantly improve the performance of the multinomial naïve Bayesian classifier with respect to the searching approach.

1. Introduction

Many data stored in computers are represented as documents, such as web pages, emails, news, and books. Naïve Bayesian classifiers are a popular and efficient tool for document classification [1–3]. Wong [4] has shown that noninformative generalized Dirichlet priors can be an alternative to improve the performance of the naïve Bayesian classifier with multinomial models, called multinomial naïve Bayesian classifier, in classifying documents. The number of distinct words in a document set is generally more than ten thousand, and the dimension of a generalized Dirichlet prior is the number of distinct words minus one. This implies that estimating the occurrence probability of a distinct word for the multinomial naïve Bayesian classifier can be computationally intensive, as will be described in Section 2.2. This will definitely affect the applicability of generalized Dirichlet priors in the multinomial naïve Bayesian classifier. To accelerate the computation of the mean occurrence probability of a distinct word from a generalized Dirichlet distribution, the distinct words in a document set are divided into groups, and the distinct words in the same group are assumed to have the same normalized variance [4]. This arrangement can greatly reduce the computational cost of estimating the mean occurrence probabilities of distinct words.

Since the parameters of priors have impact on the performance of the multinomial naïve Bayesian classifier, Wong [4] proposed a searching mechanism to determine the best noninformative generalized Dirichlet priors that can achieve the highest prediction accuracy. That mechanism searches many possible parameter values for each word group, and the multinomial naïve Bayesian classifier must be executed to obtain a classification accuracy for evaluation whenever the value of a parameter is changed. The computational cost of the searching approach proposed by Wong [4] is therefore still high. Computational efficiency remains a critical advantage for training a multinomial naive Bayesian classifier, if the best parameters are computed from the documents rather than obtained through a search procedure.

A document belonging to a class value contains many distinct words that are considered as attributes, and the occurrence probability of every distinct word corresponds to a variable in a random vector. This means that for any given class value, the correlation between any pair of distinct words can be calculated from documents. It is possible to directly determine the parameters of generalized Dirichlet priors from those correlations. In this paper, we will propose an estimation method that solves for the parameters of noninformative generalized Dirichlet priors from the mean values and the statistics in the covariance matrices of word groups summarized from all available documents. Since

* Corresponding author.
E-mail address: tzutsung@mail.ncku.edu.tw (T.-T. Wong).

http://dx.doi.org/10.1016/j.patcog.2016.04.019
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

the candidate values for a parameter can be more than one, four strategies are introduced to determine the value of the parameter corresponding to a word group. The experimental results on two document sets will show that our methods not only is far more efficient than the parameter-searching method proposed by Wong [4], but also can achieve a significantly higher prediction accuracy for document classification.

This paper is organized as follows. Section 2 briefly introduces the operation of the multinomial naïve Bayesian classifier and the generalized Dirichlet distribution. The reason why distinct words must be divided into groups will be explained in this section. When the means and the covariance matrices of random variables are known, the method for solving the parameters of a generalized Dirichlet distribution is presented in Section 3. Section 4 introduces our parameter estimation method that is composed of three steps: feature extraction, word grouping, and parameter estimation. Our parameter estimation method is tested by two document sets in Section 5 to show that it can efficiently solve for the parameters of generalized Dirichlet priors to significantly improve the performance of the multinomial naïve Bayesian classifier. The conclusions and future directions for research are summarized in Section 6.

2. Multinomial naïve Bayesian classifiers and generalized Dirichlet distributions

Naïve Bayesian classifiers are a very efficient tool for document classification [5], and several techniques have thus been proposed to improve its prediction accuracy [6–8]. Two popular models for applying the naïve Bayesian classifier on document classification are the binomial model [9] and the multinomial model [10], and the later one adopted in this study is very competitive in both prediction accuracy and computational efficiency [10]. Wong [4] showed that noninformative generalized Dirichlet priors can be an effective way to improve the performance of the multinomial naïve Bayesian classifier. In this section, we will briefly review the operation of the multinomial naïve Bayesian classifier and introduce some basic properties of the generalized Dirichlet distribution. For the sake of clarity, Table 1 summarizes the notation used through this paper. Since parameter estimation for a class value is independent of that for the other class values, the subindex for a class value will be included in a symbol only when necessary.

2.1. Multinomial naïve Bayesian classifiers

Let $w_1, w_2, \ldots, w_\nu$ be the ν distinct words in a document set, and let x_i be the frequency of word w_i in document d. Then this document can be represented as $d=(x_1, x_2, \ldots, x_\nu)$, and let $|d|=x_1+x_2+\cdots+x_\nu$ be the number of words in document d. The naïve Bayesian classifier calculates the classification probability $p(c_j|d)$ for every possible class value c_j, and the one with the largest conditional probability will be the predicted class value of document d. When the multinomial model is used, every occurrence of a distinct word must be taken into account, while the order of words can be arbitrary. Thus, the classification probability can be rewritten as

Table 1
Notation used through this paper.

Symbol	Meaning
ν	The number of distinct words.
w_i	The ith distinct word.
x_i	The frequency of word w_i in document d.
c_j	The jth class value.
D_j	The document set for class value c_j.
θ_i	Occurrence probability of word w_i for class value c_j
α_i and β_i	Prior parameters corresponding to distinct word w_i for class value c_j.
G_m	The mth word group for class value c_j.
ϕ_m	Occurrence probability of the words in group G_m for class value c_j
λ_m and δ_m	Prior parameters corresponding to word group G_m for class value c_j.
Ψ_m	The set of the positive values resulting from the statistics in row m of the covariance matrix for class value c_j.
Q	The number of word groups for class value c_j.
$[s_{i,j}]$	Covariance matrix for word groups obtained from documents.
$[\sigma_{i,j}]$	Covariance matrix for word groups calculated from generalized Dirichlet prior.

$$p(c_j|d) \propto p(c_j)p(d|c_j) = p(c_j)\frac{|d|!}{\prod_{i=1}^{\nu} x_i!}\prod_{i=1}^{|d|} p(w_i|c_j) \propto p(c_j)\prod_{i=1}^{\nu} p(w_i|c_j)^{x_i}, \tag{1}$$

because $p(d)$ and the multinomial constant are the same for all class values. The value of $p(c_j)$ is generally estimated by the proportion of the documents with class value c_j in the training data. Laplace's estimate is often used to calculate $p(w_i|c_j)$ as $(N_{ij}+1)/(N_j+\nu)$ to ensure that this probability estimate is positive, where N_j and N_{ij} are the number of words and the total number of occurrences of distinct word w_i in the training documents with class value c_j, respectively.

Before observing any document with class value c_j, let θ_i represent the probability of word w_i occurring in a document belonging to this class value. Then $\Theta=(\theta_1, \theta_2, \ldots, \theta_{\nu-1})$ can be assumed to have a multivariate distribution, called a prior, defined on the unit simplex; i.e., all variables are nonnegative and the sum of all variables is less than or equal to one. For the training documents with class value c_j, let y_i be the total number of occurrences of distinct word w_i, and the training documents for this class value can be represented as $\mathbf{y}=(y_1, y_2, \ldots, y_\nu)$. Then $E(\theta_i|\mathbf{y})$ calculated from the posterior distribution of $\Theta|\mathbf{y}$ will be an estimate of $p(w_i|c_j)$ for the multinomial naïve Bayesian classifier. For the sake of ease of use, priors are assumed to be noninformative; i.e., all distinct words have the same mean probability. When the Laplace estimate with constant $\tau>0$ is used to ensure that every probability estimate is positive, $p(w_i|c_j)$ is calculated as $(y_i+\tau)/(y_1+y_2+\cdots+y_\nu+\nu\tau)$.

2.2. Generalized Dirichlet distributions

Definition 1. A random vector $\Theta=(\theta_1, \theta_2, \ldots, \theta_k)$ has a k-variate generalized Dirichlet distribution with parameters $\alpha_i>0$ and $\beta_i>0$ for $i=1, 2, \ldots, k$ if it has density

$$f(\Theta) = \prod_{i=1}^{k}\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}\theta_i^{\alpha_i-1}(1-\theta_1-\cdots-\theta_i)^{\gamma_i} \tag{2}$$

for $\theta_1+\theta_2+\cdots+\theta_k \le 1$ and $\theta_i \ge 0$ for $i=1, 2, \ldots, k$, where $\gamma_i=\beta_i-\alpha_{i+1}-\beta_{i+1}$ for $i=1, 2, \ldots, k-1$ and $\gamma_k=\beta_k-1$. This distribution will be denoted $\mathrm{GD}_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k)$.

The generalized Dirichlet distribution is first derived by Connor and Mosimann [11], and Wong [12] established several properties of this distribution in Bayesian analysis. Based on the general moment function derived by Wong [12], the expected value of θ_i is calculated as

$$E(\theta_i) = \frac{\alpha_i}{\alpha_i+\beta_i}\prod_{m=1}^{i-1}\frac{\beta_m}{\alpha_m+\beta_m} \tag{3}$$

for $i=1, 2, \ldots, k$ and

$$E(\theta_{k+1}) = \prod_{m=1}^{k}\frac{\beta_m}{\alpha_m+\beta_m}. \tag{4}$$

Let random vector $\Theta=(\theta_1, \theta_2, \ldots, \theta_{\nu-1})$ represent the probabilities of distinct words w_1 through $w_{\nu-1}$ occurring in a document with class value c_j, and let the training documents for this class value be represented as $\mathbf{y}=(y_1, y_2, \ldots, y_\nu)$. When Θ follows