Learning Global-to-Local Discrete Components with Nonparametric Bayesian Feature Construction

Min-Oh Heo
School of Computer Sci. & Eng.
Seoul National University
Seoul, 151-742, Korea
moheo@bi.snu.ac.kr

Sang-Woo Lee
School of Computer Sci. & Eng.
Seoul National University
Seoul, 151-742, Korea
slee@bi.snu.ac.kr

Jaeseon Lee
Field Robot R&D Group
Korea Institute of Industrial Technology
Ansan, 426-910, Korea
jslee@bi.snu.ac.kr

Byoung-Tak Zhang
School of Computer Sci. & Eng.
Seoul National University
Seoul, 151-744, Korea
btzhang@bi.snu.ac.kr

Abstract

Finding common latent components from data is an important step in many data mining applications. These latent variables are typically categorical and there are many sources of categorical variables, including dichotomous, nominal, ordinal, and cardinal values. Thus it is important to be able to represent the discrete components (categories) in a flexible way. Here we propose a nonparametric Bayesian approach to learning "plastic" discrete components by considering the uncertainty of the number of components with the Indian buffet processes (IBP). As observation models, we use the product of experts (PoE) to utilize sharper representation power and sparse over-completeness. We apply the proposed method to optical hand-written digit datasets and demonstrate its capability of finding flexible global-to-local components that can be used to describe and generate the observed digit images faithfully.

1 Introduction

Finding common latent components from data is an important step in many data mining applications. These latent variables are typically categorical and there are many sources of categorical variables, including dichotomous, nominal, ordinal, and cardinal values. So far, for discrete variables, latent Dirichlet allocation (LDA) [2] and a relatively small number of works (extensions [3], [4] from PCA, and NMF [1]) provide the methodologies for count data on discrete features (e.g., term frequency of language data). To deal with arbitrary categorical data, it is necessary to represent feature-value pairs on the components capable of expressing rule-like common patterns of variable feature-value pairs.

In this paper, we propose a discrete-value component learning method via expressing explicit feature membership on components from overall (global) patterns to local ones. For this, Indian Buffet Process (IBP) [15] is applied to represent feature-to-component relationships in a non-parametric Bayesian (NPB) way. This approach can automatically choose an unbounded number of components and has already been applied to several tasks ([12]-[14]).

In this method, we assume that data are generated from a small number of components where the components are linked with a probabilistic AND operation, which can connect them as a

construction process. Product of experts (PoE) [5] nicely demonstrates a probabilistic model for this situation by formulating component (or expert) values as a weighted *product* of themselves. PoE model is advantageous from several perspectives: 1) By multiplying marginal distributions of individual experts over and over, PoE model can make a sharp joint distribution. 2) PoE can be used for sparse over-complete learning of representation where the number of experts exceeds the number of features [6]. People extended conventional PoE and built models of dictionary using population coding, where the extended PoE is widely used in encoding-related research ([7]-[11]).

Additionally, from the point of view of constructive machine learning, the experts in PoE with the above properties perhaps can be seen as basic elements for construction of original data instances, and we can regenerate data via composition of them. In [11], data fragments are used as basic components for procedural construction into generated data.

This paper is organized as follows. Section 2 introduces the proposed model and the inference method. In section 3, the experimental results will be explained. Finally, section 4 concludes this paper.

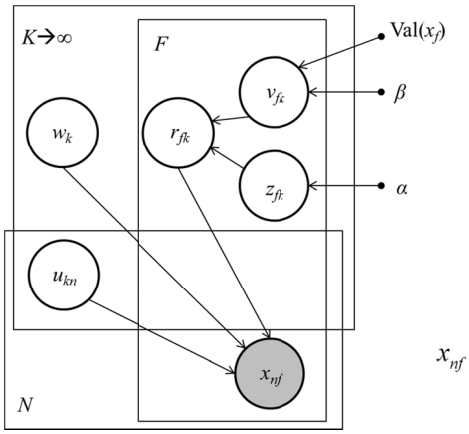
2 The Proposed Method

2.1 Nonparametric Bayesian Model for Learning Discrete Components

In [14], a nonparametric Bayesian feature construction method for IRL was introduced. The method was designed to learn reward functions in RL. We extend the model for PoE as shown in Figure 1.

At first, we assume that $\boldsymbol{X}$ is a data matrix (composed of *i.i.d* N instances and F features) generated from K components and their corresponding weights w. The prior distribution of each w is univariate Gaussian. $\boldsymbol{U}$ is a $K \times N$ random matrix to indicate the participation of components (component usage) to generate each instance. The prior for $\boldsymbol{U}$ is Bernoulli distribution. Also, $\boldsymbol{V}$ is a Dirichlet-categorical (or beta-Bernoulli) $F \times K$ random matrix to indicate the value on each component from the set of possible values of x_f, $val(x_f)$. $\boldsymbol{Z}$ is a $F \times K$ 0-1 binary random matrix to indicate that each feature is used for each component generated from the IBP. Note that K can be increased to infinity regarding to the property of dataset. $R = V \otimes Z$ is a sparse matrix to represent rule-like component description through Hadamard product (element-wise product) of value matrix $\boldsymbol{V}$ and selected feature matrix $\boldsymbol{Z}$. We can convert R_f and X_f (the f-th column in $\boldsymbol{R}$ and $\boldsymbol{X}$) into R_f^v and X_f^v using 1-of-k representation to express values explicitly.

The observation model of the proposed method is based on PoE. Observable variables in Figure 1 are shaded nodes x. In general case, we can use softmax function to represent categorical distribution for observed variables with the energy function [9], [10]:



$$v_{f,\cdot} \mid \beta \sim \begin{cases} Dirichlet - Categorical, & |Val(x_f)| > 2 \\ Beta - Bernoulli, & |Val(x_f)| = 2 \end{cases}$$

$$Z \mid \alpha \sim IBP(\alpha)$$

$$R = V \otimes Z$$

$$w_k \sim P(w_k)$$

$$u_{kn} \sim Bernoulli(0.5)$$

$$x_{nf} \mid R, W, U \sim \frac{\exp(\sum_k \sum_v r_{fk}^v \cdot w_k \cdot u_{kn} \cdot x_{nf}^v)}{\sum_{x_{nf} \in val(x_f)} \exp(\sum_k \sum_v r_{fk}^v \cdot w_k \cdot u_{kn} \cdot x_{nf}^v)}$$

Figure 1. A graphical model for learning discrete components and weights. Val(x_f) represent the set of possible values of x_f

$$E(X,R,W,U) = -\sum_{f=1}^{F}\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V_f} R_{fk}^{v} \cdot W_k \cdot U_{kn} \cdot X_{nf}^{v}$$

For simplicity, we use binary dataset where the values on $\boldsymbol{V}$ of features are -1 and 1. Then, the conditional distribution for one observed variable x_{nf} is as follows:

$$P(x_{nf} \mid R,W,U) = \frac{\exp(\sum_{k} r_{fk} \cdot w_k \cdot u_{kn} \cdot x_{nf})}{\sum_{x_{nf} \in val(x_f)} \exp(\sum_{k} r_{fk} \cdot w_k \cdot u_{kn} \cdot x_{nf})}$$

Also, the distribution can be simplified as follows:

$$P(x_{nf} = 1 \mid R,W,U) = \frac{1}{1+\exp(-2 \cdot \sum_{k} r_{fk} \cdot w_k \cdot u_{kn})}$$

2.2 MCMC Posterior Inference as a Learning Method

There is no general analytic posterior inference method for arbitrary probabilistic graphical models (PGM). So, approximate methods such as variational methods and Markov chain Monte Carlo (MCMC) approach are often used. As MCMC approaches, we use Gibbs sampling with a few Metropolis-Hastings (MH) updates similar to the method in [14], [16]. The posterior distribution over components R, the weights W and the component usages U is as

$$P(R,W,U \mid X,\alpha,\beta) \propto P(X \mid R,W,U)P(W)P(U)P(R \mid \alpha,\beta)$$

where $P(X \mid R,W,U)$ is the likelihood and $P(W)$ is the prior on the weights and $P(U)$ is the prior on the usage. α,β are hyperparameters to control the total number of components on IBP and value selection each.

Following the ancestral order, we use MH update for $\boldsymbol{V}$, $\boldsymbol{Z}$, $\boldsymbol{W}$ and $\boldsymbol{U}$ consecutively to infer the posterior. To do this, after random initialization we iteratively update values on each variable by sampling from the probability distribution conditioned on all of the other random variables. In the part of sampling Z using the IBP, we take 2-step procedure on each feature. Firstly, Gibbs sampling for $Z_{f,:}$ is performed for all components. After that, MH update for new components is sampled with Hastings acceptance ratio of new components involved in the model over the model without them. For $\boldsymbol{W}$, we use Gaussian distribution for proposing update weight and accept it with MH update.

To get the learned components and weights, we choose the sample with the maximum posterior.

3 Experimental Results and Discussion

In this section, we show the qualitative results from the experiments and discuss their meaning. For applications, each feature can be used to describe the special characteristics to represent the linkable parts to others considering relational, spatial or temporal features.

To show the concept of construction with components, we use simple binary image dataset. So, we binarize optical hand-written digit data [17] from UCI repository where the number of training data instances is 3823 and the size of each instances is 8×8 (Figure 2). The dataset contains approximately equal number of each digit 0 ~ 9. They are then vectorized to form instance rows in the data matrix $\boldsymbol{X}$ of which elements are 1 or -1.

Using the proposed model, we learn discrete components with the parameter α=6, β=(1,1). The components can have some selected features with 2 values {1, -1} and not-selected features, which are regarded as *don't care condition* getting from IBP naturally. Figure 2 shows one component example specifying the values and not-selected features with '×'.

The weights of local patterns have a tendency to increase very high. Rules in local patterns

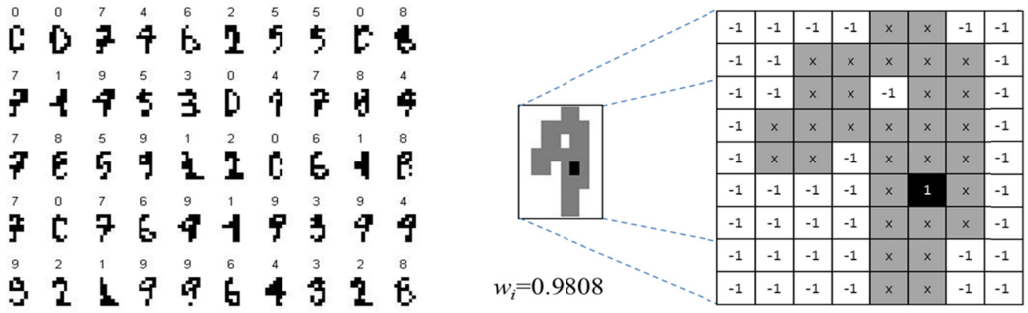


Figure 2. Optical hand-written digit dataset examples (left). 1 is assigned on the black, -1 is assigned on the white. One component example (right). Component specifies some features with values, but the other features are not specified (marked with '×'). One component has the corresponding fixed weight.

are so simple to be used averagely following the overall data statistics. On the other hand, the global patterns mostly are used in the special pattern of digits as Figure 3. Note that positive components specify instances and negative components filter them. Also, the size of component is so various that global components cover overall shape and local patterns described in small detail. The proposed model tries to build the balanced dictionary automatically to explain the dataset. Giving more sparsity, composition of global and local components can construct instances with arbitrary properties on each feature in a probabilistic manner.

Note that this approach for image modeling looks for the unified set of components which can have from 1 up to the number of all features for construction process, while mcRBM does not utilize more than 2 pixel dependencies [18]. And, Adams *et al*. studied NPB-based graphical model structure learning using cascading IBP [19]. It is different from our work that they do not consider the meaning of hidden variables and their interpretation as components.

4 Conclusion

This paper suggests nonparametric Bayesian approaches to extract global-to-local discrete components. While the assumption of exchangeability on features is relaxed, they provide automatically balanced dictionary for construction using the observation models with product of experts. This work is still ongoing to show the feasibility with experiments on sequential many-valued dataset: e.g. 1) music and 2) mobile behavior lifelogs with smartphone sensors. Also, based on [11], we will seek to utilize the idea of randomly segmented data fragment as initial component candidates as future works to expect quicker learning and apply incremental learning easily.

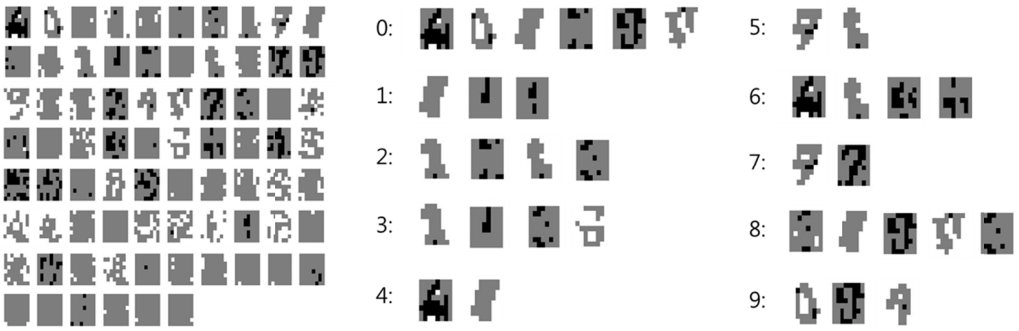


Figure 3. All components as learning results (left). Global and local components are extracted together. Each instance can be generated from the combination of these components probabilistically. The set of frequently used components on each digit (right). Some same components are used on similar digits: (0, 9), (2, 3) and more.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2010-0017734-Videome), supported in part by KEIT grant funded by the Korea government (MKE) (KEIT-10035348-mLife, KEIT-10044009, KEIT-10037352). The authors thank Jaesik Choi and Jaedeug Choi for helpful discussion.

References

[1] Lee, Daniel D. and Seung, H. Sebastian (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755):788-791.

[2] Blei, David M., Ng, Andrew Y., Jordan, Michael I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**(4–5):993–1022.

[3] Buntine, Wray and Jakulin, Aleks (2004) Applying Discrete PCA in Data Analysis. In Proceedings of the *Uncertainty in Artificial Intelligence* 2004, pp.59-66.

[4] Buntine, Wray and Jakulin, Aleks (2006) Discrete Component Analysis. In *Subspace, Latent structure and feature selection: statistical and optimization perspectives workshop*, pp.1-33.

[5] Hinton, Geoffrey E. (2002) Training products of experts by minimizing contrastive divergence. *Neural Computation* **14**(8):1771-1800.

[6] Teh, Y. W., Welling, M., Osindero, S., and Hinton, G. E. (2003) Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research* **4**:1235-1260.

[7] Hinton, Geoffrey E., Salakhutdinov, Ruslan R. (2006) Reducing the dimensionality of data with neural networks. Science **313**(5786):504-507

[8] Lee, Honglak, Grosse, Roger, Ranganath, Rajesh, Ng, Andrew Y (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the *International Conference on Machine Learning* 2009, pp. 609-616.

[9] Salakhutdinov, Ruslan, Mnih, Andriy and Hinton, Geoffrey (2007) Restricted Boltzmann machines for collaborative filtering. In Proceedings of the *International Conference on Machine Learning* 2007, pp. 791-798.

[10] Dahl, George E., Adams, Ryan P. and Larochelle, Hugo (2012) Training restricted Boltzmann machines on word observations. In Proceedings of the *International Conference on Machine Learning* 2012, pp. 791-798.

[11] Zhang, B.-T. (2008) Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory. *IEEE Computational Intelligence Magazine* **3**(3):49-63.

[12] Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological)* **59**:731–792.

[13] Ahmed, Amr, Ho, Qirong, Teo, Choon H., Eisenstein, Jacob, Smola, Alex J. and Xing, Eric P. (2011) Online Inference for the Infinite Topic-Cluster Model: Storylines from Streaming Text. In Proceedings of the *International Conference on Artificial Intelligence and Statistics 2011*, pp.101-109.

[14] Choi, J. and Kim, K.-E. (2013) Bayesian Nonparametric Feature Construction for Inverse Reinforcement Learning. In *International Joint Conference on Artificial Intelligence 2013*.

[15] Gharamani, Z., Griffiths, T. and Sollich, P. (2007) Bayesian Nonparametric Latent Feature Models. *Bayesian Statistics* **8**:1-25.

[16] Rai, Piyush and Daume III, Hal (2008) The infinite hierarchical factor regression model, In Proceedings of *Advances in Neural Information Processing Systems 2008*, pp. 1321-1328.

[17] Kaynak, C. (1995) Methods of Combining Multiple Classifiers and Their Applications to Handwritten Digit Recognition. MSc Thesis, Institute of Graduate Studies in Science and Engineering, Bogazici University.

[18] Ranzato, M. and Hinton, G. E. (2010) Modeling pixel means and covariance using factorized third-order Boltzmann machines, In CVPR 2010.

[19] Adams, Ryan P., Wallach, Hanna M, Ghahramani, Zoubin. (2009) Learning the Structure of Deep Sparse Graphical Models, http://arxiv.org/pdf/1001.0160.pdf