Simultaneous Customer Segmentation and Behavior Discovery

Siqi Zhang[1], Ling Luo[2], Zhidong Li[1(✉)], Yang Wang[1], Fang Chen[1], and Richard Xu[1]

[1] School of Software, The University of Technology Sydney, Sydney, NSW 2006, Australia
siqi.zhang-3@student.uts.edu.au,
{Zhidong.Li,Yang.Wang,Fang.Chen,Richards.Xu}@uts.edu.au
[2] School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC 3010, Australia
ling.luo@unimelb.edu.au

Abstract. Customer purchase behavior segmentation plays an important role in the modern economy. We proposed a Bayesian non-parametric (BNP)-based framework, named Simultaneous Customer Segmentation and Utility Discovery (UtSeg), to discover customer segmentation without knowing specific forms of utility functions and parameters. For the segmentation based on BNP models, the unknown type of functions is usually modeled as a non-homogeneous point process (NHPP) for each mixture component. However, the inference of these models is complex and time-consuming. To reduce such complexity, traditionally, economists will use one specific utility function in a heuristic way to simplify the inference. We proposed to automatically select among multiple utility functions instead of searching in a continuous space. We further unified the parameters for different types of utility functions with the same prior distribution to improve efficiency. We tested our model with synthetic data and applied the framework to real-supermarket data with different products, and showed that our results can be interpreted with common knowledge.

Keywords: Customer behavior · Bayesian non-parametric model · Utility function · Segmentation

1 Introduction

Customer segmentation is a common technique that allows a business to make better use of their resources and make more profits by studying the activities associated with the purchase behavior [4,9]. The traditional framework of customer segmentation is based on their demographic data. There are two drawbacks to this setting: 1) the collection of demographic data is wandering between the balance of completion and customer's privacy, and the missing of key attributes leads to unreliable segmentation. 2) The utility functions need to

H. Yang et al. (Eds.): ICONIP 2020, CCIS 1332, pp. 122–130, 2020.
https://doi.org/10.1007/978-3-030-63820-7_14

be determined before segmentation, whilst in our provided framework, we jointly estimate the segmentation of customers and the form and parameters of utility functions at the same time. Without explicitly setting the number of clusters, we employ the Bayesian non-parametric (BNP) framework to conduct the segmentation. Several challenges are hampering the inference when we jointly estimate the utility functions, parameters, and segmentation in the BNP framework. First, the optimization of utility function is a functional problem requiring modeling of stochastic process. In the previous work, Luo et al. describe customer purchase behavior by non-homogeneous point process (NHPP) [4,9], using a base measure of the polynomial and trigonometric functions. Such models are usually mathematically complex, requiring careful design for compatibility, robustness, and scalability. Instead, we proposed an approximation model, but it can still keep the flexibility to model the utility functions with different parameters in different types. Second, parameters for different types of utility functions are inconsistent for both of their prior and likelihood, given the diversity of their meanings. The inconsistency of posteriors can cause enormous amount of heuristic work, as the modeling heavily relies on expert knowledge. The re-parametric solution is to wrap up the identity of utility function with a parameter-free nonlinear function to keep the consistency of parameters. Such designation drastically reduces the inconvenience of designing utility functions by assuming all the parameters are generated from the same posterior distribution. As a result, the model has higher generalizability and is easier to interpret.

There are three main **contributions** of this work. First, we proposed an automatic and generalizable framework based on the BNP model, which can simultaneously segment the customers considering their behavior, discover their utility type, and how their purchase behavior is influenced by the product price (more external factors can be directly involved). Second, our framework overcomes the complex modeling and inefficient inference for NHPP. We unify the parameter estimation for different utility functions so that the predefined conjugate priors can be used to drastically simplify the inference. At last, we conduct case studies on the real-world supermarket data and show the patterns discovered using our method.

2 Related Work

To group customers, there are different segmentation or clustering approaches, which include decision tree, K-means, mixture model, and density peaks. Using decision trees for customer behavior research is easy to understand but the main drawback is that it must be learned with labeled data. The other methods are unsupervised but their generalizability is limited by parameter setting. Hence, we focus on the BNP model.

The NHPP can naturally fit the problem of describing events (e.g. purchase) based on observations, with different types of utility functions. However, the inference of NHPP under the BNP framework is an extremely complicated task. Previous research focused on the inference of intensity function without considering grouping at the same time. The tractable inference was proposed in [1], then

the major research is to find faster ways to infer the intensity function [6,12]. There are also some models [7] considering the change of customer segmentation over time. The requirement of inference design is also the main obstacle to generalize these models. We propose a effective model to approximate NHPP with a unified framework to generalize.

Exploring purchase behavior based on price sensitivity is a common topic of customer analysis. Similar problem attracts research in a variety of applications, such as tourism and airline industry [8] . The utility function is the main assumption used in such studies to assess customer purchase behavior [13] to describe the relationship between price and purchase. The model determines one utility function for each customer, then analyses/segments customers according to the assumed utility function.

3 Methodology

3.1 Problem Definition

Given a customer $i = \{1...N\}$, we know a series of their purchase events $\boldsymbol{y_i}$, where $\boldsymbol{y_i} = \{y_{i,j} | y_{i,j} \in \mathbb{N}^+\}$ is the number of products that customer i purchased for their j^{th} purchase. The corresponding price is $\boldsymbol{x_i} = \{x_{i,j} | x_{i,j} \in \mathbb{R}^+\}$. We assume that there are M_i observations for i in total. Here we use discount rate as the price to normalize the price value of different types of products. Details will be shown in the data pre-processing in Sect. 5.3. Our target is to segment customers into unknown number of groups, so that a group index k_i needs to be obtained for i. The grouping is based on the function f_i that can map price $x_{i,j}$ with purchase behavior $y_{i,j}$.

Considering the efficiency and generalizability, we covert the problem into the semi-parametric model but it can approximately cover the space that a full BNP model can cover. Here we assume that three utility functions can represent most types of relationships between customers' purchase behavior and price. As we do not know which function should be used for each customer, we use a latent variable $u_i \in \{1, 2, 3\}$ to denote the function selected for customer i. We hope to jointly estimate group index k_i and utility function type u_i, without heuristically selecting for each customer, by the *integration of a Bayesian semi-parametric model for utility function selection and BNP model for grouping.*

The flow of our work is that all the customers will be compared with different utility functions. Then our algorithm will determine the best latent utility function used to describe the customer behavior and customer group. Suppose two people are associated with utility function 1, and one person is associated with utility function 2. The customers in the same group have the same parameters in their utility functions. Since we have the unified utility function, the parameters for different forms of utility functions can still have the same parameter values.

3.2 Utility Functions

These three functions can explain most of the relation between purchase behavior and price [2]. The traditional way to represent different functions is based

on different parameters. This will cause many heuristic settings, which requires domain knowledge to determine what priors are needed for each different setting, and what likelihood function is suitable for each parameter, otherwise, it could cause intractability with non-conjugated priors. This setting limits the generalizability and scalability of the system.

To overcome such an obstacle, we use $\boldsymbol{y} = \boldsymbol{a}g(\boldsymbol{x}) + \boldsymbol{b}$ as a general representation of any relation between demand and price change. The parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ are unified, which can be interpreted in the same way in different utility functions, such as $\boldsymbol{a}$ as the coefficient and $\boldsymbol{b}$ as adjustment. Then we can represent such setting with fixed prior functions. The common utility functions can be reduced to our general representation, and directly fit our model. Such a setting provides the generalization capability so that it can be easily extended to incorporate a variety of utility functions. Specifically, the utility functions explored in this work are: f_i^1: $y_i = a_{i,1}x_i + b_{i,1}$, f_i^2: $y_i = a_{i,2}\log(x_i) + b_{i,2}$, f_i^3: $y_i = a_{i,3}e^{x_i} + b_{i,3}$.

The utility function is selected based on the negative log-likelihood loss function. It is a common way to measure if the utility function can fit the data points well or not.

3.3 Simultaneous Customer Segmentation and Utility Estimation Model

This section describes the details of the Simultaneous Customer Segmentation and Utility Estimation (UtSeg) model and introduces the generative process of parameters, latent variables, and observations of UtSeg. The model is mainly used for cluster customers into groups and infer their utility functions.

In the UtSeg, α_0 is the hyperparameter for the CRP. Then, each customer i will get $\boldsymbol{a}_i = [a_{i1}, a_{i2}, a_{i3}]$ and $\boldsymbol{b}_i = [b_{i1}, b_{i2}, b_{i3}]$ by using curve fitting function based on price and purchase information. We assume that $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ follow Gamma distribution. This is because of the Poisson distribution used to estimate the purchase number, given by $N_i(x_i) \sim Poi(y_i(x_i))$. The Poisson distribution can be decomposed as a superposition of multiple Poisson distributions with the summation of frequencies as the overall frequency. Therefore, b_i is also the parameter of a Poisson distribution.

This is the first property that can drastically reduce the complexity of inference, because the conjugated prior can be set for both gamma distribution, without sampling in high dimension space for stochastic process.

For each customer i, the generative process of UtSeg can be represented as follows:

$$\begin{aligned} &k_i \sim CRP(\alpha_0), u_i \sim Mul(u_0), \alpha_{k_i} \sim Gam(\theta_a), a_{i,u_i} \sim Gam(\alpha_{k_i}), \\ &\beta_{k_i} \sim Gam(\theta_b), b_{i,u_i} \sim Gam(\beta_{k_i}), l_{i,u_i} \sim N(\mu_{i,u_i}, \sqrt{M_i}\sigma_0) \\ &\mu_{i,u_i} = \sum_{j=1\cdots M_i} a_{i,u_i} g_{u_i}(x_i) + b_{i,u_i} - y_i \end{aligned} \tag{1}$$

– We generate table index k_i based on CRP, using the hyperparameter α_0;

- For utility function selection, an function index u_i is generated for customer i with Multinomial distribution parameterized by u_0;
- We generate a latent variable for each table k_i, for both coefficient variable α_{k_i} and offset variable β_{k_i} with the base measure parameterized by $gamma(\theta_a)$ and $gamma(\theta_b)$[1];
- a_i and b_i are generated based on α_{k_i} and β_{k_i} using Gamma distributions;
- The selected function should fit the observations, so the minimised loss l_{iu_i} can be learned based on Sect. 3.2. l_{iu_i} is assumed to be Gaussian distributed[2] loss, with variance σ_0, mean $\mu_{i,u_i} = \sum_j a_{i,u_i} g_{u_i}(x_i) + b_{i,u_i} - y_i$, based on a_i and b_i as a_{i,u_i} and b_{i,u_i} respectively.

Therefore, the joint probability of the model is:

$$\begin{aligned}
&P(k_{1\cdots n}, u_{1\cdots n}, \alpha_0, \theta_a, \theta_b, l_{1\cdots n,1\cdots 3}, a_{1\cdots n,1\cdots 3}, b_{1\cdots n,1\cdots 3}) \\
\propto &\prod_i P(k_i|\alpha_0)P(\alpha_{k_i}|\theta_a)P(\beta_{k_i}|\theta_b)P(u_i|u_0)P(a_{i,u_i}|\alpha_{k_i}, k_i)\cdot \\
&P(b_{i,u_i}|\beta_{k_i}, k_i)P(l_{i,u_i}|u_i, a_{i,u_i}, b_{i,u_i}, g_{u_i}, x_i, y_i, \sigma_0, M_i) \\
= &\prod_i CRP(k_i|\alpha_0)Gam(\alpha_{k_i}|\theta_a)Gam(\beta_{k_i}|\theta_b)Mul(u_i|u_0)Gam(a_{i,u_i}|\alpha_{k_i}) \\
&Gam(b_{i,u_i}|\beta_{k_i})N(l_{i,u_i}|\mu_{i,u_i}, \sqrt{M_i}\sigma_0)
\end{aligned} \tag{2}$$

4 Inference: Gibbs Sampling for UtSeg Model

Gibbs Sampling is a Markov Monte Carlo method (MCMC) [1,4], which is widely used in the inference. In the UtSeg model, each customer is assigned to a utility function based on the multinomial prior and Gaussian likelihood for the loss function. The parameters θ_a and θ_b are randomly initialized and u_i from the last step is used. The possible sampling result can be any existing table or starting a new table. For each customer i, the posterior probability to select a table k_i. Where k_{i^-} represents the current table assignments except for customer i. Similarly we sample the form of utility u_i by: Eq. (4). By sampling all the k_i and u_i iteratively, we can get the utility function allocation for all customers.

$$\begin{aligned}
&p(k_i = k|k_{i^-}, u_{1\cdots n}, \alpha_0, \theta_a, \theta_b, l_{1\cdots n,1\cdots 3}, a_{1\cdots n,1\cdots 3}, b_{1\cdots n,1\cdots 3}) \\
\propto &CRP(k|k_{i^-}, \alpha_0)Gam(\alpha_k|\theta_a)Gam(\beta_k|\theta_b)Gam(a_{i,u_i}|\alpha_k) \\
&Gam(b_{i,u_i}|\beta_k)N(l_{i,u_i}|\mu_{i,u_i}, \sqrt{M_i}\sigma_0).
\end{aligned} \tag{3}$$

$$\begin{aligned}
&p(u_i = u|k_{1\cdots n}, u_{i^-}, \alpha_0, \theta_a, \theta_b, l_{1\cdots n,1\cdots 3}, a_{1\cdots n,1\cdots 3}, b_{1\cdots n,1\cdots 3}) \\
\propto &Mul(u|u_0)Gam(\alpha_k|\theta_a)Gam(\beta_k|\theta_b)Gam(a_{i,u_i}|\alpha_k) \\
&Gam(b_{i,u_i}|\beta_k)N(l_{i,u_i}|\mu_{i,u_i}, \sqrt{M_i}\sigma_0).
\end{aligned} \tag{4}$$

[1] For a Gamma distribution, we simplify both actual parameters into one parameter.
[2] This can be determined by the loss used. We use the quadratic loss, but Gaussian distribution is used to approximate the Chi-square distribution when data volume is large.

5 Experiments

5.1 Experiment Setup

Baseline Models

UtSeg-(1-3): This is a simplified model from UtSeg, which is based on CRP and one utility function to describe customer purchase behavior. Each method corresponds to one of the utility functions [2].
CRP-GM: This baseline is CRP with Gaussian mixture component [1]. We use $x_{i,j}$ and $y_{i,j}$ to compute the likelihood of CRP.
NHPP: This model is based on [3], which assumes that the mixture component is NHPP with different types of intensity functions.
Clustering: This model is parametric segmentation, which includes classic clustering models **K-Means (KM)** and **Density Peak (DP)** [5].

Evaluation Measurements

Confusion matrix: With ground truth data, we can use the confusion matrix (CM) to show the true and learned grouping.
Clustering distance: The average distance inside groups (ADIG) and the average distance between groups(ADBG) are used. The ADIG refers to the average distance between sample points of the same group, *lower* is better. The ADBG refers to the distance between groups, which the *greater* distance means the larger difference between groups.
Segmentation Log-Likelihood (LL): Segmentation LL can compare the fitness of different models. A *higher* log-likelihood value means the model fits better. To compare all models, we use the obtained group index. The likelihood is obtained by the Poisson distribution parameter on the average of the selected coefficients in each group. For CRP-GM, we double the likelihood as it only has one parameter [6].

5.2 Synthetic Data Set

We follow (1) to generate synthetic data for experiment. We generated pairs of purchased number and price for 100k customers. The evaluation results are shown in Table 1. Firstly, UtSeg has the best result for both distances. Naturally, using the parametric method can obtain better results when the chosen parameters happen to be similar to the true parameter. However, without special design or knowledge, the parameters are unknown, which is the largest obstacle to use those parametric methods. On the contrary, our method can be generalized to unseen cases without such settings. In terms of computation time, the UtSeg model is compared with the optimized NHPP model as implemented in

Table 1. Evaluation results on synthetic data.

	ADIG	ADBG	Segmentation LL	CM accuracy
UtSeg	0.563	0.670	−1321500	0.383
UtSeg-1	0.626	0.616	−1523572	0.372
UtSeg-2	0.856	0.709	−2651849	0.301
UtSeg-3	1.077	0.816	−3247918	0.193
CRP-GM	0.687	0.532	−1650943	0.348
NHPP	0.572	0.730		0.427
K-means K = 5	0.696	0.616		0.533
K-means K = 7	0.580	0.690		0.394
Density peak k = 5	0.670	0.689		0.405
Density peak k = 7	0.665	0.765		0.376

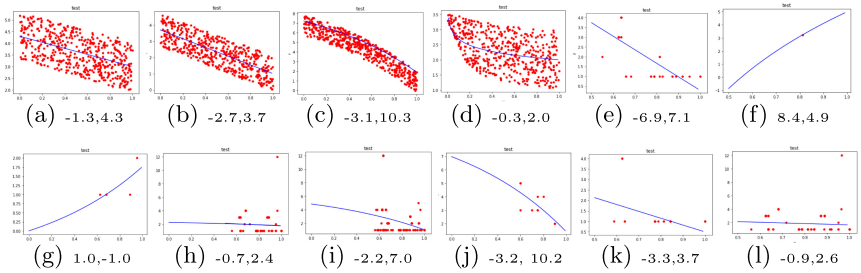

Fig. 1. Utility functions for different data sets: (a)–(d) are for synthetic data, (e)–(i) are parts for four product types. 5 utility functions and observed points are shown for each type. The learned parameters α_k and β_k are under the plots.

5.3 Case Study

In this section, we present a case study on a real data set of 1,529,057 purchase transaction records collected by an Australian national supermarket chain in 2014 and we use 41,210 of them. Base on the variation of customer purchase quantities (normalized according to product volume) and purchase price (normalize as a discount rate, which is in the range of [0, 1]). We run the algorithm on 4 products separately to see how the customers are segmented for their behavior. Our result can support the product providers to set promotion or stimulation. In the prepossessing, we further normalize quantity by purchased quantity per week subtracting the average amount throughout the year. This normalization can remove the influence of demand levels.

The results are given in Table 2. In the study, the UtSeg model provides better results in most cases, without setting functions and parameters using domain knowledge. The results are consistent with what we observed in the synthetic data.

Table 2. Evaluation results on real data.

	Measure	UtSeg	UtSeg-1	UtSeg-2	UtSeg-3	NHPP
Milk	ADIG	0.61	0.74	0.86	1.08	0.65
	ADBG	0.90	0.88	0.86	0.67	0.83
	Segment LL	−1.3E+05	−1.5E+05	−1.9E+05	−2.5E+05	
Chips	ADIG	0.63	0.73	0.83	1.01	0.68
	ADBG	0.80	0.67	0.71	0.76	0.82
	Segment LL	−1.1E+05	−2.1E+05	−2.4E+05	−2.6E+05	
Chocolate	ADIG	0.55	0.59	0.71	0.75	0.58
	ADBG	0.83	0.79	0.74	0.77	0.79
	Segment LL	−4.5E+04	−1.8E+05	−2.0E+05	−2.4E+05	
Softdrinks	ADIG	0.56	0.69	0.74	0.83	0.54
	ADBG	0.67	0.62	0.57	0.52	0.59
	Segment LL	−6.9E+04	−1.4E+05	−1.9E+05	−2.1E+05	
	Measure	CRP-GM	KM-3	KM-7	DP-3	DP-7
Milk	ADIG	0.90	1.06	0.55	1.07	0.59
	ADBG	0.83	0.60	0.72	0.79	0.89
	Segment LL	−2.0E+05	–	–	–	–
Chips	ADIG	0.73	0.97	0.78	0.95	0.67
	ADBG	0.61	0.88	0.53	0.89	0.54
	Segment LL	−1.8E+05	–	–	–	–
Chocolate	ADIG	0.66	0.83	0.55	1.31	0.69
	ADBG	0.75	0.58	0.72	0.58	0.81
	Segment LL	−1.8E+05	–	–	–	–
Softdrinks	ADIG	0.79	0.93	0.49	1.34	0.70
	ADBG	0.53	0.56	0.77	0.53	0.90
	Segment LL	−1.6E+05	–	–	–	–

6 Conclusion

In this paper, we propose a BNP framework to segment customers without knowing their utility functions of purchase behavior. Using the semi-parametric method, we unify the parameters of different types of utility functions into the same representation so they can be generated by the same distribution. This proposed technique significantly saved the effort to design a special inference algorithm so that we can efficiently learn the latent variables. The setting also makes it easier to generalize our method to comply with more utility functions.

References

1. Adams, R.P., Murray, I., MacKay, D.J.: Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 9–16 (2009)
2. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Parzen, E., Tanabe, K., Kitagawa, G. (eds.) Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics), pp. 199–213. Springer, Heidelberg (1998). https://doi.org/10.1007/978-1-4612-1694-0_15
3. Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B., Blei, D.M.: Hierarchical topic models and the nested Chinese restaurant process. In: Advances in Neural Information Processing Systems, pp. 17–24 (2004)
4. Kamakura, W.A., Russell, G.J.: A probabilistic choice model for market segmentation and elasticity structure. J. Market. Res. **26**(4), 379–390 (1989)
5. Kamen, J.M., Toman, R.J.: Psychophysics of prices. J. Market. Res. **7**(1), 27–35 (1970)
6. Lloyd, C., Gunter, T., Osborne, M., Roberts, S.: Variational inference for Gaussian process modulated Poisson processes. In: International Conference on Machine Learning, pp. 1814–1822 (2015)
7. Luo, L., Li, B., Koprinska, I., Berkovsky, S., Chen, F.: Discovering temporal purchase patterns with different responses to promotions. In: Proceedings of the 25thACM International on Conference on Information And Knowledge Management, pp. 2197–2202. ACM (2016)
8. Masiero, L., Nicolau, J.L.: Tourism market segmentation based on price sensitivity: finding similar price preferences on tourism activities. J. Travel Res. **51**(4), 426–435 (2012)
9. McDonald, M., Christopher, M., Bass, M.: Market segmentation. In: McDonald, M., Christopher, M., Bass, M. (eds.) Marketing, pp. 41–65. Springer, Heidelberg (2003). https://doi.org/10.1007/978-1-4039-3741-4_3
10. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., Welling, M.: Fast collapsed Gibbs sampling for latent Dirichlet allocation. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 569–577. ACM (2008)
11. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **344**(6191), 1492–1496 (2014)
12. Samo, Y.L.K., Roberts, S.: Scalable nonparametric Bayesian inference on point processes with Gaussian processes. In: International Conference on Machine Learning, pp. 2227–2236 (2015)
13. Sirvanci, M.B.: An empirical study of price thresholds and price sensitivity. J. Appl. Bus. Res. (JABR)**9**(2), 43–49 (1993)
14. Smyth, P.: Clustering sequences with hidden Markov models. In: Advances in Neural Information Processing Systems, pp. 648–654 (1997)