J. Vis. Commun. Image R. 21 (2010) 613–626

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Discrete visual features modeling via leave-one-out likelihood estimation and applications

Nizar Bouguila *, Mukti Nath Ghimire

Concordia Institute for Information Systems Engineering, Faculty of Engineering and Computer Science, Concordia University, Montreal, Que., Canada H3G 2W1

ARTICLE INFO

Article history:
Received 26 January 2009
Accepted 17 March 2010
Available online 13 April 2010

Keywords:
Discrete features
Finite mixture models
Multinomial
Dirichlet
Generalized Dirichlet
Leave-one-out likelihood
SVM
Generative/discriminative
Scene classification
Visual words

ABSTRACT

Discrete data are an important component in many image processing and computer vision applications. In this work we propose an unsupervised statistical approach to learn structures of this kind of data. The central ingredient in our model is the introduction of the generalized Dirichlet distribution as a prior to the multinomial. An estimation algorithm, based on leave-one-out likelihood and empirical Bayesian inference, for the parameters is developed. This estimation algorithm can be viewed as a hybrid expectation–maximization (EM) which alternates EM iterations with Newton–Raphson iterations using the Hessian matrix. We propose then the use of our model as a parametric basis for support vector machines within a hybrid generative/discriminative framework. In a series of experiments involving scene modeling and classification using visual words, and color texture modeling we show the efficiency of the proposed approach.

1. Introduction

Nowadays with the huge amount of digital data such as images and videos, an important and challenging problem is to develop approaches and models to automatically process, manage and categorize large collections of this kind of data. A recurring subject in machine learning and data mining, to reach this goal, is the separation of data into homogeneous clusters. This topic has been extensively studied and different approaches and algorithms have been proposed and applied to several problems such as image categorization and retrieval. In particular, statistical models are widely used and have major challenges namely the choice of appropriate model structure to capture the characteristics of the data. These models should be dedicated to the type of features that we extract in order to represent a given image or video in a way suitable for its automatic processing. Discrete features appear in many computer vision, image processing and pattern recognition applications [1–3]. Some examples are color histograms [4], co-occurrence matrices [5], correlograms [6], color coherent vectors [7], and the recently proposed keyblocks (i.e. visual keywords) as an analogy to dictionaries in the case of text documents [8–10]. One of the most used statistical approaches is finite mixture of distributions which have been long studied [11]. An important problem, in the case of finite mixture models, is the choice of the distribution, since an irrelevant choice may degrade the performance of the model. Different assumption have been made in the case of discrete data. The multinomial represents, however, the state-of-the-art distribution for discrete data modeling.

In spite of the popularity of the multinomial, recent researches have shown that it has some drawbacks such as considering that the events to model are independent [1,12–14]. Another important problem is the parameters estimation in the case of sparse data[1] (i.e. the estimation of the probabilities of rarely observed or unobserved occurrences) [16,17]. The severity of this problem, which lead generally to poor biased estimates, has been widely studied by the natural language processing community, but generally ignored by image processing and computer vision researchers.[2] Different smoothing techniques have been proposed to overcome these problems [19]. The most successful approach is the use of the Dirichlet distribution as a prior, reflecting a certain background knowledge, to the multinomial which results in a completely formal statistical

* Corresponding author.
E-mail addresses: bouguila@ciise.concordia.ca (N. Bouguila), m_ghimir@encs.concordia.ca (M.N. Ghimire).

[1] This problem is also known as the zero-frequency problem and arises when dealing with observations that never occurred in the training data [15].

[2] A main assumption generally considered in image processing and computer vision applications is the Gaussian distribution. This assumption, however, is not realistic when dealing with discrete data. Moreover, it is well-known that the normal assumption limits the ability to analyze rare events [18].

1047-3203/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2010.04.001

model [1,13]. Indeed, we have previously proposed a framework in which finite mixture of Dirichlet distributions was used as a prior to the multinomial and applied to different applications such as texture modeling and narrowing the semantic gap for content-based image summarization and retrieval [1,20,21]. Recently, we have noticed that even the Dirichlet has some problems such as its very restrictive negative covariance structure which makes its use as a prior in the case of positively correlated data inappropriate (see [22,23] for more details and discussions). These problems can be overcame by the consideration of the generalized Dirichlet distribution which is more general and offers more flexibility [22,23]. This specific choice, however, has its problems namely the estimation of the parameters, which appears to be a laborious task when we consider the maximum likelihood approach, as we will show in Section 3.

In this paper, we consider the use of generalized Dirichlet mixtures as prior to the multinomial to model and cluster discrete visual feature vectors in the case of some interesting image representation applications. We propose a novel approach to enhance the estimation and the learning of our statistical framework parameters. Our approach is based on the maximization of the leave-one-out (LOO) likelihood through a hybrid expectation maximization algorithm which alternates EM iterations with Newton–Raphson iterations using the Hessian matrix. The proposed model is also used to generate SVM kernels within a generative/dicriminative framework involving both mixture models and SVM in a way that combines their respective advantages in order to take into account the discrete nature of the data. Indeed, mixing generative and discriminative approaches has attracted a lot of attention and some theoretical studies have shown its several advantages such as providing lower test error than both generative and discriminative techniques [24]. Moreover, generative/discriminate approaches have been found to be useful in many practical applications [25]. Our experiments involve a number of interesting applications such as scene modeling and classification using visual words and color texture modeling.

The rest of the paper is organized as follows. In Section 2, we review the multinomial assumption and both Dirichlet and generalized Dirichlet distributions used as priors for smoothing purposes. Section 3 gives a new approach for the estimation and selection of multinomial generalized Dirichlet mixtures. In Section 4, we present a generative/discriminative framework based on our developed model and SVM. Section 5, details our experiments. Finally, Section 6 summarizes our contributions and outlines future directions.

2. The discrete statistical model

Let $\vec{X}_i = (X_{i1}, \dots, X_{iD_i}),\ i = 1, \dots, N$, be a discrete vector representing a given image, D_i is the number of visual features in the image, and each variable X_{id}, $d = 1, \dots, D_i$, takes on values on a V-size visual corpus (or dictionary) that is a finite set of discrete values. Then, a classic assumption is that $\vec{X}_i$ is generated by the following model:

$$p(\vec{X}_i \,|\, \vec{\pi}) = \prod_{d=1}^{D_i} \prod_{v=1}^{V} \pi_v^{\delta(X_{id}=v)} = \prod_{v=1}^{V} \pi_v^{f_{iv}} \tag{1}$$

where $\delta(X_{id} = v)$ is an indicator function, $\{f_{iv}\}$ are the frequencies of values v in $\vec{X}_i$ and represent the sufficient statistics, $\vec{\pi} = (\pi_1, \dots, \pi_V)$ is the parameter vector of a multinomial, $\sum_{v=1}^{V} \pi_v = 1$.

Recent machine learning researches[3] have shown, however, that the multinomial assumption as a naive Bayes approach has several drawbacks and suffers from the zero counts which create serious obstacles [1,12–14]. For instance, data sparseness problem makes the maximum likelihood (ML) approach to estimate the π_v parameters unreliable [30]. Indeed, it is easy to show that the ML estimate is simply

$$\hat{\pi}_v = \frac{f_{iv}}{\sum_{v=1}^{V} f_{iv}} \tag{2}$$

Moreover, it is clear that $\hat{\pi}_v$ is zero for any feature that does not appear in $\vec{X}_i$, since the probabilities are estimated by the fraction of times the feature occurs over the total number of opportunities. The unreliability of ML estimates can be generalized for features which appear rarely (i.e. with small frequency). In order to adjust the ML estimates, a widely used approach is to modify the sample counts by augmenting them with some chosen values (i.e. pseudo-counts) and a common choice is to add 1 to all frequencies[4]:

$$\hat{\pi}_v = \frac{1 + f_{iv}}{V + \sum_{v=1}^{V} f_{iv}} \tag{3}$$

This adjustment is actually a special case of another classic approach to prevent zero probabilities which is the consideration of a Dirichlet prior for $\vec{\pi}$:

$$p(\vec{\pi}|\vec{\alpha}) = \frac{\Gamma(\sum_{v=1}^{V} \alpha_v)}{\prod_{v=1}^{V} \Gamma(\alpha_v)} \prod_{v=1}^{V} \pi_v^{\alpha_v - 1}$$

where $\vec{\alpha} = (\alpha_1, \dots, \alpha_V)$. The Dirichlet distribution depends on V parameters $\alpha_1, \dots, \alpha_V$, which are all real and positive. The choice of the Dirichlet distribution is motivated by the fact that it is closed under multinomial sampling (i.e. the Dirichlet is a conjugate prior for the multinomial) [33]. Using the Dirichlet as a prior, we can show that [1]:

$$\hat{\pi}_v = \frac{\alpha_v + f_{iv}}{\sum_{v=1}^{V} \alpha_v + \sum_{v=1}^{V} f_{iv}} \tag{4}$$

where $\sum_{v=1}^{V} \alpha_v$ is generally called *equivalent sample size*, since it can be interpreted as augmenting the actual frequencies by $\sum_{v=1}^{V} \alpha_v$ virtual ones [34]. Note that the last equation is reduced to Eq. (3) when we consider a symmetric Dirichlet, as a prior, with parameters $\alpha_v = 1$, $v = 1, \dots, V$. In spite of its flexibility and the fact that it is conjugate to the multinomial which have led to its application in different learning approaches and techniques, the Dirichlet has a very restrictive negative covariance matrix which violates generally experimental observations [35–37]. Another restriction of the Dirichlet is that the variables with the same mean must have the same variance as shown in [38]. These problems can be handled by the consideration of a generalized Dirichlet as a prior [3]:

$$p(\vec{\pi}|\xi) = \prod_{v=1}^{V-1} \frac{1}{B(\alpha_v, \beta_v)} \pi_v^{\alpha_v - 1} \left(1 - \sum_{l=1}^{v} \pi_l\right)^{\gamma_v}$$

where $B(\alpha_v, \beta_v) = \frac{\Gamma(\alpha_v)\Gamma(\beta_v)}{\Gamma(\alpha_v + \beta_v)}$. The generalized Dirichlet depends on $2(V-1)$ parameters $\xi = (\alpha_1, \beta_1, \dots, \alpha_{V-1}, \beta_{V-1})$, which are all real and positive, and $\gamma_v = \beta_v - \alpha_{v+1} - \beta_{v+1}$ for $v = 1, \dots, V-2$ and $\gamma_{V-1} = \beta_{V-1} - 1$. Note that the generalized Dirichlet is reduced to a Dirichlet with parameters $(\alpha_1, \dots, \alpha_{V-1}, \alpha_V = \beta_{V-1})$ when $\beta_v = \alpha_{v+1} + \beta_{v+1}$. The particular choice of the generalized Dirichlet as a prior has several advantages which we have widely discussed in our previous works [22] such as its general covariance matrix and the fact that it is also conjugate to the multinomial. Using this prior, we can show that [22]:

[3] Note that the drawbacks underlying the multinomial assumption have been discussed a long time ago by statisticians (see [26–29], for instance).

[4] This choice is usually referred to as Jeffrey's estimate [31,32, p. 293] or Laplace smoothing [19].