Learning Multi-faceted Activities from Heterogeneous Data with the Product Space Hierarchical Dirichlet Processes

Thanh-Binh Nguyen, Vu Nguyen, Svetha Venkatesh, and Dinh Phung

Centre for Pattern Recognition and Data Analytics, Deakin University, Australia
thanhbi@deakin.edu.au

Abstract. Hierarchical Dirichlet processes (HDP) was originally designed and experimented for a single data channel. In this paper we enhanced its ability to model heterogeneous data using a richer structure for the base measure being a product-space. The enhanced model, called Product Space HDP (PS-HDP), can (1) simultaneously model heterogeneous data from multiple sources in a Bayesian nonparametric framework and (2) discover multilevel latent structures from data to result in different types of topics/latent structures that can be explained jointly. We experimented with the MDC dataset, a large and real-world data collected from mobile phones. Our goal was to discover *identity–location–time* (a.k.a *who-where-when*) patterns at different levels (globally for all groups and locally for each group). We provided analysis on the activities and patterns learned from our model, visualized, compared and contrasted with the ground-truth to demonstrate the merit of the proposed framework. We further quantitatively evaluated and reported its performance using standard metrics including F1-score, NMI, RI, and purity. We also compared the performance of the PS-HDP model with those of popular existing clustering methods (including K-Means, NNMF, GMM, DP-Means, and AP). Lastly, we demonstrate the ability of the model in learning activities with missing data, a common problem encountered in pervasive and ubiquitous computing applications.

1 Introduction

Big data is providing us with data not only large in the amount but also multiform: text, image, graphic, video, speech, and so forth. Besides, data are often disrupted, irregular and disparate, resulting in missing data which is difficult to deal with. For example, while a post on Facebook may have texts and emotion tags but not a photo, another post may have texts, a video, but without any emotion tag. This diversity, heterogeneity and incompleteness of data cause trouble to data scientists as machine learning methods are typically designed to work with only one data type and/or not designed to work with missing elements.

Usually, useful patterns in heterogeneous data, such as daily life activities from mobile data, are not appearing in form of raw data, but have to be learned or inferred by exploiting the rich dependency between multiple channels of data.

Extracting these hidden patterns has been challenging the data mining and machine learning fields for the last few decades – a field known as latent variable modelling.

A recent tremendously successful latent model is the Latent Dirichlet Allocation (LDA) [1] and its Bayesian nonparametric version, the Hierarchical Dirichlet Processes (HDP) [22]. These are Bayesian topic models, which extract the latent patterns in form of probability distributions. However, although the statistical foundation is generic, these models were originally developed to deal with a single data channel, often to model words in document corpus. In this paper, we extend the machinery of HDP to extract richer, high-order latent patterns from heterogeneous data through the use of a richer base measure distribution being the product-space. This method allows heterogeneous data from multiples sources can be exploited simultaneously and take their correlated information into account of the learning process. We term our model the Product Space HDP (PS-HDP) model. Although the model has all advantages of hierarchical nonparametric modeling which automatically discover the space of latent patterns and activities from the data, its major strength lies in the product-space approach which can deal with multiple heterogeneous data sources and with missing elements.

We experiment our proposed PS-HDP model on the Nokia Mobile Data Challenge (MDC) dataset [9] to discover interesting *identity–location–time* (a.k.a *who-where-when*) patterns, which are known to be useful for mobile context-aware applications [21], or for human dynamics understanding [18]. First, we extract only the Bluetooth and WiFi-related data along with their timestamps from the MDC dataset. After that, we feed these data into our PS-HDP model to discover *identity–location–time* patterns at multiple levels including (1) global level (i.e. patterns occur across all participants) and (2) local level (i.e. patterns occurs for a specific participant). Those patterns are then visualized and analyzed against the ground truth to show that the model can automatically discover interesting patterns from data. Next, we evaluate quantitatively the performance of the proposed model over the standard metrics including F1-score, normalized mutual information (NMI), rand index (RI) and purity. These metrics are then compared with those from popular clustering methods, including K-Means, Non-negative Matrix Factorization (NNMF) [10], Gaussian Mixture Model (GMM) [13], DP-Means [8], and Affinity Propagation (AP) [6]. Finally, we demonstrate the ability of the PS-HDP model in dealing with missing data. The experimental results show that when more data is observed (i.e. less missing data), the F1-score and purity increase consistently.

In short, our main contributions in this paper include: (1) a novel Bayesian nonparametric model which can simultaneously model heterogeneous data from multiple sources to discover multilevel latent structures from data that can be explained jointly; (2) an application of learning hidden activities from heterogeneous data collected from multiple sensors, which is important in the area of pervasive and ubiquitous computing, using the proposed model on MDC dataset, a large and real world dataset collected from mobile phones.

2 Related Works

2.1 Pattern Discovery from Heterogeneous Data

Discovery of hidden patterns from heterogeneous data has been a challenge in machine learning and data mining. As machine learning algorithms are typically designed to work with only one specific data type (e.g. continuous, discrete), most of the previous works treat each data channel separately. For example, in the combined mining method [2], K different miners has been applied on K data sources to get K corresponding sets of patterns. After that, a merger is used to combine these sets to get global patterns (i.e. patterns from all data sources). Unfortunately, this approach is usually time-consuming, and more importantly, unable to exploit the correlating information between these data sources during the learning process to create better patterns.

Recently, researchers have been trying to leverage the advances of Bayesian nonparametric approach to propose unifying models to learn and discover patterns from multiple data sources. One common strategy is to model one data source as the primary data (called *content*), while treating other data sources as secondary data (called *contexts*). Contexts are viewed as distributions over some index space and both contents and contexts are modelled jointly. For example, Phung et al. [19] proposed an integration of the HDP and the nested Dirichlet process (nDP) with shared mixture components to jointly model contexts and contents respectively, where mixture components could be integrated out under a suitable parameterization. The authors showed that their model achieved good results in the field of computer vision. In another topic modelling work, Nguyen et al. [16] used documents as primary data, and used other information (e.g. time, authors) as contexts. In this model, secondary data channels should be collected in group-level, called group-level contexts [17]. More recently, Huynh et al. [7] have developed a full Bayesian nonparametric approach to model correlation structures among multiple and heterogeneous data sources. Choosing a data source as the primary data (*content*), they induce a mixture distribution over the data using HDP. Other data sources, also generated from HDP, are treated as *context(s)* and are assumed to be mutually independent given the *content*. However, in some applications, choosing one data source to be the primary data is not an easy task. To our best knowledge, this work is the most similar to our work. However, our approach differs from this one as it treats all data channels equally and hence does not require to specify *contexts* and *content*.

2.2 Discovery of Interaction and Mobility Patterns from Bluetooth and WiFi Data

Bluetooth data is widely used in the proximity detection problems. In particular, mobile Bluetooth data is usually used to detect surrounding people to discover interaction patterns of a person. Do et al. [3] used a large daily life Bluetooth data captured by smartphones to create a dynamic social network. Then, by using their proposed probabilistic model, they discovered different social contexts

(such as group meeting or dinner with family) and interaction types (e.g. office interaction, personal interaction). In another approach, Nguyen et al. [14] used HDP to discover interaction types from Bluetooth data from honest social signals captured by sociometric badges. Then, they clustered mixtures proportions from HDP using Affinity Propagation algorithm to extract contexts and communities.

While Bluetooth data is useful for discover surrounding people, WiFi data is usually used to infer locations. A smartphone is able to scan for surrounding WiFi hotspots. Each WiFi hotspot has a unique identified fingerprint (i.e. its MAC address). Assuming that the location of a WiFi hotspot is never changed, one can use its fingerprint as an indicator of the location where a person is at. From the raw WiFi scans, different approaches are used to discover locations. Dousse et al. [4] used the OPTICS clustering to group similar scans to a cluster representing a place. However, using OPTICS algorithm, the number of clusters must be provided beforehand. Nguyen et al. [15], in contrast, using the AP algorithm to discover interesting locations without the need of specifying the number of clusters. Furthermore, from interesting locations, they can also learn daily routines and mobility patterns of mobile phone users.

3 Framework

3.1 Hierarchical Dirichlet Processes (HDP)

Let J be the number of groups and $\left\{x_{j1}, \ldots, x_{jN_j}\right\}$ be N_j observations associated with group j which are assumed to be exchangeable within the group. Under HDP framework, each group j is endowed with a random group-specific mixture distribution G_j which is statistically connected with other mixture distributions via another Dirichlet Process (DP) sharing the same base probability measure G_0:

$$G_j \mid \alpha, G_0 \overset{\text{iid}}{\sim} \text{DP}\left(\alpha, G_0\right) \qquad (j = 1, \ldots, J) \tag{1}$$

$$G_0 \mid \gamma, H \sim \text{DP}\left(\gamma, H\right). \tag{2}$$

This generative process further indicates that G_j (s) are exchangeable at the group level and conditionally independent given the base measure G_0, which is also a random probability measure distributed according to another DP.

It is clear from the definition of the HDP that G_j's, G_0 and H share the same support Θ. Then the local atoms in group j is draw as $\theta_{ji} \overset{\text{iid}}{\sim} G_j$ and the observation is generated following $x_{ji} \sim F\left(\theta_{ji}\right)$.

We present the stick-breaking representation of HDP for posterior inference which can be summarized following. We draw a global mixing weight $\beta \sim \text{GEM}(\gamma)$, then generate the topics $\phi_k \overset{\text{iid}}{\sim} H(\lambda)$. The global atom G_0 in Eq. 2 can be characterized as $G_0 = \sum_{k=1}^{\infty} \delta_{\phi_k} \times \beta_m$. We next sample the mixing proportion for each document j such that $\pi_j \overset{\text{iid}}{\sim} \text{DP}\left(\alpha\beta\right)$. The local atom in each document is represented as $G_j = \sum_{k=1}^{\infty} \pi_{j,k} \times \delta_{\phi_k}$. Finally, we draw the latent assignment $z_{ji} \overset{\text{iid}}{\sim} \text{Mult}(\pi_j)$ and observation $x_{ji} \overset{\text{iid}}{\sim} F(\phi_{z_{ji}})$ accordingly.

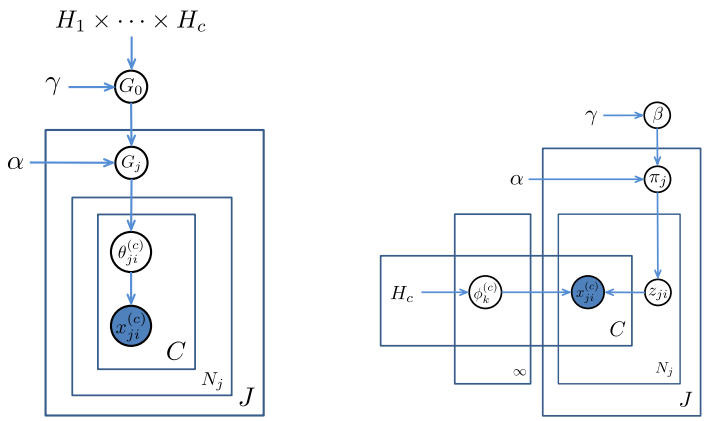
HDP is originally applied for nonparametric text modelling, but it is also applied in the field of natural language processing to detect how many grammar symbols exist in a particular set of sentences [11]. Another extension of HDP as Dynamic HDP [20] is for modelling time series documents. Evolutionary HDP (EvoHDP) [23] is for multiple correlated time-varying corpora by adding time dependencies to the adjacent epochs. In EvoHDP, each HDP is built for multiple corpora at each time epoch, and the time dependencies are incorporated into adjacent epochs under the Markovian assumption. Specifically, the dependency is formulated by mixing two distinct Dirichlet processes (DPs) (one is the DP model for the previous epoch, and the other is an updating DP model). The model inference is implemented by a cascaded Gibbs sampling scheme.

3.2 Product-space HDP (PS-HDP)

HDP is a powerful topic model to learn latent topics and patterns; however, although the statistical background is generic, it was originally designed and experimented for a single data channel. The Product Space HDP (PS-HDP) is an extension of the HDP model using a richer structure for the base measure being a product-space. It can be described as follow. The data consist of J groups, each of which contain N_j data points. Each data point i (in group j) is a collection of observations, denoted as $\left\{x_{ji}^1, x_{ji}^2, ..., x_{ji}^C\right\}$ from C data sources which can be heterogeneous. To have matters concrete, we assume a group is a user and a data point is an interaction between users. Each interaction includes three channels of information such that x_{ji}^1 is a timestamp (Gaussian distribution), x_{ji}^2 and x_{ji}^3 are a Bluetooth and a WiFi signal respectively (Multinomial distribution).

Stochastic Representation Let $H_1, H_2, ..., H_C$ be the base measure (for each data sources) generating the global atom G_0 from Dirichlet Process with concentration parameter γ as $G_0 = \left\langle G_0^1 G_0^2 ... G_0^C \right\rangle \sim \mathrm{DP}\left(\gamma, H_1 \times H_2 \times ... H_C\right)$ where $H_1 \times H_2 \times ... H_C$ is the product of base measure. Then, each group j will have a local atom G_j drawn from DP with the concentration parameter α and the global atom G_0 as $G_j = \left\langle G_j^1 G_j^2 ... G_j^C \right\rangle \sim \mathrm{DP}\left(\alpha, G_0\right)$. In other words, each random group-specific mixture distribution G_j sharing the same base probability measure G_0. Next, the atom for each data point i in group j at channel c is iid drawn as $\theta_{ji}^c \overset{\text{iid}}{\sim} G_j^c$ and the observation is generated subsequently $x_{ji}^c \sim F^c\left(\theta_{ji}^c\right)$.

Stick-breaking Representation For posterior inference, we characterize the above stochastic view using stick-breaking representation. We draw the global weight $\beta \sim \mathrm{GEM}\left(\gamma\right)$ and the topics (or patterns) $\phi_k^c \overset{\text{iid}}{\sim} H_c, \forall k = 1, ..., \infty, \forall c = 1...C$ such that the global atom $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\left\langle \phi_k^1, ..., \phi_k^C \right\rangle}$. For each group j, the local weight $\pi_j \overset{\text{iid}}{\sim} \mathrm{DP}\left(\alpha, \beta\right)$ such that $G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\left\langle \phi_k^1, ..., \phi_k^C \right\rangle}$. Then, the data point label $z_{ji} \sim \mathrm{Mult}\left(\pi_j\right)$. Finally, the observation is generated using the corresponding topic $\phi_{z_{ji}}^c$ as $x_{ji}^c \sim F^c\left(\phi_{z_{ji}}^c\right)$, $\forall j = 1, 2..., J$, $\forall i = 1..N_j$, $\forall c = 1...C$. Fig. 1 shows the graphical models of the PS-HDP from the stochastic view and the stick-breaking view.



(a) Stochastic view (b) Stick-breaking view

Fig. 1: The stochastic view and stick-breaking view of the Product Space HDP.

Posterior inference Our framework is a Bayesian model. For posterior inference, we utilize collapsed Gibbs sampling [12]. To integrate out π_j and ϕ_k due to conjugacy property, the two latent variables z_{ji} and β need to be sampled.

Sampling z_{ji}. We assign a data point x_{ji} to its component ϕ_k. The conditional distribution for z_{ji} is influenced by a collection of words associated with topic k across documents:

$$\begin{aligned} p\left(z_{ji}=k \mid \boldsymbol{x}, \mathbf{z}_{-ji}, \alpha, \beta, H\right) &= p\left(z_{ji}=k \mid \boldsymbol{z}_{-j}, \alpha, \beta\right) \times \\ &\quad p\left(x_{ji} \mid z_{ji}=k, \{x_{j'i'} \mid z_{j'i'}=k, \forall\left(j'i' \neq ji\right)\}, H\right) \\ &= \begin{cases} \left(n_{jk}^{-ji}+\alpha\beta_k\right) \times f_k^{-x_{ji}}\left(x_{ji}\right) & \text{used } k \\ \alpha \times \beta_{\text{new}} \times f_{k^{\text{new}}}^{-x_{ji}}\left(x_{ji}\right) & \text{new } k. \end{cases} \end{aligned}$$

The first term is recognized as the Chinese Restaurant Franchise (number of data points in group j follows topic k) while the second term is the predictive likelihood $f_k^{-x_{ji}}\left(x_{ji}\right)$.

Sampling $\boldsymbol{\beta}$. We sample the global mixing weight follow the approach presented in [22] that we sample $\boldsymbol{\beta}$ jointly with the auxiliary variable $\boldsymbol{m}$ $(0 \leq m \leq n_{jk})$:

$$p\left(m_{jk}=m \mid \boldsymbol{z}, \boldsymbol{m}_{-jk}, \boldsymbol{\beta}\right) \propto \text{Stirl}\left(n_{ij}, m_{jk}\right)\left(\alpha\beta_k\right)^m$$

$$p\left(\boldsymbol{\beta} \mid \boldsymbol{m}, \boldsymbol{z}, \alpha, \gamma\right) \propto \beta_{\text{new}}^{\gamma-1} \prod_{k=1}^{K} \beta_k^{\sum_j m_{jk}-1}$$

Sampling hyperparameters α and γ. To make the model robust in identifying the unknown number of clusters, we resample hyperparameters in each Gibbs iteration. The lower concentration parameter α is described in [22]. The upper concentration parameter γ is followed the techniques of [5].

4 Experiments

Our experiments are run on a PC equipped with a Intel Xeon E5-2460 CPU (8 cores, 2.6Ghz), 16GB RAM. We implemented the PS-HDP model using C#. For posterior inference, we do a Gibbs sampling over 500 iterations. In average, each iteration costs around 19 seconds.

4.1 Experimental Dataset

The Mobile Data Challenge (MDC) dataset [9] is a very large and real-world dataset collected from mobile phones and frequently used in pervasive and ubiquitous computing researches. In this paper, as we aim to demonstrate our framework to discover *identity–location–time* patterns, we limit the use of the MDC dataset to Bluetooth and WiFi data, in which the former is used for identifying nearby people (*identity*) and the latter is used for inferring the significant locations of users (*location*). All data have also associated with timestamps that can be used to infer the change of patterns over *time*. The dataset also includes information about places where a user stayed at least 20 minutes, which can be used to extract the location ground-truth of each Bluetooth/WiFi scan, and the IMEI number of the phones, which are used to extract the identity ground-truth. The data were extracted as described below.

Bluetooth data. For the convenience of the evaluation step, we keep only scans that include at least one identifiable device from the ground truth. Thus, only scans that capture one (or more) of participants' phones are retained.

WiFi data. For WiFi data, we eliminate access points that has been scanned less than 10 times as they are not statistically meaningful and do not affect the location discovery. Empty scans after this elimination are also removed.

Matching Bluetooth and WiFi scans. The Bluetooth and WiFi scans have sometimes different timestamps due to the design of the data collecting application. In our experiments, a Bluetooth and a WiFi scan is considered to be from the same scan if they have exactly the same timestamp.

Timestamp data. As we aim to discover context patterns during a day, we only use the hour, minute and second information of timestamps. We also convert them to a real number of hour (e.g. 10 hours and 30 minutes to 10.5 hours).

The MDC dataset has millions of scans. As we aim to demonstrate the ability to deal with multiple sources of the PS-HDP model, we want to keep the experimental dataset in an appropriate size. Thus, we select a month (Feb 2011 in our experiments), and then extract all scans in this month. Moreover, the MDC dataset is imbalanced as some users have much higher number of scans than others. To keep it balanced, we randomly choose at most 200 scans for each user. The final dataset includes 98 users with 16,950 scans (data points) in total, including 107 unique Bluetooth IDs and 1385 unique WiFi hotspot IDs.

4.2 Discovery of Joint *Identity–Location–Time* Patterns with Complete Data

We feed the extracted experimental dataset in section 4.1 to our PS-HDP model, where the Bluetooth and WiFi data are modelled by multinomial distributions,

whilst the Time data is modelled by a Gaussian distribution. At the end, there were totally 46 patterns (clusters) discovered. Each pattern has information of the time (represented by a mean and a variance) at which the pattern occurs; the probability of each Bluetooth ID presenting in the pattern; and the probability of each WiFi hotspot presenting in the pattern.

We display each pattern by three plots. The top plot displays the mean and the variance of the time. The middle plot display the top ten Bluetooth IDs. The bottom plot display the place ID at which the top 10 WiFi hotspot IDs are located. Both Bluetooth ID and place ID are displayed using tag cloud technique, where the size of each number reflects the contribution of the corresponding ID to the cluster. As a result, there could be some IDs with very small probability cannot be seen on the plot. This situation occurs frequently for the Bluetooth IDs. Fig. 2 shows an example pattern, where users who are corresponding to Bluetooth ID 24, 77, 22, 78, 84, and 94 often meet each other around 19:00 at a location where the WiFi hotspot is located.

Among 46 discovered patterns, we can find some interesting ones. For example, Fig. 3 shows two patterns 23 and 25 in which the Bluetooth ID 48 and 30 (which are corresponding to two specific participants) are usually co-exist at the place ID number 1, but at different time. The pattern 23 shows that they are usually at the place ID 1 at around 4:00, whereas the pattern 25 shows that they also meet each other at the same place at around 21:00. Furthermore, we can see that at 4:00 the person corresponding to the Bluetooth ID 48 (person 48) is more likely at the place ID 1 than the person corresponding to the Bluetooth ID 30 (person 30). The situation change at 21:00, where the person 30 presents at the place ID 1 more frequent than the person 48. These two patterns would not be discovered if one treat each data channel (i.e. Bluetooth and WiFi) independently. It shows the advantage of our model over traditional approaches.

4.3 Multilevel Pattern Analysis

We further analyze multilevel patterns at global (for all participants) and local (for each participant) levels. We aim to discover which patterns are regular for all users (e.g., what users often interact with each other, where the locations are,

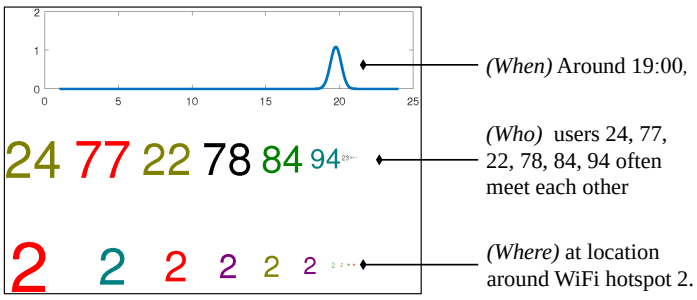
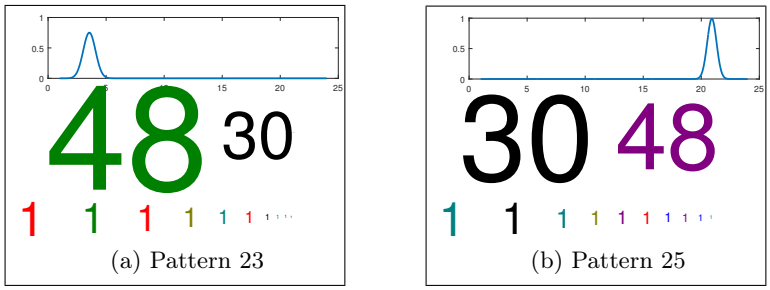


Fig. 2: Representation of *Who-When-Where* pattern.



(a) Pattern 23 (b) Pattern 25

Fig. 3: Two patterns with the same Bluetooth IDs and place, but different time.

and when they meet). Similarly, we learn the local regular pattern for individuals (e.g., who are this user often interacting with, at what locations, and when). We will ask similar questions for infrequent patterns at multilevel. To answer these questions, we utilize the global weight β to analyze the level of global patterns and the mixing proportion π_j to analyze the level of local patterns for user j.

Global Pattern Analysis We rely on β to find interesting patterns globally. Fig. 5 shows the pattern that has the largest β_k value (pattern number 6) and the pattern that has a smallest β_k value (pattern number 12), as well as the β_k values of all 46 patterns.

Local Pattern Analysis After running the PS-HDP model, we got a mixing proportion vector for each participant and cluster assignment for each data point. Using these information, we can rebuild patterns for each participant using only data points scanned by his phone. To help finding interesting patterns for each participant, we built a small application which let user choose a user (represented by a Bluetooth ID) and the application visualizes the two most interesting patterns (i.e. one has the largest and one has the smallest π_{ji} value) inferred from the data of that participant. Fig. 4 shows the 2 most interesting patterns for the user corresponding to the Bluetooth ID 77.

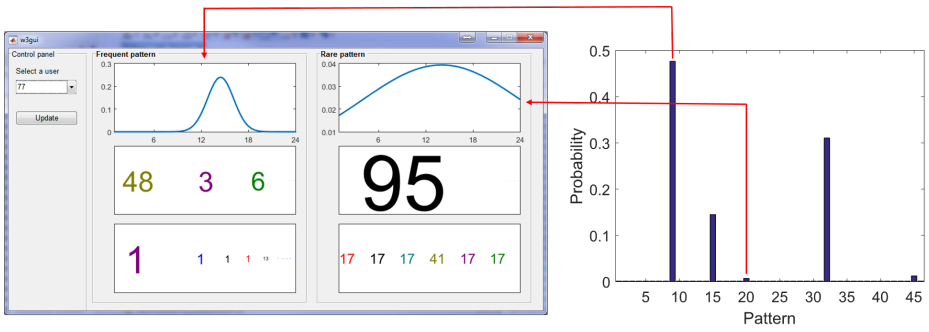


Fig. 4: Local patterns corresponding to the largest *(left)* and the smallest *(right)* π_{ji} values of the participant 77.

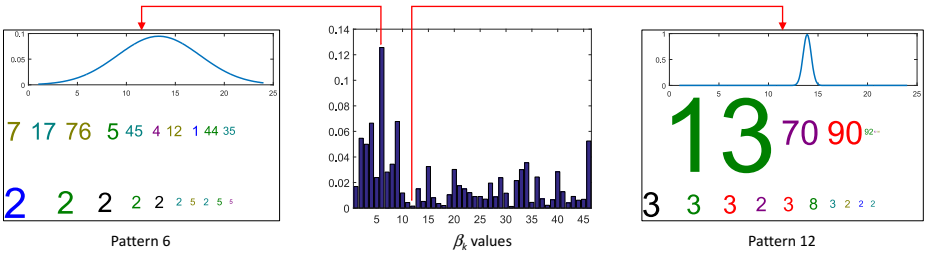


Fig. 5: Global patterns. *Left* and *Right*: the pattern corresponding to the largest and smallest β_k values respectively. *Middle*: β_k values of all patterns.

Furthermore, we compute a similarity matrix between 98 users using the Euclidian distance of user's mixture proportions. After that, we feed this similarity matrix into the Affinity Propagation clustering algorithm [6] to automatically group users together. We also compute the linear correlation coefficients from the users' mixture proportions as shown in the Fig. 6. We can see that users with similar correlation coefficients have been grouped together, showing that the mixture proportions produced by our PS-HDP model are meaningful.

4.4 Evaluation of Performance

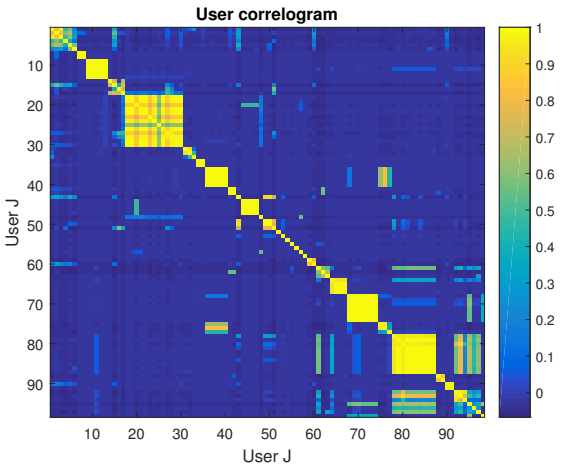


Fig. 6: Correlogram of users' mixture proportions.

After running the PS-HDP model, each data point will be assigned to a cluster through the indicator z_{ji}. Besides that, from the ground truth, the location of each data point could be known. We rely on this information to evaluate the clustering performance of our approach. To get the baselines, we run some quantitative clustering algorithms including K-Means, Non-negative Matrix Factorization (NNMF), Gaussian Mixture Model (GMM), DP-Means, and Affinity Propagation (AP) on the same dataset and use their performance results to compare with the performance of PS-HDP. We note that the K-Means, NNMF, and GMM methods require a number of cluster K to be specified beforehand, while our model can automatically learn the number of patterns from data. Therefore, to be fair, we run K-Means with different K values and calculate the average performance

scores. The range of K is selected around the values of K which are automatically discovered by DP-means, AP, and PS-HDP. Here, we vary K from 30 to 60. Table 1 shows the performance scores of these algorithms. Overall, the PS-HDM model achieves better scores in comparison with other clustering algorithms.

Algo.	Clusters	F1	NMI	RI	purity
K-Means	30-60	0.081	0.155	0.780	0.439
NNMF	30-60	0.094	0.320	0.790	0.543
GMM	30-60	0.116	0.345	**0.793**	0.545
DP-Means	48	0.122	0.103	0.757	0.399
AP	37	0.069	0.143	0.785	0.417
PS-HDP	46	**0.154**	**0.367**	0.781	**0.573**

Table 1: Performance of different clustering algorithms on the MDC dataset (note that F1-scores are usually low on pervasive data).

Amount	F1	purity
70%	0.109	0.253
50%	0.114	0.259
30%	0.134	0.288
10%	0.155	0.289

Table 2: F1-score and purity w.r.t. different amount of missing data.

4.5 Discovery of Joint *Identity–Location–Time* Patterns with Missing Data

To demonstrate the pattern discovery ability of our framework on data with missing values, we create 4 settings with different amount of missing data. More specifically, from the complete dataset in section 4.1, we randomly choose m data points (with m repeatedly set to 70%, 50%, 30% and 10% of total data points) and set their Bluetooth data as missing values. Each of these generated datasets is then inputted to the PS-HDP model and performance metrics are calculated (similar to the complete data settings). The experimental results show that F1-score and the purity increase when less missing data occurs (as shown in table 2). In other words, the more data is observed, the better patterns are. It means that the PS-HDP model can deal well with data even if they have missing values.

5 Conclusion

We presented a full Bayesian nonparametric model, called the Product Space HDP, which is extended from the HDP model by using a richer structure for the base measure being a product-space. Its major strengths and advantages firstly inherited from the Bayesian nonparametric approach including the ability to automatically grow the model complexity, hence does not require to know number of clusters beforehand; and secondly lie in the product-space approach which can deal with multiple heterogeneous data sources and with missing elements. The difference of our approach over existing ones is that it treats data from all sources equally, therefore does not require to specify the *content* and *contexts* channels. We apply the proposed model to one of the most fundamental problems in pervasive and ubiquitous computing: learning hidden activities from heterogeneous data collected from multiple sensors. Experimental results showed advantages of the PS-HDP model, including the ability of learning complex hidden patterns with good performance over popular existing clustering methods.

References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Cao, L., Zhang, H., Zhao, Y., Luo, D., Zhang, C.: Combined mining: discovering informative knowledge in complex data. Trans. on SMC 41(3), 699–712 (2011)
3. Do, T.M.T., Gatica-Perez, D.: Human interaction discovery in smartphone proximity networks. Personal and Ubiquitous Computing 17(3), 413–431 (2013)
4. Dousse, O., Eberle, J., Mertens, M.: Place learning via direct wifi fingerprint clustering. In: Mobile Data Management (MDM), 2012. pp. 282–287. IEEE (2012)
5. Escobar, M., West, M.: Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90(430), 577–588 (1995)
6. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science 315(5814), 972–976 (2007), http://www.sciencemag.org/content/315/5814/972
7. Huynh, V., Phung, D., Nguyen, L., Venkatesh, S., Bui, H.: Learning conditional latent structures from multiple data sources. In: ACML, pp. 343–354 (2015)
8. Kulis, B., Jordan, M.I.: Revisiting k-means: New algorithms via bayesian nonparametrics. In: Proc. ICML (2012)
9. Laurila, J.K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T.M.T., Dousse, O., Eberle, J., Miettinen, M., et al.: The mobile data challenge: Big data for mobile computing research. In: Pervasive Computing (2012)
10. Lee, D.D., Seung, H., et al.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
11. Liang, P., Petrov, S., Jordan, M.I., Klein, D.: The infinite pcfg using hierarchical dirichlet processes. In: EMNLP'07. pp. 688–697 (2007)
12. Liu, J.: The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. American Statistical Association 89, 958–966 (1994)
13. McLachlan, G., Peel, D.: Finite mixture models (2004)
14. Nguyen, T.C., Phung, D., Gupta, S., Venkatesh, S.: Extraction of latent patterns and contexts from social honest signals using hierarchical Dirichlet processes. In: PERCOM. pp. 47–55 (2013)
15. Nguyen, T.B., Nguyen, T.C., Luo, W., Venkatesh, S., Phung, D.: Unsupervised inference of significant locations from wifi data for understanding human dynamics. In: Proc. of MUM 2014. pp. 232–235 (2014)
16. Nguyen, T., Phung, D., Venkatesh, S., Nguyen, X., Bui, H.: Bayesian nonparametric multilevel clustering with group-level contexts. In: ICML. pp. 288–296 (2014)
17. Nguyen, V., Phung, D., Venkatesh, S., Bui, H.: A bayesian nonparametric approach to multilevel regression. In: Proc. of PAKDD. pp. 330–342 (2015)
18. Pentland, A.: Automatic mapping and modeling of human networks. Physica A: Statistical Mechanics and its Applications 378(1), 59–67 (2007)
19. Phung, D., Nguyen, X., Bui, H., Nguyen, T., Venkatesh, S.: Conditionally dependent Dirichlet processes for modelling naturally correlated data sources. Tech. rep., Pattern Recognition and Data Analytics, Deakin University (2012)
20. Ren, L., Dunson, D.B., Carin, L.: The dynamic hierarchical dirichlet process. In: Proceedings of the 25th ICML'08. pp. 824–831. ACM, New York, NY, USA (2008)
21. Schilit, B.N., Theimer, M.M.: Disseminating active map information to mobile hosts. Network, IEEE 8(5), 22–32 (1994)
22. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical Dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)
23. Zhang, J., Song, Y., Zhang, C., Liu, S.: Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In: SIGKDD. pp. 1079–1088 (2010)