

Online Data Clustering Using Variational Learning of a Hierarchical Dirichlet Process Mixture of Dirichlet Distributions

Wentao Fan and Nizar Bouguila(✉)

Concordia Institute for Information Systems Engineering,
Concordia University, Montreal, QC, Canada
{wenta_fa,nizar.bouguila}@encs.concordia.ca

Abstract. This paper proposes an online clustering approach based on both hierarchical Dirichlet processes and Dirichlet distributions. The deployment of hierarchical Dirichlet processes allows to resolve difficulties related to model selection thanks to its nonparametric nature that arises in the face of unknown number of mixture components. The consideration of the Dirichlet distribution is justified by its high flexibility for non-Gaussian data modeling as shown in several previous works. The resulting statistical model is learned using variational Bayes and is evaluated via a challenging application namely images clustering. The obtained results show the merits of the proposed statistical framework.

Keywords: Mixture models · Dirichlet distribution · Variational inference · Hierarchical Dirichlet process · Online learning · Image clustering

1 Introduction

With the ubiquity of new information technology and media, the amount of multimedia data generated everyday has increased exponentially. Handling the resulting massive data sets is a difficult problem [19, 20, 24, 33]. Fortunately, advances in statistics and computing have made available several data modeling tools and approaches in many areas such as pattern recognition, computer vision, and data mining. Among these approaches finite mixture models play a crucial role and have become fundamental tools for data analysis [9]. The efficient adoption of finite mixture models, however, presents itself serious challenges related mainly to the important model selection problem (i.e. automatic determination of the model complexity without under- or over-fitting). Thus, much recent research has been directed at data modeling using infinite mixtures rather than finite ones. Indeed, as we can see from advances in the area of machine learning, Bayesian nonparametric approaches have been widely studied and adopted recently [26, 31]. This is especially true for Dirichlet process (DP) mixtures of distributions [10, 11, 19, 25].

DP mixtures of Gaussian distributions have been largely adopted in the past. In a previous work, however, we have shown that DP mixtures of Dirichlet distributions could be a better alternative especially in the case of non-Gaussian

data [4]. A DP mixture of Dirichlet distributions can be viewed as a learning machine which estimates a given probability density function as an infinite weighted sum of Dirichlet distributions. This learning machine has been shown to be effective in several data mining and computer vision applications and has been proposed as an alternative to overcome the drawbacks of finite Dirichlet mixture models [4]. In this paper, we go a step further by taking advantage of the flexibility that hierarchical Bayesian modeling offers via the development of a hierarchical DP process mixture of Dirichlet distributions. A hierarchical DP [32] is actually a dependency model for multiple Dirichlet processes. It has been shown to be an efficient nonparametric Bayesian approach to the problem of model-based clustering of grouped data with sharing clusters [8, 30]. It is an extension to the conventional DP with a Bayesian hierarchy where the base measure for a set of Dirichlet processes is itself distributed according to a DP. Learning technique for DP-based models are generally designed to be run over already observed collections of objects. In several real applications, however, the collection grows over time which makes the use of batch learning algorithms infeasible. In this case, we should consider online learning algorithms, which allow to update the model’s parameters each time new objects are observed, by maintaining high-quality inference for new introduced data [7]. We develop then an online variational algorithm for the learning of our hierarchical DP mixture of Dirichlet distributions model. The adoption of variational Bayesian inference [1] is motivated by the fact that it has been shown to be an efficient alternative to purely Bayesian inference in the case of several nonparametric Bayesian models [13] and especially in the case of Dirichlet mixture models [14].

The paper is organized as follows. In Sect. 2 we present our hierarchical non-parametric model. In Sect. 3, an online variational approach is developed for the learning of the proposed model. Section 4 outlines the experimental setup involving the challenging problem of images categorization and presents the obtained results. The paper is concluded in Sect. 5.

2 Hierarchical DP Mixture of Dirichlet Distributions

In this section, we start by briefly reviewing Dirichlet processes and then we present in details our hierarchical model.

2.1 Dirichlet Process

The DP is a stochastic process whose sample paths are probability measures with probability one [16, 21]. Given a random distribution G , it is distributed according to a DP if its marginals follow Dirichlet distributions. More specifically, let H be a distribution over some probability space Θ and γ be a positive real number, then G is a DP with the base distribution H and concentration parameter γ , denoted as $G \sim \text{DP}(\gamma, H)$, if

$$(G(A_1), \dots, G(A_t)) \sim \text{Dir}(\gamma H(A_1), \dots, \gamma H(A_t)) \quad (1)$$

where (A_1, \dots, A_t) is the set of the finite partitions of Θ , and $\text{Dir}(\gamma H(A_1), \dots, \gamma H(A_t))$ is a finite-dimensional Dirichlet distribution with parameters $(\gamma H(A_1), \dots, \gamma H(A_t))$.

2.2 Hierarchical DP Mixture Model of Dirichlet Distributions

Hierarchical Dirichlet Process. A hierarchical DP is a distribution over a set of random probability measures over a probability space Θ . Recently, it has been shown to be an effective framework for modeling grouped data where observations are organized into groups that are allowed to remain statistically linked [30, 32]. Assuming that we have a data set which is separated into M groups. A hierarchical DP involves an indexed set of DPs $\{G_j\}$, one of each group, that share a base distribution G_0 , which is itself distributed as a DP:

$$G_0 \sim \text{DP}(\gamma, H) \quad G_j \sim \text{DP}(\lambda, G_0) \quad \text{for each } j, j \in \{1, \dots, M\} \quad (2)$$

where j is an index for each group of data. A hierarchical Dirichlet process can be represented in a more intuitive and straightforward way using two stick-breaking constructions [18, 29] containing a base-level and a group-level construction. In the base-level construction, since the base distribution G_0 is distributed according to the Dirichlet process $\text{DP}(\gamma, H)$, it can be expressed using a stick-breaking representation as

$$\beta'_k \sim \text{Beta}(1, \gamma) \quad \alpha_k \sim H \quad \beta_k = \beta'_k \prod_{s=1}^{k-1} (1 - \beta'_s) \quad G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\alpha_k} \quad (3)$$

where $\{\alpha_k\}$ are independent random variables distributed according to H , and where δ_{α_k} is an atom at α_k . The variables $\{\beta_k\}$ are known as the stick-breaking weights that satisfy $\sum_{k=1}^{\infty} \beta_k = 1$, and are obtained by recursively breaking a unit length stick into an infinite number of pieces such that the size of each successive piece is proportional to the rest of the stick. It is noteworthy that since G_0 is discrete and has a stick-breaking representation as in Eq. (3) according to the property of DP, the atoms α_k are shared among all G_j and differ only in weights. In this work, we apply the stick-breaking representation [34] to construct each group-level DP G_j :

$$\pi'_{jt} \sim \text{Beta}(1, \lambda) \quad \varpi_{jt} \sim G_0 \quad \pi_{jt} = \pi'_{jt} \prod_{s=1}^{t-1} (1 - \pi'_{js}) \quad G_j = \sum_{t=1}^{\infty} \pi_{jt} \delta_{\varpi_{jt}} \quad (4)$$

where $\delta_{\varpi_{jt}}$ is a group-level atom at ϖ_{jt} , and where $\{\pi_{jt}\}$ are the stick-breaking weights which satisfy $\sum_{t=1}^{\infty} \pi_{jt} = 1$. Since ϖ_{jt} is distributed according to the base distribution G_0 , it takes on the value α_k with probability β_k . We may also represent this using a binary latent variable $\mathbf{C}_{jt} = (C_{jt1}, C_{jt2}, \dots)$ as an indicator variable, such that $C_{jtk} \in \{0, 1\}$, $C_{jtk} = 1$ if ϖ_{jt} maps to the base-level atom α_k which is indexed by k ; otherwise, $C_{jtk} = 0$. Accordingly, we have $\varpi_{jt} = \alpha_k^{C_{jtk}}$. Consequently, group-level atoms ϖ_{jt} do not need to be explicitly represented

which further simplifies the inference process as it shall be clearer in the next section. The indicator variable \mathbf{C}_{jt} is distributed according to β :

$$p(\mathbf{C}|\beta) = \prod_{j=1}^M \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \beta_k^{C_{jtk}} \quad (5)$$

Since β is a function of β' according to the stick-breaking construction of the Dirichlet process as shown in Eq. (3), $p(\mathbf{C})$ can then be represented in the following form

$$p(\mathbf{C}|\beta') = \prod_{j=1}^M \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} [\beta'_k \prod_{s=1}^{k-1} (1 - \beta'_s)]^{C_{jtk}} \quad (6)$$

The prior of β' is a Beta distribution according to Eq. (3):

$$p(\beta') = \prod_{k=1}^{\infty} \text{Beta}(1, \gamma_k) = \prod_{k=1}^{\infty} \gamma_k (1 - \beta'_k)^{\gamma_k - 1} \quad (7)$$

One significant application of hierarchical DP is its consideration as a non-parametric prior over the factors for grouped data. More specifically, let i indexes the observations within each group j , we assume that each variable θ_{ji} is a factor corresponding to an observation X_{ji} , and the factors $\theta_j = (\theta_{j1}, \theta_{j2}, \dots)$ are distributed according to G_j , for each j . Thus, we can have the likelihood in the following form

$$\theta_{ji}|G_j \sim G_j \quad X_{ji}|\theta_{ji} \sim F(\theta_{ji}) \quad (8)$$

where $F(\theta_{ji})$ denotes the distribution of the observation X_{ji} given θ_{ji} , the prior for the factors θ_{ji} is the base distribution H of G_0 . This setting forms the definition of a *hierarchical DP mixture model*, where each group is associated with a mixture component, and the components are shared among these mixture models due to the sharing of atoms α_k among all G_j . Moreover, since each factor θ_{ji} is distributed according to G_j , it takes the value ϖ_{jt} with probability π_{jt} . Next, we introduce a binary latent variable $\mathbf{Z}_{ji} = (Z_{ji1}, Z_{ji2}, \dots)$ as an indicator variable. That is, $Z_{jit} \in \{0, 1\}$, we have $Z_{jit} = 1$ if θ_{ji} is associated with component t and maps to the group-level atom ϖ_{jt} ; otherwise, $Z_{jit} = 0$. Thus, we have $\theta_{ji} = \varpi_{jt}^{Z_{jit}}$. Since ϖ_{jt} also maps to the base-level atom α_k , we then have $\theta_{ji} = \varpi_{jt}^{Z_{jit}} = \alpha_k^{C_{jtk} Z_{jit}}$. The indicator variable \mathbf{Z}_{ji} is distributed according to π as

$$p(\mathbf{Z}|\pi) = \prod_{j=1}^M \prod_{i=1}^N \prod_{t=1}^{\infty} \pi_{jt}^{Z_{jit}} \quad (9)$$

According to the stick-breaking construction of the Dirichlet process in Eq. (4), π is a function of π' . Then, we have

$$p(\mathbf{Z}|\pi') = \prod_{j=1}^M \prod_{i=1}^N \prod_{t=1}^{\infty} [\pi'_{jt} \prod_{s=1}^{t-1} (1 - \pi'_{js})]^{Z_{jit}} \quad (10)$$

As shown in Eq. (4), the prior distribution of π' is a Beta:

$$p(\pi') = \prod_{j=1}^M \prod_{t=1}^{\infty} \text{Beta}(1, \lambda_{jt}) = \prod_{j=1}^M \prod_{t=1}^{\infty} \lambda_{jt} (1 - \pi'_{jt})^{\lambda_{jt} - 1} \quad (11)$$

The Hierarchical Infinite Dirichlet Mixture Model. We focus on a specific form of hierarchical DP mixture model where each observation within a group is drawn from a mixture of Dirichlet distributions. Since DP mixture models are often considered as infinite mixture models, we refer to the proposed model as the hierarchical infinite Dirichlet mixture model. The consideration of Dirichlet mixtures is motivated by their superior performance in modeling proportional data (i.e. normalized histograms) that are naturally generated by many applications [3, 6, 14]. Although the Dirichlet distribution is a multivariate distribution which is often used as a conjugate prior to the multinomial distribution in Bayesian statistics, it will be considered as parent distribution to model the data directly in this work. Furthermore, since we adopt the hierarchical DP mixture model framework, the problem of determining the number of mixture components is avoided by assuming that there is a countably infinite number of components.

Now let us consider a data set \mathcal{X} containing N random vectors and separated into M groups. We suppose that each vector $\mathbf{X}_{ji} = (X_{ji1}, \dots, X_{jiD})$ is represented in a D -dimensional space and is drawn from a hierarchical infinite Dirichlet mixture model. Then, the corresponding likelihood function of the proposed model with latent variables can be written as

$$\begin{aligned} p(\mathcal{X}|\mathbf{Z}, \mathbf{C}, \boldsymbol{\alpha}) &= \prod_{j=1}^M \prod_{i=1}^N \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \text{Dir}(X_{jit}|\boldsymbol{\alpha}_k)^{Z_{jit}C_{jtk}} \\ &= \prod_{j=1}^M \prod_{i=1}^N \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \left[\frac{\Gamma(\sum_{l=1}^D \alpha_{kl})}{\prod_{l=1}^D \Gamma(\alpha_{kl})} \prod_{l=1}^D X_{jil}^{\alpha_{kl}-1} \right]^{Z_{jit}C_{jtk}} \end{aligned} \quad (12)$$

Next, we need to place a prior distribution over the parameter $\boldsymbol{\alpha}$. In our case, conjugate prior is preferred since it greatly simplifies the mathematics in the learning process. Since $\boldsymbol{\alpha}$ is positive and the formal conjugate prior for the Dirichlet distribution is intractable, a Gamma distribution $\mathcal{G}(\cdot)$ is adopted to approximate the conjugate prior with an assumption that the Dirichlet parameters are statistically independent [14]:

$$p(\boldsymbol{\alpha}) = \mathcal{G}(\boldsymbol{\alpha}|\mathbf{u}, \mathbf{v}) = \prod_{k=1}^{\infty} \prod_{l=1}^D \frac{v_{kl}^{u_{kl}}}{\Gamma(u_{kl})} \alpha_{kl}^{u_{kl}-1} e^{-v_{kl}\alpha_{kl}} \quad (13)$$

where \mathbf{u} and \mathbf{v} are positive hyperparameters.

3 Online Variational Model Learning

First, we propose a batch variational inference method for learning the proposed hierarchical infinite Dirichlet mixture model based on a natural gradient method. Then, an online extension is proposed to account for large-scale or streaming data. The consideration of Variational inference [1] is motivated by the excellent results that it has provided when applied to finite Dirichlet mixtures [14]. In order to simplify notations, in this section, we define $\Omega = (\mathbf{Z}, \Lambda)$ as the set of latent and unknown random variables where $\Lambda = (\mathbf{C}, \boldsymbol{\pi}', \boldsymbol{\beta}', \boldsymbol{\alpha})$.

3.1 Batch Variational Inference

The goal of variational inference is to find an appropriate approximation, in terms of Kullback-Leibler (KL) divergence, $q(\Omega)$ for the true posterior distribution $p(\Omega|\mathcal{X})$. This problem can be tackled by adopting a factorization assumption for restricting the form of $q(\Omega)$ which is known as *mean field theory* [1]. Moreover, we adopt a truncation technique proposed in [2] to truncate the variational approximations of base and group levels at K and T , such that

$$\beta'_K = 1, \quad \sum_{k=1}^K \beta_k = 1, \quad \beta_k = 0 \text{ when } k > K \quad (14)$$

$$\pi'_{jT} = 1, \quad \sum_{t=1}^T \pi_{jt} = 1, \quad \pi_{jt} = 0 \text{ when } t > T \quad (15)$$

Notice that the truncation levels K and T are variational parameters which can be freely initialized and will be optimized automatically during the learning process. By adopting the truncated stick-breaking representation and the factorization assumption, the approximated posterior distribution $q(\Omega)$ can be fully factorized into disjoint distributions as

$$q(\Omega) = q(\mathbf{Z})q(\mathbf{C})q(\boldsymbol{\pi}')q(\boldsymbol{\beta}')q(\boldsymbol{\alpha}) \quad (16)$$

The approach that we consider for deriving our optimization solutions is based on a gradient method [28] and that can be easily extended to online settings as we shall see in the next section. The idea of the gradient-based variational inference approach is that, since the model has conjugate priors, the functional form of the factors in the variational posterior distribution is known. Thus, the lower bound $\mathcal{L}(q)$ can be considered as a function of the parameters of these distributions by taking their general parametric forms. The optimization of variational factors is then obtained by maximizing the lower bound with respect to these parameters. In our case, the functional form for each variational factor is the same as its conjugate prior distribution, namely Discrete for \mathbf{Z} and \mathbf{C} , Beta for $\boldsymbol{\beta}'$ and $\boldsymbol{\pi}'$, and Gamma for $\boldsymbol{\alpha}$. Therefore, the parametric forms for these variational posterior distributions can be defined as the following

$$q(\mathbf{Z}) = \prod_{j=1}^M \prod_{i=1}^N \prod_{t=1}^T \rho_{jit}^{Z_{jit}} \quad q(\mathbf{C}) = \prod_{j=1}^M \prod_{t=1}^T \prod_{k=1}^K \vartheta_{jtk}^{C_{jtk}} \quad (17)$$

$$q(\boldsymbol{\pi}') = \prod_{j=1}^M \prod_{t=1}^T \text{Beta}(\pi'_{jt}|a_{jt}, b_{jt}) \quad q(\boldsymbol{\beta}') = \prod_{k=1}^K \text{Beta}(\beta'_k|g_k, h_k) \quad (18)$$

$$q(\boldsymbol{\alpha}) = \prod_{k=1}^K \prod_{l=1}^D \mathcal{G}(\alpha_{kl}|u_{kl}^*, v_{kl}^*) \quad (19)$$

By Maximizing the lower bound $\mathcal{L}(q)$, we obtain $\rho_{jit} = \frac{\exp(\tilde{\rho}_{jit})}{\sum_{f=1}^T \exp(\tilde{\rho}_{jif})}$, where

$$\tilde{\rho}_{jit} = \sum_{k=1}^K \langle C_{jtk} \rangle [\tilde{\mathcal{R}}_k + \sum_{l=1}^D (\bar{\alpha}_{kl} - 1) \ln X_{jil}] + \langle \ln \pi'_{jt} \rangle + \sum_{s=1}^{t-1} \langle \ln(1 - \pi'_{js}) \rangle \quad (20)$$

$$\tilde{\mathcal{R}}_k = \ln \frac{\Gamma(\sum_{l=1}^D \bar{\alpha}_{kl})}{\prod_{l=1}^D \Gamma(\bar{\alpha}_{kl})} + \sum_{l=1}^D \bar{\alpha}_{kl} [\Psi(\sum_{l=1}^D \bar{\alpha}_{kl}) - \Psi(\bar{\alpha}_{kl})] [\langle \ln \alpha_{kl} \rangle - \ln \bar{\alpha}_{kl}] \quad (21)$$

$$+ \frac{1}{2} \sum_{l=1}^D \bar{\alpha}_{kl}^2 [\Psi'(\sum_{l=1}^D \bar{\alpha}_{kl}) - \Psi'(\bar{\alpha}_{kl})] \langle (\ln \alpha_{kl} - \ln \bar{\alpha}_{kl})^2 \rangle$$

$$+ \frac{1}{2} \sum_{c=1}^D \sum_{\substack{d=1 \\ (d \neq c)}}^D \alpha_{kc} \alpha_{kd} \left[\Psi'(\sum_{l=1}^D \bar{\alpha}_{kl}) (\langle \ln \alpha_{kc} \rangle - \ln \bar{\alpha}_{kc}) (\langle \ln \alpha_{kd} \rangle - \ln \bar{\alpha}_{kd}) \right]$$

$$\vartheta_{jtk} = \frac{\exp(\tilde{\vartheta}_{jtk})}{\sum_{f=1}^K \exp(\tilde{\vartheta}_{jtf})} \quad (22)$$

$$\tilde{\vartheta}_{jtk} = \sum_{i=1}^N \langle Z_{jit} \rangle [\tilde{\mathcal{R}}_k + \sum_{l=1}^D (\bar{\alpha}_{kl} - 1) \ln X_{jil}] + \langle \ln \beta'_k \rangle + \sum_{s=1}^{k-1} \langle \ln(1 - \beta'_s) \rangle \quad (23)$$

$$a_{jt} = 1 + \sum_{i=1}^N \langle Z_{jit} \rangle, \quad b_{jt} = \lambda_{jt} + \sum_{i=1}^N \sum_{s=t+1}^T \langle Z_{jis} \rangle \quad (24)$$

$$g_k = 1 + \sum_{j=1}^K \sum_{t=1}^T \langle C_{jtk} \rangle, \quad h_k = \gamma_k + \sum_{j=1}^M \sum_{t=1}^T \sum_{m=k+1}^K \langle C_{jtm} \rangle \quad (25)$$

$$u_{kl}^* = u_{kl} + \sum_{j=1}^M \sum_{t=1}^T \langle C_{jtk} \rangle \sum_{i=1}^N \langle Z_{jit} \rangle \bar{\alpha}_{kl} [\Psi(\sum_{s=1}^D \bar{\alpha}_{ks}) - \Psi(\bar{\alpha}_{kl})]$$

$$+ \sum_{s \neq l}^D \bar{\alpha}_{ks} \Psi'(\sum_{s=1}^D \bar{\alpha}_{ks}) (\langle \ln \alpha_{ks} \rangle - \ln \bar{\alpha}_{ks}) \quad (26)$$

$$v_{kl}^* = v_{kl} - \sum_{j=1}^M \sum_{t=1}^T \langle C_{jtk} \rangle \sum_{i=1}^N \langle Z_{jit} \rangle \ln X_{jil} \quad (27)$$

where $\Psi(\cdot)$ is the digamma function. The expected values in the above formulas are defined as

$$\bar{\alpha}_{kl} = \frac{u_{kl}^*}{v_{kl}^*} \quad \langle Z_{jit} \rangle = \rho_{jit} \quad \langle C_{jtk} \rangle = \vartheta_{jtk} \quad \langle \ln \alpha_{kl} \rangle = \Psi(u_{kl}^*) - \ln v_{kl}^* \quad (28)$$

$$\langle \ln \pi'_{jt} \rangle = \Psi(a_{jt}) - \Psi(a_{jt} + b_{jt}) \quad \langle \ln(1 - \pi'_{jt}) \rangle = \Psi(b_{jt}) - \Psi(a_{jt} + b_{jt}) \quad (29)$$

$$\langle \ln \beta'_k \rangle = \Psi(g_k) - \Psi(g_k + h_k) \quad \langle \ln(1 - \beta'_k) \rangle = \Psi(h_k) - \Psi(g_k + h_k) \quad (30)$$

$$\langle (\ln \alpha_{kl} - \ln \bar{\alpha}_{kl})^2 \rangle = [\Psi(u_{kl}^*) - \ln v_{kl}^*]^2 + \Psi'(u_{kl}^*) \quad (31)$$

The batch variational inference for hierarchical infinite Dirichlet mixture model can be considered as an EM-like algorithm and is summarized in Algorithm 1.

Algorithm 1. Batch variational learning.

-
- 1: Choose the initial truncation levels K and T .
 - 2: Initialize the values for hyperparameters λ_{jt} , γ_k , u_{kl} and v_{kl} .
 - 3: Initialize the value of ρ_{jit} by K -Means algorithm.
 - 4: **repeat**
 - 5: *The variational E-step:*
 - 6: Estimate the expected values in Eqs. (28)–(31), use the current distributions over the model parameters.
 - 7: *The variational M-step:*
 - 8: Update the variational solutions for each factor using Eqs. (17)–(19) and the current values of the moments.
 - 9: **until** Convergence.
-

3.2 Online Variational Inference

Inspired from the online learning framework proposed in [28] and tested successfully in [34], we develop an online variational inference framework for learning our model. In contrast with batch learning algorithms, online algorithms are more efficient when dealing with large-scale or streaming data which are naturally present in many real-world applications. In our case, let r denotes the amount of observed data that we currently have. Then, the current lower bound for the observed data can be calculated by

$$\mathcal{L}^{(r)}(q) = \frac{N}{r} \sum_{i=1}^r \int q(\Lambda) d\Lambda \sum_{\mathbf{Z}_i} Q(\mathbf{Z}_i) \ln \left[\frac{p(\mathbf{X}_i, \mathbf{Z}_i | \Lambda)}{q(\mathbf{Z}_i)} \right] + \int q(\Lambda) \ln \left[\frac{p(\Lambda)}{q(\Lambda)} \right] d\Lambda \quad (32)$$

where $\Lambda = (\mathbf{C}, \boldsymbol{\pi}', \boldsymbol{\beta}', \boldsymbol{\alpha})$. The main idea of the online variational inference is to successively maximize the current variational lower bound as in Eq. (32) with respect to each variational factor. Consider that we have already observed a data set $\{\mathbf{X}_1, \dots, \mathbf{X}_{(r-1)}\}$. Then, after obtaining a new observation \mathbf{X}_r , we can maximize the current lower bound $\mathcal{L}^{(r)}(q)$ with respect to $q(\mathbf{Z}_r)$, while other variational factors remain fixed to $q^{(t-1)}(\mathbf{C})$, $q^{(r-1)}(\boldsymbol{\alpha})$, $q^{(r-1)}(\boldsymbol{\pi}')$ and $q^{(r-1)}(\boldsymbol{\beta}')$. Therefore, we can update the variational solution to $q(\mathbf{Z}_r)$ as

$$q(\mathbf{Z}_r) = \prod_{j=1}^M \prod_{t=1}^T \rho_{jtr}^{Z_{jtr}} \quad (33)$$

where $\rho_{jtr} = \frac{\exp(\tilde{\rho}_{jtr})}{\sum_{j'=1}^T \exp(\tilde{\rho}_{j'tr})}$, and $\tilde{\rho}_{jtr} = \sum_{k=1}^K \langle C_{jtk}^{(r-1)} \rangle [\tilde{\mathcal{R}}_k^{(r-1)} + \sum_{l=1}^D (\bar{\alpha}_{kl}^{(r-1)} - 1) \ln X_{jrl}] + \langle \ln \pi_{jt}^{(r-1)} \rangle + \sum_{s=1}^{t-1} \langle \ln(1 - \pi_{js}^{(r-1)}) \rangle$. In the following step, we maximize the current lower bound $\mathcal{L}^{(r)}(q)$ with respect to $q^{(r)}(\mathbf{C})$, while $q(\mathbf{Z}_r)$ is fixed and other variational factors remain at their $(r-1)$ th values. Thus, the variational factor $q^{(r)}(\mathbf{C})$ can be updated as

$$q^{(r)}(\mathbf{C}) = \prod_{j=1}^M \prod_{t=1}^T \prod_{k=1}^K (\vartheta_{jtk}^{(r)})^{C_{jtk}^{(r)}} \quad (34)$$

where the hyperparameter $\vartheta_{jtk}^{(r)}$ is defined by

$$\vartheta_{jtk}^{(r)} = \vartheta_{jtk}^{(r-1)} + \xi_r \Delta \vartheta_{jtk}^{(r)} \quad (35)$$

where ξ_r is the learning rate. In this work, we adopt a learning rate function introduced in [34], such that $\xi_r = (\eta_0 + r)^{-w}$, subject to the constraints $w \in (0.5, 1]$ and $\eta_0 \geq 0$. In Eq. (35), $\Delta \vartheta_{jtk}^{(r)}$ is the natural gradient of the hyperparameter $\vartheta_{jtk}^{(r)}$. The natural gradient of a hyperparameter is obtained by multiplying the gradient by the inverse of Riemannian metric, which cancels the coefficient matrix for the posterior parameter distribution. Thus, we can obtain the natural gradient $\Delta \vartheta_{jtk}^{(r)}$ as

$$\Delta \vartheta_{jtk}^{(r)} = \vartheta_{jtk}^{(r)} - \vartheta_{jtk}^{(r-1)} = \frac{\exp(\tilde{\vartheta}_{jtk}^{(r)})}{\sum_{f=1}^K \exp(\tilde{\vartheta}_{jtf}^{(r)})} - \vartheta_{jtk}^{(r-1)} \quad (36)$$

$$\tilde{\vartheta}_{jtk}^{(r)} = N \rho_{jtr} [\tilde{\mathcal{R}}_k^{(r-1)} + \sum_{l=1}^D (\bar{\alpha}_{kl}^{(r-1)} - 1) \ln X_{jrl}] + \langle \ln \beta_k'^{(r-1)} \rangle + \sum_{s=1}^{k-1} \langle \ln(1 - \beta_s'^{(r-1)}) \rangle \quad (37)$$

Next, the current lower bound $\mathcal{L}^{(r)}(q)$ is maximized with respect to $q^{(r)}(\boldsymbol{\pi}')$, $q^{(r)}(\boldsymbol{\beta}')$ and $q^{(r)}(\boldsymbol{\alpha})$:

$$q^{(r)}(\boldsymbol{\pi}') = \prod_{j=1}^M \prod_{t=1}^T \text{Beta}(\pi_{jt}'^{(r)} | a_{jt}^{(r)}, b_{jt}^{(r)}) \quad (38)$$

$$q^{(r)}(\boldsymbol{\beta}') = \prod_{k=1}^K \text{Beta}(\beta_k'^{(r)} | g_k^{(r)}, h_k^{(r)}) \quad q^{(r)}(\boldsymbol{\alpha}) = \prod_{k=1}^K \prod_{l=1}^D \mathcal{G}(\alpha_{kl}^{(r)} | u_{kl}^{*(t)}, v_{kl}^{*(t)}) \quad (39)$$

where the hyperparameters are given by

$$a_{jt}^{(r)} = a_{jt}^{(r-1)} + \xi_r \Delta a_{jt}^{(r)}, \quad b_{jt}^{(r)} = b_{jt}^{(r-1)} + \xi_r \Delta b_{jt}^{(r)} \quad (40)$$

$$g_k^{(r)} = g_k^{(r-1)} + \xi_r \Delta g_k^{(r)}, \quad h_k^{(r)} = h_k^{(r-1)} + \xi_r \Delta h_k^{(r)} \quad (41)$$

$$u_{kl}^{*(r)} = u_{kl}^{*(r-1)} + \xi_r \Delta u_{kl}^{*(r)}, \quad v_{kl}^{*(r)} = v_{kl}^{*(r-1)} + \xi_r \Delta v_{kl}^{*(r)} \quad (42)$$

The corresponding natural gradients can be calculated as

$$\Delta a_{jt}^{(r)} = 1 + N \rho_{jtr} - a_{jt}^{(r-1)} \quad \Delta b_{jt}^{(r)} = \lambda_{jt} + N \sum_{s=t+1}^T \rho_{jsr} - b_{jt}^{(r-1)} \quad (43)$$

$$\Delta g_k^{(r)} = 1 + \sum_{j=1}^M \sum_{t=1}^T \vartheta_{jtk}^{(r)} - g_k^{(r-1)} \quad \Delta h_k^{(r)} = \gamma_k + \sum_{j=1}^M \sum_{t=1}^T \sum_{m=k+1}^K \vartheta_{jtk}^{(r)} - h_k^{(r-1)} \quad (44)$$

$$\begin{aligned} \Delta u_{kl}^{*(t)} &= u_{kl} + N \sum_{j=1}^M \sum_{t=1}^T \vartheta_{jtk}^{(r)} \rho_{jtr} \bar{\alpha}_{kl}^{(r-1)} [\Psi(\sum_{s=1}^D \bar{\alpha}_{ks}^{(r-1)}) - \Psi(\bar{\alpha}_{kl}^{(r-1)})] \\ &+ \sum_{s \neq l}^D \bar{\alpha}_{ks}^{(r-1)} \Psi'(\sum_{s=1}^D \bar{\alpha}_{ks}^{(r-1)}) (\langle \ln \alpha_{ks}^{(r-1)} \rangle - \ln \bar{\alpha}_{ks}^{(r-1)})] - u_{kl}^{*(t-1)} \end{aligned} \quad (45)$$

$$\Delta v_{kl}^{(r)} = v_{kl} - N \sum_{j=1}^M \sum_{t=1}^T \vartheta_{jtk}^{(r)} \rho_{jtr} \ln X_{jrl} - v_{kl}^{(r-1)} \quad (46)$$

It is noteworthy that the hyperparameters of $q^{(r)}(\boldsymbol{\pi}')$, $q^{(r)}(\boldsymbol{\beta}')$ and $q^{(r)}(\boldsymbol{\alpha})$ can be updated in parallel. This online variational inference procedure is repeated until all the variational factors are updated with respect to the current arrived observation. The online variational inference for hierarchical infinite Dirichlet mixture model is summarized in Algorithm 2. The proposed online learning algorithm is much more computationally efficient than its batch counterpart. This is because the batch algorithm updates the variational factors by using the whole data set in each iteration, and thus its estimation quality is improved more slowly than in the case of the online one.

Algorithm 2. Online variational learning.

- 1: Choose the initial truncation levels K and T .
 - 2: Initialize the values for hyperparameters λ_{jt} , γ_k , u_{kl} and v_{kl} .
 - 3: **for** $r = 1 \rightarrow N$ **do**
 - 4: *The variational E-step:*
 - 5: Update the variational solution to $q(\mathbf{Z}_r)$ using Eq. (33).
 - 6: *The variational M-step:*
 - 7: Compute learning rate $\xi_r = (\eta_0 + r)^{-w}$.
 - 8: Calculate the natural gradient $\Delta \vartheta_{jtk}^{(r)}$ using Eq. (36).
 - 9: Update the variational factor $q^{(r)}(\mathbf{C})$ as shown in Eq. (34).
 - 10: Calculate the natural gradients of the remaining hyperparameters using Eqs. (43)–(46).
 - 11: Update variational factors $q^{(r)}(\boldsymbol{\pi}')$, $q^{(r)}(\boldsymbol{\beta}')$ and $q^{(r)}(\boldsymbol{\alpha})$ through Eqs. (38)–(39).
 - 12: Repeat the *E-* and *M-steps* until new data are observed.
 - 13: **end for**
-

4 Experimental Results: Online Images Categorization

4.1 Experimental Design

In this section, we evaluate the effectiveness of the proposed online hierarchical infinite Dirichlet mixture (referred to as *OnHIDM*) model through a challenging real-world application namely online images categorization. The tackled problem is a fundamental task in computer vision and has drawn significant attention during the last decade [12, 15, 17, 35]. This problem, however, remains challenging due to the difficulty of capturing the variability of appearance and shape of diverse objects belonging to the same class, while avoiding confusing objects from different classes [23]. In our experiments, we demonstrate the advantages of our *OnHIDM* model by comparing its performance with three other mixture models involving the batch hierarchical infinite Dirichlet mixture

(*BaHIDM*) model, the online hierarchical infinite Gaussian mixture (*OnHIGM*) model and the online finite Dirichlet mixture (*OnFDM*) model. To make a fair comparison, all of these models are learned using variational inference. It is noteworthy that our goals are mainly to demonstrate the advantages of using online variational inference learning framework over the batch one, and using hierarchical infinite mixture model over the finite one, as well as using Dirichlet over the Gaussian mixture. In our experiments, the testing data are supposed to arrive sequentially in an online manner except for the *BaHIDM* model. We initialize the base truncation level K to 50, and the group truncation level T to 15. The parameters w and η_0 of the learning rate are set to 0.65 and 64, respectively. The hyperparameters involved in our model are initialized as $(\lambda_{jt}, \gamma_k, u_{kl}, v_{kl}) = (0.05, 0.05, 0.1, 0.01)$. Our simulations have supported these specific choices.

4.2 Methodology and Results

We apply the proposed *OnHIDM* to the problem of online images clustering using the following methodology. First, 128-dimensional scale-invariant feature transform (SIFT) [22] descriptors¹ are extracted from each image using the Difference-of-Gaussians (DoG) interest point detectors and then normalized. Next, these features are modeled using the proposed approach. Specifically, each image \mathcal{I}_j is considered as a “group” and is therefore associated with a Dirichlet process mixture (infinite mixture) model G_j . Thus, each extracted SIFT feature vector X_{ji} from image \mathcal{I}_j is supposed to be drawn from an infinite mixture model G_j , in which mixture models can be viewed as a representation of “visual words”. A global vocabulary is constructed and is shared among all groups (images) through the introduction of the common global infinite mixture model G_0 . This setting matches the desired design of a hierarchical Dirichlet process mixture model. An important step in image categorization approaches with bag-of-visual words representation is the construction of a visual vocabulary. The majority of these approaches need to use a separate vector quantization algorithm (such as K -means) to build the visual dictionary, where the vocabulary size is normally manually selected. In our approach, the construction of the visual vocabulary is part of the hierarchical Dirichlet process mixture framework, and the size of the vocabulary (number of mixture components in the global level mixture model) can be automatically inferred from the data thanks to its Bayesian nonparametric nature. Since our goal is to determine automatically the category to which a testing image \mathcal{I}_j should be assigned, our hierarchical Dirichlet process mixture framework needs to be augmented by an indicator variable B_{jm} associated with each image (or group). B_{jm} means that \mathcal{I}_j is generated from category m and then is drawn from an other infinite mixture model which is truncated at level J . This means that we need to add a new hierarchy level to our hierarchical infinite mixture model with a sharing vocabulary among all image categories. In this

¹ Other state-of-the-art local visual descriptors may provide better results, however, this is not the focus of this work.

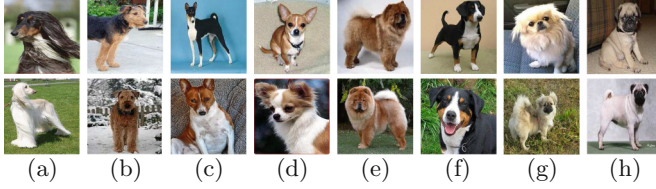


Fig. 1. Samples from the Dogs database. (a) Afghan hound, (b) Airedale, (c) Basenji, (d) Chihuahua, (e) Chow, (f) Entlebucher, (g) Pekinese, (h) Pug.

experiment, we truncate J to 20 and initialize the hyperparameter of the mixing probability of B_{jm} as 0.05. Finally, a testing image is affected to the category which has the highest posterior probability according to Bayes' decision rule.

Table 1. The average categorization accuracy rate (Acc) (%) obtained over 30 runs using different methods. The numbers in parenthesis are the standard deviation of the corresponding quantities.

| Method | <i>OnHIDM</i> | <i>BaHIDM</i> | <i>OnFDM</i> | <i>OnHIGM</i> |
|---------|---------------|---------------|--------------|---------------|
| Acc (%) | 80.87 (1.19) | 81.32 (1.02) | 76.18 (1.54) | 75.43 (1.31) |

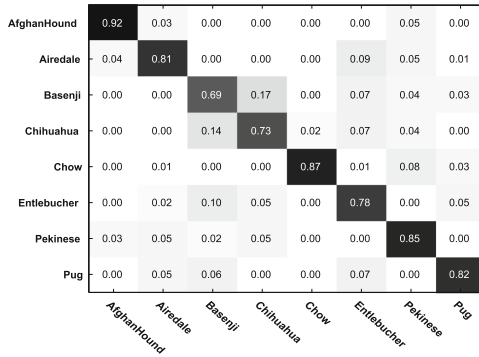


Fig. 2. Accuracy as a function of the number of images in the training set.

In our experiments, we consider a challenging public available database known as the Stanford Dogs database². This database contains 20,580 images of 120 breeds of dogs from around the world. The images are characterized by large scale, pose and light variations. The large intra-class and the small inter-class

² Database available at: <http://vision.stanford.edu/aditya86/ImageNetDogs>.

variabilities make this data set more challenging. In our experiments, we use a subset of this database consisting of 8 classes of dogs: Afghan hound (239 images), Airedale (202 images), Basenji (209 images), Chihuahua (152 images), Chow (196 images), Entlebucher (202 images), Pekinese (149 images) and Pug (200 images). Thus, we have 1,549 images in total. Sample images from each class are displayed in Fig. 1. We evaluated the categorization performance of the proposed algorithm by running it 30 times. We quantified the performance of our categorization approach using a confusion matrix as well as the rate of overall categorization accuracy. Each entry (i, j) of the confusion matrix denotes the percentage of images in category i that are assigned to category j . Figure 2 shows the confusion matrix computed by the proposed *OnHIDM* for our Dogs database. According to this matrix, the average categorization accuracy obtained by using *OnHIDM* was 80.87 % (error rate of 19.13 %). For comparison, we have also applied three other mixture-based approaches as mentioned earlier: *BaHIDM*, *OnHIGM* and *OnFDM*. The average performances of all tested approaches are given in Table 1. According to the results shown in this table, it is clear that the proposed *OnHIDM* and its batch counterpart (the *BaHIDM*) behave similarly (i.e., a Students t -test shows that the difference in performance between the *BaHIDM* and *OnHIDM* is not statistically significant: p -values between 0.1364 and 0.2237 for different runs) by providing better results than other two tested approaches. In this case, *OnHIDM* is a better choice over the *BaHIDM*, since *OnHIDM* is significantly faster, thanks to its online learning property, than the *BaHIDM*. According to our results, the *BaHIDM* required 2 h and 32 min to categorize all images while the *OnHIDM* only needed 47 min to do so on a computer with Intel’s Core i7 processor 2.00 GHz. Furthermore, the advantage of using a hierarchical infinite mixture model over a finite mixture model is clear by observing that better performance was obtained by *OnHIDM* (80.87 %) than by *OnFDM* (76.18 %) in terms of categorization accuracy rate. It is also worth mentioning that, as we can see from Table 1, the proposed *OnHIDM* (80.87 %) outperformed *OnHIGM* (75.43 %) which shows again the fact that the Dirichlet model has better modeling capability than the Gaussian for normalized data.

5 Conclusion

Nonparametric Bayesian models have been quite popular recently in many pattern recognition and computer vision problems due to their high accuracy and potential for data modeling. The success of these techniques rests largely on good choices of the distributions. This paper has presented and evaluated a hierarchical DP mixture model of Dirichlet distributions learned within a variational framework. The approach strives to achieve a high accuracy of online data clustering and has been validated through a challenging application namely images categorization. Further efficiency improvements are possible by performing several extensions such as introducing feature selection within the proposed model or considering Beta-Liouville distribution that has been shown to be a good alternative to the Dirichlet recently [5]. The consideration of the proposed model

with other learning approaches such as transfer learning [27] or its application to other challenging problems such as images annotation or objects recognition are interesting avenues for future research, also.

Acknowledgment. The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
2. Blei, D.M., Jordan, M.I.: Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1**, 121–144 (2005)
3. Bouguila, N., Wang, J.H., Hamza, A.B.: Software modules categorization through likelihood and bayesian analysis of finite dirichlet mixtures. *J. Appl. Stat.* **37**(2), 235–252 (2010)
4. Bouguila, N., Ziou, D.: A dirichlet process mixture of dirichlet distributions for classification and prediction. In: *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 297–302. IEEE (2008)
5. Bouguila, N.: Infinite liouville mixture models with application to text and texture categorization. *Pattern Recogn. Lett.* **33**(2), 103–110 (2012)
6. Bouguila, N., Ziou, D.: Using unsupervised learning of a finite dirichlet mixture model to improve pattern recognition applications. *Pattern Recogn. Lett.* **26**(12), 1916–1925 (2005)
7. Bouguila, N., Ziou, D.: Online clustering via finite mixtures of dirichlet and minimum message length. *Eng. Appl. Artif. Intell.* **19**(4), 371–379 (2006)
8. Boyd-Graber, J.L., Blei, D.M.: Syntactic topic models. In: *NIPS*, pp. 185–192. Curran Associates, Inc. (2008)
9. Bradley, P.S., Fayyad, U., Reina, C.A.: Clustering very large databases using em mixture models. In: *Proceedings of ICPR*, vol. 2, pp. 76–80. IEEE (2000)
10. Carbonetto, P., Kisynski, J., de Freitas, N., Poole, D.: Nonparametric bayesian logic. In: *Proceedings of UAI*, pp. 85–93 (2005)
11. Caron, F., Davy, M., Doucet, A., Duflos, E., Vanheeghe, P.: Bayesian inference for linear dynamic models with dirichlet process mixtures. *IEEE Trans. Sign. Proces.* **56**(1), 71–84 (2008)
12. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: *Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–12. Springer (2004)
13. Doshi, F., Miller, K., Gael, J.V., Teh, Y.W.: Variational inference for the indian buffet process. *J. Mach. Learn. Res. Proc. Track* **5**, 137–144 (2009)
14. Fan, W., Bouguila, N., Ziou, D.: Variational learning for finite Dirichlet mixture models and applications. *IEEE Trans. Neural Netw. Learn. Syst.* **23**(5), 762–774 (2012)
15. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.* **106**(1), 59–70 (2007)
16. Ferguson, T.S.: Bayesian density estimation by mixtures of normal distributions. *Recent Adv. Stat.* **24**, 287–302 (1983)

17. Frome, A., Singer, Y., Sha, F., Malik, J.: Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: Proceedings of ICCV, pp. 1–8. IEEE (2007)
18. Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.* **96**, 161–173 (2001)
19. Jin, L.C., Wan, W.G., Cui, B., Yu, X.Q.: A new multimedia classification approach: Bayesian of inductive cognition algorithm based on dirichlet process. *Imaging Sci. J.* **58**(6), 331–339 (2010)
20. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence and wordnet. In: Proceedings of the 13th ACM International Conference on Multimedia, pp. 706–715 (2005)
21. Korwar, R.M., Hollander, M.: Contributions to the theory of Dirichlet processes. *Ann. Probab.* **1**, 705–711 (1973)
22. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
23. Malisiewicz, T., Efros, A.A.: Improving spatial support for objects via multiple segmentations. In: Proceedings of BMVC, pp. 1–10 (2007)
24. Malisiewicz, T., Efros, A.A.: Recognition by association via learning per-exemplar distances. In: Proceedings of CVPR, pp. 1–8. IEEE (2008)
25. Nott, D.J.: Predictive performance of dirichlet process shrinkage methods in linear regression. *Comput. Stat. Data Anal.* **52**(7), 3658–3669 (2008)
26. Oppor, M., Winther, O.: Gaussian processes for classification: mean-field algorithms. *Neural Comput.* **12**(11), 2655–2684 (2000)
27. Quattoni, A., Collins, M., Darrell, T.: Transfer learning for image classification with sparse prototype representations. In: Proceedings of CVPR, pp. 1–8. IEEE (2008)
28. Sato, M.: Online model selection based on the variational Bayes. *Neural Comput.* **13**, 1649–1681 (2001)
29. Sethuraman, J.: A constructive definition of Dirichlet priors. *Stat. Sin.* **4**, 639–650 (1994)
30. Teh, Y.W., Jordan, M.I.: Hierarchical Bayesian nonparametric models with applications. In: Hjort, N., Holmes, C., Müller, P., Walker, S. (eds.) *Bayesian Nonparametrics: Principles and Practice*, pp. 158–207. Cambridge University Press (2010)
31. Teh, Y.W., Görür, D., Ghahramani, Z.: Stick-breaking construction for the indian buffet process. *J. Mach. Learn. Res. Proc. Track* **2**, 556–563 (2007)
32. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* **101**(476), 1566–1581 (2006)
33. Volkmer, T., Smith, J.R., Natsev, A.: A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. In: Proceedings of the 13th ACM International Conference on Multimedia, pp. 892–901 (2005)
34. Wang, C., Paisley, J.W., Blei, D.M.: Online variational inference for the hierarchical Dirichlet process. *J. Mach. Learn. Res. Proc. Track* **15**, 752–760 (2011)
35. Zhang, W., Yu, B., Zelinsky, G.J., Samaras, D.: Object class recognition using multiple layer boosting with heterogeneous features. In: Proceedings of the CVPR, pp. 323–330. IEEE (2005)

Database Systems for Advanced Applications
19th International Conference, DASFAA 2014,
International Workshops: BDMA, DaMEN, SIM³, UnCrowd;
Bali, Indonesia, April 21--24, 2014, Revised Selected
Papers

Han, W.-S.; Lee, M.L.; Muliantara, A.; Sanjaya, N.A.;

Thalheim, B.; Zhou, S. (Eds.)

2014, XXI, 430 p. 184 illus., Softcover

ISBN: 978-3-662-43983-8