Engineering Applications of Artificial Intelligence 49 (2016) 19–42

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A statistical framework for online learning using adjustable model selection criteria

Taoufik Bdiri [a], Nizar Bouguila [b,*], Djemel Ziou [c]

[a] Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada H3G 1T7
[b] The Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada H3G 1T7
[c] DI, Faculté des Sciences, Université de Sherbrooke, Sherbrooke, QC, Canada J1K 2R1

ARTICLE INFO

Article history:
Received 26 March 2014
Received in revised form
5 June 2015
Accepted 16 October 2015
Available online 11 December 2015

Keywords:
Mixture models
Generalized inverted Dirichlet
User perception
Model updating
Probabilistic metrics
Object classification

ABSTRACT

Model-based approaches have been for long an effective method to model data and classify it. Recently they have been used to model users interactions with a given system in order to satisfy their needs through adequate responses. The semantic gap between the system and the user perception for the data makes this modeling hard to be designed based on the features space only. Indeed the user intervention is somehow needed to inform the system how the data should be perceived according to some ontology and hierarchy when new data are introduced to the model. Such a task is challenging as the system should learn how to establish the update according to the user perception and representation of the data. In this work, we propose a new methodology to update a mixture model based on the generalized inverted Dirichlet distribution, that takes into account simultaneously user's perception and the dynamic nature of real-world data. Experiments on synthetic data as well as real data generated from a challenging application namely visual objects classification indicate that the proposed approach has merits and provides promising results.

1. Introduction

One of the most important success keys to design a system that is dedicated to serve human beings is to ensure that the system's responses are approaching the satisfaction of the real needs and intentions of its users. Designing such systems is a challenging task especially when their responses should change according to the users interactions that differ from one user to another, which discards the option of designing predefined/deterministic systems' behaviors (Lu et al., 2003). A typical example of those systems are the search engines and object recognition applications where a user is looking for an object of interest among a huge amount of data (Hu and Bagga, 2004). Naturally a given property of data/object can be perceived by a user as being “interesting” while it could be “meaningless” for another user, also, an object of interest can be a target for two different users who have different behavior patterns, which should lead the system to close results using different interactions patterns. During the recent years, and with the tremendous evolution of multimedia devices, these users interactions/behaviors can generate a huge amount of data e.g. via internet through the recommendation systems and users feedbacks (Boutemedjet and Ziou, 2010, 2008). Many researchers considered model-based approaches to analyze users-systems interactions, such as in Boutemedjet and Ziou (2008), and Mironica et al. (2013). Indeed, this family of approaches is based on a rigorous mathematical background, and has been used extensively in classification and users feedbacks modeling, in order to feed inference engines that are capable of generating knowledge which enables a given system to learn and model the users' needs. Still, a semantic gap between the system representation of data, and the user mental representation of the same data, is considerably affecting the quality of the responses of the systems that are not usually aware of the users intentions when representing the data in the features' space. Adopting a model-based approach to solve this problem, we have concluded in our previous work in Bdiri et al. (2014, 2013) that the system should not represent a given class via a single mixture mode but rather model it using different components in a flexible hierarchical way which confirms several previous studies (Tantrum et al., 2004;Krishnamachari and Abdel-Mottaleb, 1999a,b; Garcia and Nock, 2010; Zhao and Karypis, 2005; Malik et al., 2010).

In this work we treat the problem of new coming data and how it should be perceived by the system in order to be able to satisfy the users needs. This issue is crucial as discussed in different research works (see, for example, Chen et al., 2000; Kleinberg, 2002). Consider, for instance, that the system receives new coming data online that

* Corresponding author. Tel.: +1 5148482424; fax: +1 5148483171.
E-mail addresses: t_bdiri@encs.concordia.ca (T. Bdiri), nizar.bouguila@concordia.ca (N. Bouguila), djemel.ziou@usherbrooke.ca (D. Ziou).

http://dx.doi.org/10.1016/j.engappai.2015.10.011
0952-1976/

should update the model at the system level without loosing the flexibility and the degree of dissimilarity/similarity perceived by the users. Thus, the system should take an important decision whether new classes should be created to represent the new coming data or not. In unsupervised learning, many model selection criteria have been developed such as the minimum description length (MDL) (Rissanen, 1999) and minimum message length (MML) (Wallace, 2005) criteria. Also, the model selection problem was treated through a nonparametric Bayesian technique namely the Dirichlet Process (DP) by assuming that there is an infinite number of mixture components such as in Neal (2000), but as we have discussed in Bdiri et al. (2014, 2013) those approaches have a serious drawback as the semantic meaning of the mixture components in the selected model does not necessarily fit with a human comprehensible semantic or intentions. The system should learn how to model the data according to a certain degree of similarity/dissimilarity tolerance. For example, consider that we have a class representing red apples, and the new coming data is representing tomatoes, should the system consider the red apples and tomatoes as being one single class, or differ between them by creating a new class for the tomatoes. Naturally the decision (creating a new class or not) depends on the application itself and the purposes of the users. In this paper, we propose a new methodology that can control how the system should perceive new coming data over time. We propose a statistical framework based on the generalized inverted Dirichlet (GID) distribution (Bourouis et al., 2014) which is the generalization of the inverted Dirichlet (ID) distribution that we have considered in the past (Bdiri et al., 2014, 2013). It is noteworthy to mention that the proposed framework can consider any other distribution. However, we consider the GID distribution because of an interesting property that enables the transformation of its representation into a space where the features are independent and follow each an inverted Beta distribution. We propose in this work an algorithm to learn a mixture of GID, update it and create new components depending on the users perception. We perform different simulations on synthetic and real data in order to validate our methodology.

The rest of this paper is organized as follows; in Section 2 we introduce the GID mixture model. Then, we propose an estimation algorithm for its parameters by considering both batch and online settings in Section 3. In Section 4 we define some model selection criteria for choosing the adequate number of components of a GID mixture using an unsupervised approach. Then, in Section 5 we introduce a new updating scheme for a growing mixture whose number of components can increase according to users perceptions, we define some dissimilarity/similarity metrics, and we introduce the complete algorithm. We introduce our experimental results on synthetic and real data in Section 6, and we conclude this paper in Section 7.

2. Finite generalized inverted Dirichlet mixture model

Let us consider a set $\mathcal{Y}$ of N D-dimensional vectors, such that $\mathcal{Y}=(\vec{Y}_1,\vec{Y}_2,\ldots,\vec{Y}_N)$. Let M denote the number of different components forming a flat mixture model (McLachlan and Peel, 2000; Zhang and Kwok, 2006) at the system level. We assume that $\mathcal{Y}$ is controlled by a mixture of GID distributions such that the vectors follow a common probability density function $p(\vec{Y}_i|\Theta)$, where Θ is the set of its parameters. Let $Z=\{\vec{Z}_1,\vec{Z}_2,\ldots,\vec{Z}_N\}$ denote the missing group indicator, where $\vec{Z}_i=(z_{i1},z_{i2},\ldots,z_{iM})$ is the label of $\vec{Y}_i$, such that $z_{ij}\in\{0,1\}$, $\sum_{j=1}^{M} z_{ij}=1$ and z_{ij} is equal to one if $\vec{Y}_i$ belongs to class j and zero, otherwise. Then, the distribution of $\vec{Y}_i$ given the class label $\vec{Z}_i$ is:

$$p(\vec{Y}_i|\vec{Z}_i,\Theta)=\prod_{j=1}^{M} p(\vec{Y}_i|\theta_j)^{z_{ij}} \tag{1}$$

with $\Theta=\{\theta_1,\theta_2,\ldots,\theta_M\}$ and θ_j is the set of parameters of the class j. In practice, we define $p(\vec{Y}_i|\Theta)$ which can be obtained by marginalizing the complete likelihood $p(\vec{Y}_i,\vec{Z}_i|\Theta)$ over the hidden variables. We define the prior distribution of $\vec{Z}_i$ as follows:

$$p(\vec{Z}_i|\vec{\pi})=\prod_{j=1}^{M}\pi_j^{z_{ij}} \tag{2}$$

where $\vec{\pi}=(\pi_1,\ldots,\pi_M)$, $\pi_j>0$ and $\sum_{j=1}^{M}\pi_j=1$, then we have:

$$p(\vec{Y}_i,\vec{Z}_i|\Theta,\vec{\pi})=p(\vec{Y}_i|\vec{Z}_i,\Theta)P(\vec{Z}_i|\vec{\pi})=\prod_{j=1}^{M}(p(\vec{Y}_i|\theta_j)\pi_j)^{z_{ij}} \tag{3}$$

We proceed by the marginalization of Eq. (3) over the hidden variables, which gives us the mixture for a given vector $\vec{Y}_i$:

$$p(\vec{Y}_i|\Theta,\vec{\pi})=\sum_{j=1}^{M} p(\vec{Y}_i|\theta_j)\pi_j \tag{4}$$

The GID was introduced by Lingappaiah (1976) as follows:

$$p(\vec{Y}_i|\theta_j)=\prod_{l=1}^{D}\frac{\Gamma(\alpha_{jl}+\beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})}\frac{Y_{il}^{\alpha_{jl}-1}}{(1+\sum_{k=1}^{l}Y_{ik})^{\gamma_{jl}}} \tag{5}$$

By substituting Eq. (5) in Eq. (4), we obtain:

$$p(\mathcal{Y}|\Theta,\vec{\pi})=\prod_{i=1}^{N}\left(\sum_{j=1}^{M}\pi_j\prod_{l=1}^{D}\frac{\Gamma(\alpha_{jl}+\beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})}\frac{Y_{il}^{\alpha_{jl}-1}}{\left(1+\sum_{k=1}^{l}Y_{ik}\right)^{\gamma_{jl}}}\right) \tag{6}$$

where $\Theta=(\theta_1,\theta_2,\ldots,\theta_M)$, with $\theta_j=\{\theta_{j1},\theta_{j2},\ldots,\theta_{jD}\}$, where $\theta_{jl}=(\alpha_{jl},\beta_{jl})$, $j=1,\ldots,M$, $l=1,\ldots,D$. We define γ_{jl} such that $\gamma_{jl}=\beta_{jl}+\alpha_{jl}-\beta_{j(l+1)}$ for $l=1,\ldots,D$ with $\beta_{j(D+1)}=0$.

The purpose of this work is to establish a way to (1) estimate the GID mixture's parameters, (2) update these parameters when new data are introduced, (3) select the optimal number of classes to consider when new data are introduced, (4) consider an infinite possible number of classes when we have a stream of new data that can form new classes that were not considered at the beginning. These new classes can be set up by the user or the system.

3. Parameters estimation

The posterior probabilities are a key factor to cluster data and classify it in mixture model-based representations of data. Indeed maximizing the posterior probabilities is considered an intuitive rule to assign data to its appropriate clusters. The GID has an interesting property enabling the factorization of its posterior probability such that (Mashrgy et al., 2014)

$$p(j|\vec{Y}_i,\Theta,\vec{\pi})\propto p_j\prod_{l=1}^{D}p_{iBeta}(X_{il}|\theta_{jl}) \tag{7}$$

where we have set $X_{i1}=Y_{i1}$ and $X_{il}=\frac{Y_{il}}{1+\sum_{k=1}^{l-1}Y_{ik}}$ for $l>1$. $p_{iBeta}(X_{il}|\theta_{jl})$ is an inverted Beta distribution with parameters $\theta_{jl}=(\alpha_{jl},\beta_{jl})$, $l=1,\ldots,D$, such that:

$$p_{iBeta}(X_{il}|\theta_{jl})=\frac{\Gamma(\alpha_{jl}+\beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})}X_{il}^{\alpha_{jl}-1}(1+X_{il})^{-(\alpha_{jl}+\beta_{jl})} \tag{8}$$

Thus, the clustering structure underlying $\mathcal{Y}$ is the same as the one underlying $\mathcal{X}=\{\vec{X}_1,\vec{X}_2,\ldots,\vec{X}_i\}$, where $\vec{X}_i=\{X_{i1},X_{i2},\ldots,X_{iD}\}$, $i=1,\ldots,N$, governed by the following mixture model with conditionally independent features

$$p(\mathcal{X}|\Theta,\vec{\pi})=\prod_{i=1}^{N}\left(\sum_{j=1}^{M}\pi_j\prod_{l=1}^{D}p_{iBeta}(X_{il}|\theta_{jl})\right) \tag{9}$$

The estimation of the parameters in Eq. (6) is then equivalent to the estimation of the parameters in Eq. (9).