Exploring Latent Structure Similarity for Bayesian Nonparameteric Model with Mixture of NHPP Sequence

Yongzhe Chang[1,2(✉)], Zhidong Li[1,3], Ling Luo[4], Simon Luo[1], Arcot Sowmya[2], Yang Wang[1,3], and Fang Chen[1,3]

[1] Data 61 CSIRO, Sydney, Australia
{yongzhe.chang,zhidong.li,simon.luo,yang.wang,fang.chen}@data61.csiro.au
[2] University of New South Wales, Kensington, Australia
{yongzhe.chang,arcot.sowmya}@unsw.edu.au
[3] University of Technology Sydney, Ultimo, Australia
{zhidong.li,yang.wang,fang.chen}@uts.edu.au
[4] The University of Melbourne, Melbourne, Australia
ling.luo@unimelb.edu.au

Abstract. Temporal point process data has been widely observed in many applications including finance, health, and infrastructures, so that it has become an important topic in data analytics domain. Generally, a point process only records occurrence of a type of event as 1 or 0. To interpret the temporal point process, it is important to estimate the intensity of the occurrence of events, which is challenging especially when the intensity is dynamic over time, for example non-homogeneous Poisson process (NHPP) which is exactly what we will analyse in this paper. We performed a joint task to determine which two NHPP sequences are in the same group and to estimate the intensity resides in that group. Distance dependent Chinese Restaurant Process (ddCRP) provides a prior to cluster data points within a Bayesian nonparametric framework, alleviating the required knowledge to set the number of clusters which is sensitive in clustering problems. However, the distance in previous studies of ddCRP is designed for data points, in this paper such distance is measured by dynamic time warping (DTW) due to its wide application in ordinary time series (e.g. observed values are in $\mathcal{R}$). The empirical study using synthetic and real-world datasets shows promising outcome compared with the alternative techniques.

Keywords: NHPP · ddCRP · DTW

1 Introduction

A temporal point process is basically a random list process whose observations are times of events. A simple temporal point process can be typically modeled by its intensity λ which related to time t, familiarity with Poisson process [18] and

T. Gedeon et al. (Eds.): ICONIP 2019, LNCS 11954, pp. 432–444, 2019.
https://doi.org/10.1007/978-3-030-36711-4_36

Hawkes process [9]. In the real world, many phenomena can be represented as temporal point process, for example the happening of earthquakes, customers' shopping records and patients' hospitalization records etc. For these examples, one common issue is that we do not know how many time of events may occur and at what time. However, latent pattern may behind superficial phenomenon, for instance an earthquake can cause aftershocks. An essential target is to estimate intensity function of temporal point process in order to predict the future events.

A non-homogeneous Poisson process (NHPP) is a counting process with its average rate of arrivals varying with time. In general, intensity function of NHPP can be any arbitrary function over temporal variables, due to this fact, it is difficult to recover the true underlying intensity function form in many cases. Therefore it is challenging in estimating intensity function for temporal point process directly. Moreover, estimation for intensity function may suffer from the model selection problem and the insufficient power of intensity functions to represent complex real-world problems.

For a given NHPP, we can first estimate intensity function then cluster them according to each intensity function, assign those NHPP with same or similar intensity functions, for example [14] discussed Bayesian nonparametric methods in clustering and in [19], Chinese restaurant process (CRP) Dirichlet mixture model and Hierarchical Dirichlet mixture model were used in achieving the goal. When using Bayesian nonparametric methods in clustering, the framework of Bayesian nonparametric prior and a mixture component is usually being used, such as in [7] they combined a nonparamatric prior with Dirichlet allocation in learning topic hierarchies from data.

However, Bayesian nonparametric clustering methods are complicated and slow, especially for clustering NHPP. First, in order to be more efficient, the likelihood is preferable to be conjugate to the prior, otherwise Monte Carlo Markov chain (MCMC) should be utilzied for inference, which is tractable but less efficient. Second, each object being clustered itself is a Bayesian nonparametric process, e.g. the intensity is formulated as a transformed Gaussian process (GP) which guarantees the intensity being positive. However, as a well known result, the complexity of GP inference is cubic so the inference algorithm becomes extremely inefficient. Thirdly, in Bayesian nonparametric clustering, the highly similar components can still be assigned to different clusters due to sampling randomness, which leads to a undesirable result.

Considering the problems above, in this paper, we proposed a Bayesian nonparametric model which is a DTW based ddCRP model and we shortly named it DTW-ddCRP model, to discover latent pattern of intensity function and cluster NHPP, which is much faster and easier for inference than CRP based clustering model mentioned above. In our model, we used DTW distance measurement for similarity learning and reflecting structure information within processes, then we proposed ddCRP-based model to study the partition of failure event processes. DTW distance measurement is an accurate measure in looking for the nearest neighbor and ddCRP gives partition result without knowing the number of partitions at first. However, DTW can only find out the nearest one, ddCRP has

the ability of deciding whether the two customer should be put into one cluster because of the setting of scalar α and decay function parameter. We combined ddCRP with DTW distance measure, and used Gibbs sampler in the inference part of our model. It makes contribution via tackling the challenges in the temporal point process pattern learning: clustering temporal point processes without knowing latent pattern, which is especially latent dynamic intensity here; estimating latent dynamic intensity of water mains failure bursts by inferring the probability of which cluster one failure event process would in; then according to the ddCRP inference, the label of each table (cluster) can be obtained, which is an approximation of water mains failure intensity.

2 Related Work

In this section, we give a brief introduction about related work. Section 2.1 introduces sequence similarity learning method especially what we use in our model, the DTW distance measurement. In Sect. 2.2 we will describe Bayesian nonparametric methods in clustering and parameter estimation work, distance dependent Chinese restaurant process.

2.1 Clustering for Time-Series and DTW Distance

Clustering for time-series has been shown effective in providing useful information in various domain and attracted increasing interest as a part of the effort in the temporal data mining research [12]. In [8], Han and Kamber classified clustering methods into five main categories: hierarchical methods, density-based methods, partitioning methods, grid-based methods, and model-based methods.

In clustering model for time-series, the distance measurement approach is an important part. The most popular distance measurements for time-series data include Euclidean distance, Hausdorff distance, HMM-based distance and DTW distance, etc [2]. Each of these distance measurements has its advantages and disadvantages, for example the Euclidean distance in simple but can only be used when different sequences have same length; the DTW distance is computational expensive but can treat with different length of sequences. In our model, we chose DTW distance measurement mainly because its good performance on measuring sequences with different length.

DTW algorithm is a distance learning measurement by obtaining the optimal alignment between two time series, especially when the two sequences vary in speed, and has been widely used in time series clustering not only in academic domain, but also in many industrial projects [13] in the past decades. These applications include image processing, data mining, computer graphics and so on. It is one of the best measurement in searching for the nearest neighbor [6], defined for time series by measuring the distance between temporal sequences that vary in frequency and length. As all known, the expensive calculation problem is always the key problem for distance measurements, researchers have done plenty of contribution to speed it up and have a better performance, for example [16] and [15].

2.2 Joint Model for Clustering and Parameter Estimation

The Chinese restaurant process (CRP) [4] is a probability distribution over partitions. It is a popular representation of Dirichlet process (DP) and emphasis the clustering nature of DP. CRP assumes a prior distribution over the clusters of customers and is widely used in Bayesian nonparametric mixture models. It gives the probability of which table a coming customer may sit at by assuming that there are infinite tables in the Chinese restaurant. The CRP model is exchangeable because the change of the coming order of customers does not change the distribution of partition and each customer's probability assigned to any particular table only depends on how many customers already sit at that table. A concentration parameter α determines the probability that the customer sits at a new table or not. So it is obvious that a table with more customers has higher probability on attracting a new coming customer. The CRP is a widely used clustering method especially on mixture models and open-ended problems.

However, in CRP model, there is no relation among customers, which may assign those customers that have similar pattern into different tables. Then ddCRP [3] was introduced to solve this problem that can model random partitions on non-exchangeable data. Unlike CRP, in ddCRP one customer does not choose which table to sit at directly, but choose the customer who has the nearest distance between him to sit with, which can be comprehended as the ddCRP modifies the CRP by determining the table assignment via customer relations. It is a widely used method that provide a Bayesian nonparametric prior for clustering models and mixture models. Different from k-means and other clustering methods, with ddCRP prior we can learn the number of clusters automatically from data without knowing it beforehand. Also in the preliminary work section we will expound this distribution elaborately.

3 Model and Inference

3.1 Preliminary Work

To solve NHPP clustering problem, a DP and Gaussian process (GP) mixture model can be used. Assume we have n observations that are NHPP $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, here we don't know intensity function for each process, assume intensity functions are $\Lambda = (\lambda_{t1}, \lambda_{t2}, ..., \lambda_{tK})$, then equation below can achieve estimation of each intensity and clustering these NHPP:

$$P(\Lambda, Z|X) \propto P(X|\Lambda, Z)P(Z|\alpha)P(\Lambda) \tag{1}$$

In Eq. (1), Z is vector of sitting configuration, where each element is the table $z_i = k$ for $\mathbf{x}_i$; $P(X|\Lambda, Z)$ is likelihood; $P(Z|\alpha)$ is a DP prior to figure out Z, which means DP prior determines the number of clusters; $P(\Lambda) = GP(\lambda_1)GP(\lambda_2)...GP(\lambda_n)$ that samples intensity for each NHPP. By doing this, we can not only get intensity function for each process, but also handle the clustering problem. However, the question is, for NHPP its intensity λ is a function

on time which is a smooth function, not discrete vector that means sampling intensity for each NHPP may be a infinity work. So it is extremely hard to do inference, even it can be done theoretically, this would be incredibly slow.

Since the GP-DP mixture model is very complicated in the inference part and to my knowledge, no one had used this kind of model before, we can use DP mixture of Beta distribution plus CRP to approximate the GP-DP model. In [11], the author used DP mixture of Beta distribution model to estimate intensity, which can be used in estimating the intensity function of NHPP. With the estimated intensity result, a CRP model can be used for final clustering. However, when the dataset is large, this method became very slow. Considering calculation efficiency, we need to propose another model that can solve this NHPP clustering problem more efficiently, which is exactly our DTW-ddCRP model.

In DTW distance measurement, suppose we have two time series $\mathbf{x} = \{x_i\}_{i=1}^{m}$ and $\mathbf{y} = \{y_j\}_{j=1}^{n}$. To find the alignment of two sequences using DTW, we should first construct an $n * m$ matrix where the (i, j) element of this matrix is the distance $d(x_i, y_j)$ between point x_i and y_j, and this distance can be Manhattan distance or Euclidean distance or other kind of distance.Then we define a cumulative distance $D(i, j)$ as bellow:

$$D(i,j) = d(x_i, y_j) + min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\}. \tag{2}$$

DTW algorithm begins at element $d(x_1, y_1)$ and ends at element $d(x_m, y_n)$ so the final $D(m, n)$ is the DTW distance between time series $\mathbf{x}$, $\mathbf{y}$. The time complexity of DTW algorithm is $O(mn)$.

The ddCRP model determines table assignment via customer's link, which leads to a result that each customer is more likely to be clustered with other customers that are near it in an external sense. These customer assignments are generated according to the distribution below:

$$p(c_i = j | M_D, \alpha) \propto \begin{cases} f(d_{ij}) & if \quad i \neq j \\ 3.\alpha & if \quad i = j \end{cases} \tag{3}$$

Here we set up c_i that denotes the ith sequence's assignment, which is the index of sequence i with which sequence being put into the same cluster. Let d_{ij} denote the DTW distance between sequence i and sequence j. Let M_D denote the distance matrix of all the time series sequences, and f is the decay function. The decay function mediates how the distance of two data points affects the probability that they connect to each other, for example their probability of belonging to the same cluster. According to [3], there are three kind of decay functions, the window decay $f(d) = 1[d < a]$ which only considers customers at most distance a from the current customer; the exponential decay $f(d) = e^{-d/a}$ which decays the probability of exponentially link to an earlier customer with the current customer; and the logistic decay $f(d) = exp(-d + a)/(1 + exp(-d + a))$ which is a smooth version of the window decay. In general, ddCRP shows the probability of sequence i and j being in the same cluster, which conditioned on the distance measurement.

From Eq. (3) we note that the probability of which table would one customer sit at has relation with the distance, decay function and scalar α, which means that, if one customer have a connection with another one according to the customer links, they may have high probability sitting together at one table. However, Eq. (3) is only a prior, likelihood and setting of hyperparameters also contribute much to final assignments.

3.2 Model

In this section, we will give a formal and detailed description of the proposed DTW-ddCRP model on NHPP.

In the DTW-ddCRP model, our target is to estimate intensity $\lambda_i(t)$ for each NHPP i via time t by clustering different processes, so both intensity and clustering are latent. Each NHPP is described by $\mathbf{x}_i = \{x_{t_i}\}$ where each x_{t_i} is the event time for tth event in $\mathbf{x}_i$, where $t_i = 1...n_i$. DTW algorithm is used in the first stage to obtain distances d_{ij} for any pair of pipes i and j. This distance is later input into decay function f in ddCRP and the purpose is to estimate $\lambda_i(t)$. However, if we assume that the structure information has been compared in DTW, we can use the statistic information as simpler distributions instead of the stochastic process assumption. Here we assume that each component k follows a Poisson distribution with latent parameter $\theta = (\theta_1, \theta_2, ..., \theta_k)$, and corresponding prior is a gamma distribution with hyper-parameters α^*, β. Given two point process $\mathbf{x}_i$ and $\mathbf{x}_j$ $0 \leq i, j \geq N$, we assume that $c_{ij} = 1$ indicates there is a link between them and $c_{ij} = 0$ indicates the opposite. Our target is to estimate $E(C)$ where $C \in 0, 1^{N \times N}$, and each element is $c_{i,j}$. To obtain $E(C)$, the posterior distribution $P(C|X, f, \Theta)$ will be estimted via our model. Here $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ is the dataset of all point processes, f is the decay function, and Θ includes all hyper-parameters.The connection matrix C represents which processes are linked together. Here both $\mathbf{c}$ and θ are latent. Here is the generative process:

1. For all $\mathbf{x}_i$, calculate $\mathbf{M}_D = d(\mathbf{x}_i, \mathbf{x}_j)$ as DTW distance matrix for each pair of processes in dataset X;
2. For each i, a connect $c_{ij} = 1/0$ can be generated using the probability represented in (3). For all the pipes can be linked to the same pipe, we denote the set of them to be a cluster. Then we use k to index all the clusters, and $z(x_i) = k$ to represent the transfer from clusters to the index.
3. For each cluster k, draw a cluster parameter $\theta_k \sim Gamma(\alpha^*, \beta)$. Note that θ_k does not change with t so it is much simplified version to $\lambda_i(t)$. θ_k is inferred as the model variable.
4. For each water main failure burst process, generate the number of failures by $n_i \sim Poisson(\theta_k)$.

Here we can write the posterior as: $P(C|X, \mathbf{M}_D) = \int_\lambda P(C, \lambda|X, \mathbf{M}_D)$ $P(\lambda|\alpha^*, \beta)d\lambda$. The $P(C, \lambda|X, \mathbf{M}_D) \sim P(X|C, \lambda)P(C|d(\mathbf{x}_i, \cdot))$. Here $P(X|C, \lambda)$ is the likelihood, we use Poisson distribution. $P(C|d(\mathbf{x}_i, \cdot))$ is the ddCRP prior

distribution in Eq. (3). Our model can save effort as we avoided the estimation of the likelihood using Poisson process and Gaussian process as prior. We will see the details in the inference part.

3.3 Model Inference

In this section, we will give the detailed inference method for inferring model parameters from water main failure burst process.

The inference itself analytically complicate for ddCRP-based models due to the combinatorial nature in partitions and intractable structure of the model. So, for our model we used approximated inference, the Gibbs sampler [10] particularly, due to the fact that the hyperparameters on our model are conjugate priors of our model parameters.

The key part of the inference is the inference of index for each cluster and the assignment of each customer, which in our model is the assignment of non-homogeneous Poisson processes.

As introduced in Sect. 3.2, the prior is $p(c_i|D, f, \alpha)$, the likelihood of the observation is $p(x_i|z(c_{-i} \cup c_i), G_0)$, $z(c_{-i} \cup c_i)$ is the current partition, G_0 is denote the base measure. So the posterior is then:

$$p(c_i|c_{-i}, X, \Theta) \propto p(c_i|M_D, \alpha)p(X|z(c_{-i} \cup c_i), \Theta, G_0) \tag{4}$$

where Θ is the hyperparameters of our model and $\Theta = M_D, f, G_0$. As our cluster parameter θ is draw by Gamma distribution and our observations are Possion processes, so posterior became into:

$$p(c_i|c_{-i}, X, M_D, \alpha^*, \beta) \propto p(c_i|M_D, \alpha)p(X|z(c_{-i} \cup c_i), \alpha^*, \beta) \tag{5}$$

Then we can decompose the likelihood term as below:

$$p(X|z(c_{-i} \cup c_i), \Theta, \alpha^*, \beta, \lambda) = \prod_{k=1}^{|z(C)|} p(X_{z(c_{1:n})=k}|\Theta, \alpha^*, \beta, \lambda) \tag{6}$$

In Eq. 6, we define $|z(C)|$ as the number of unique clusters and $X_{z(c_{1:n})=k}$ as the set of $\mathbf{x}_i$ that are generated from cluster k. For each particular cluster, we then marginalized out the mixture component λ because the dataset of observations from each cluster are drawn independently from the same intensity, which itself is drawn from Gamma distribution, so the marginal probability is:

$$p(X|z(c_{-i} \cup c_i), \Theta, \alpha^*, \beta) = \int \prod_{k=1}^{|z(C)|} p(X_{z(c_{1:n})=k}|\Theta, \alpha^*, \beta, \lambda)d\lambda \tag{7}$$

Then comes to the sampling part, where we use Gibbs sampling [6] that is a simple form of MCMC inference [17]. The first stage is to remove or reassign the customer link c_i, where we either leave the old cluster structure inact or split the cluster that was assigned to the coming ith customer. Then the second stage is to

consider the prior probability of such new customer and corresponding changes it takes to the likelihood term. Suppose we have l and m that represent the indices that joined with index k, then resampling customer assignments from

$$p(c_i|c_{-i}, X, \Theta) \propto \begin{cases} p(c_i|M_D, \alpha)\Phi(\boldsymbol{X}, z, \alpha^*, \beta) & if \;\; c_i \; joins \; l \; and \; m, \\ p(c_i|M_D, \alpha) & otherwise, \end{cases} \tag{8}$$

where Φ is defined as:

$$\Phi(X, z, \alpha^*, \beta) = \frac{p(X_{z(C)=k}|\alpha^*, \beta)}{p(X_{z(C)=l}|\alpha^*, \beta)p(X_{z(C)=m}|\alpha^*, \beta)} \tag{9}$$

4 Empirical Study

In this section, we conduct the comparison experiments on different datasets on both synthetic data and real-world data.

4.1 Synthetic Data

In this section we are going to test model performance on synthetic data which include three steps: the first step is data generation; the second step is to cluster generated data using our DTW-ddCRP model and hyper-parameter discussion; the third step is comparing model performance with other three baseline measures and hyperparameter discussion.

Generating NHPP. The first step is to generate NHPP as our synthetic data. We generate NHPP on temporal domain T, where $T \in R^{\mathcal{D}}$, and $\lambda(t)$ denotes the intensity function, $N(\tau)$ denotes the number of events in the subregion $\tau \subset T$ of NHPP. The number of random events follows Poisson distribution where $\lambda_\tau = \int_\tau \lambda(t)dt$. According to [1], at first we need to transform the GP prior on Poisson intensities, which indeed is to use transformation of GP to form a prior distribution and then generate random events from the intensity function drawn from the GP prior. We first randomly drawing events $\{\hat{t}_j\}_{j=1}^J$ from a HPP, then define an upper bound intensity λ^* of a HPP and randomly draw the number of events J in region τ from Poisson distribution with parameters τ, then randomly and uniformly distribute the events $\{\hat{t}_j\}_{j=1}^J$ in region τ. Given the events $\{\hat{t}_j\}_{j=1}^J$, we could sample the intensity function value $\{g(\hat{t}_j)\}_{j=1}^J$ of $\{\hat{t}_j\}_{j=1}^J$ from the GP prior. Since we have already obtained value of intensity function, then we used thinning method to sample observations, initialize $t = 0$, find a constant rate function $\lambda_u(t) = \lambda_u$, let $\lambda(t)$ be the intensity function of the entire process; T_i refers to the $i - th$ event time point and it is independently deleted with a probability $1-\lambda(t)/\lambda_u(t)$; then the remained points form a NHPP with intensity function $\lambda(t)$.

Clustering NHPP. We generated two kinds of process for synthetic data experiment, one is homogeneous Poisson process (HPP) and another one is NHPP with dynamic intensities. For homogeneous Poisson process, we generated 500 sequences with 20 different intensities. Since each process has an exact intensity, so processes with the same intensity should be in the same cluster, then we got different accurate rate for different hyperparameters showing in Table 1(a). We chose the window decay function that we mentioned in Sect. 3.1 as our f and we initialized $\alpha = 0.005$, which we found worked well. Then we generated NHPP with dynamic intensities, which is also 500 sequences with 20 different intensities. Since for each sequence, the intensity is dynamical, the DTW distance measure can have a good performance in looking for the nearest neighbour that can let our model well performed for NHPP clustering. Accurate result shows in Table 1(b).

Table 1. Accurate clustering rate for HPP and NHPP

	$a = 1$	$a = 5$	$a = 10$	$a = 15$
Accurate rate	86%	77%	49%	15%

(a) HPP

	$a = 1$	$a = 5$	$a = 10$	$a = 15$
Accurate rate	87%	72%	51%	8%

(b) NHPP

As we have discussed before, the window decay function only considers customers that are at most distance a from the current customer. When a is too large, the result of window decay function $f(d)$ will be 1 for most customers which leads to a result that most of customers will be clustered into a same cluster. So normally the setting of parameter a could influence final result significantly. However, for sparse observation processes, a can be set a little larger in order to achieve a more accurate clustering outcome.

We used three methods as baseline: the first one is the DP mixture of Beta distribution on CRP; the other two methods are traditional clustering methods which are divisive hierarchical clustering and k-means respectively. For DP mixture of Beta distribution based CRP model, we use Beta DP mixture model to estimate the intensity of each failure process, then clustering with CRP. For k-means and hierarchical we selected median performance among the whole possible condition settings. We discussed model performance by comparing in-cluter-distance and between-cluster-distance between our DTW-ddCRP model with baseline measure showing in Fig. 1.

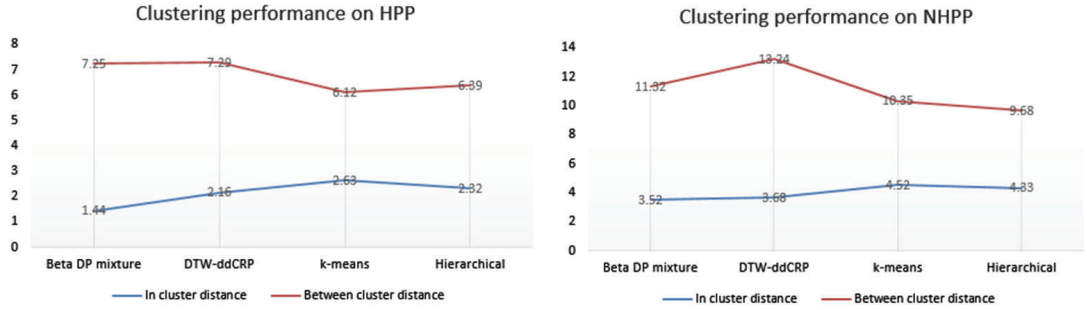

Fig. 1. Model performance comparing with other measures

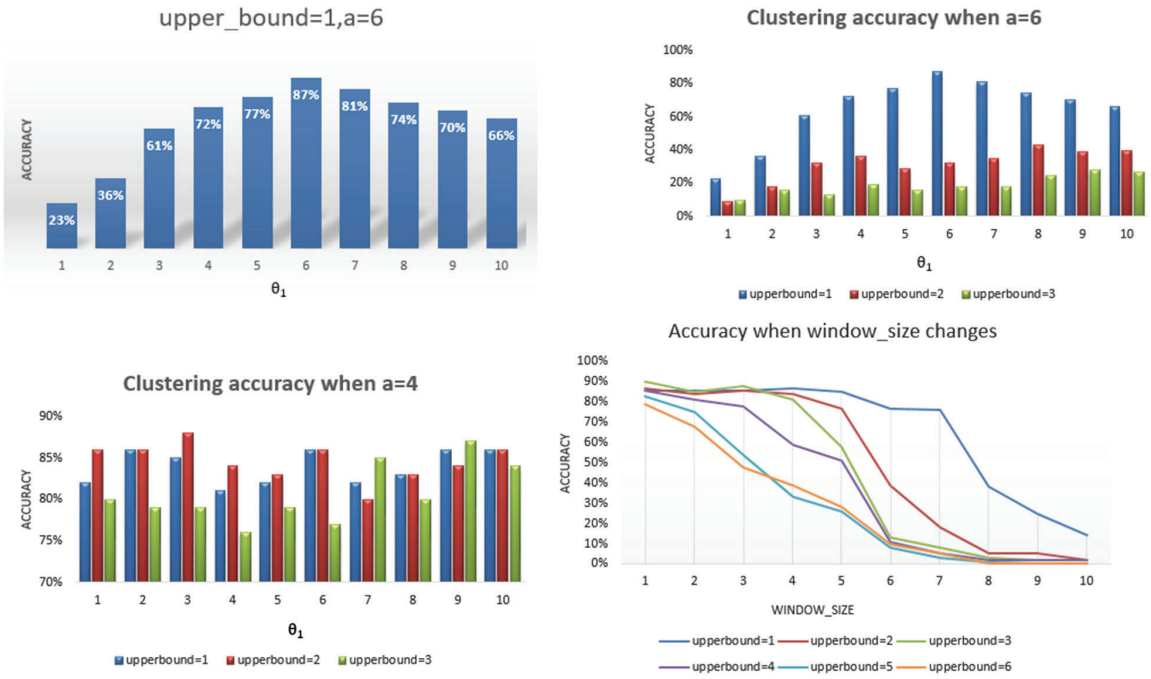

Fig. 2. Clustering accuracy with different hyper-parameters

Hyper-parameter Discussion. We tested all the influence from hyper-parameters in our model, after hundreds of experiments, we found out that several hyper-parameters have great influence on our model performance and others do not, so here we mainly discuss those have great impacts.

In Fig. 2, *upperbound* is a parameter that controls the intensity of each NHPP, the higher of *upperbound* the higher of point density in NHPP. We found that when *upperbound* is too high it may cause a fact that even two NHPP have different intensity function may have too many points which leads to a very low DTW distance, in others words they shows a similar structure information though they do not have same intensity function. Another important parameter is θ_1, which is a parameter that influence the structure of intensity function of NHPP. When θ_1 increases, structure of intensity function changes from volatile to smooth, according to which we found both too low and too high of θ_1 would leads to a bad performance when the other parameters remain the same. Another

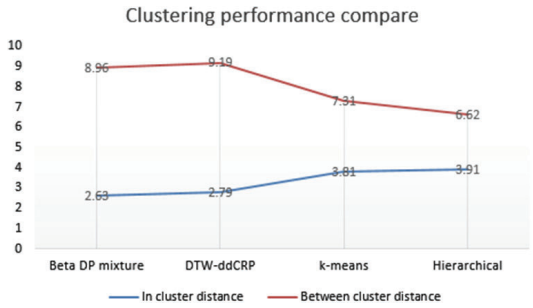

Fig. 3. Model performance comparing with baseline measures

important parameter is the size a of window decay function, an appropriate value can lead us to a perfect clustering result. Generally speaking, when our data set is sparse we need a high value of window size a.

4.2 Real World Data

We use water main failure data as real world data for empirical study. In order to expound water main failure problem, we first need to introduce water main failure data itself. We selected water main failure data that records all the failure events happened from the beginning of 2001 to the end of 2016, which includes over 80 thousand failure events. For each water main, failure event had happened more than once in the sixteen years. We set one month as a time interval, for one water main if failure event happened, log 1 for the certain cell and 0 for those cells that no failure event happened, then for each water main there is a point process which is a NHPP. What we did is to cluster the water mains by clustering these NHPP, after which we can get a general intensity for each water main that can be used to predict probable time the next failure event happens. Then our water main failure problem became NHPP clustering problem, which we have introduced in the introduction section. The original dataset collected 80315 failure events from 2001 to 2016, transform into point process, there came out with 2450 point processes. We found that for some water main, failure process is very sparse, as a result there DTW distance came out to be very small that can not be used to accurately analyze water mains' pattern so we removed those failure processes that the total number of failure events was lower than 10, then we got 593 failure processes.

Model performance comparing on real data is similar with on synthetic data, for DP mixture of Beta distribution based CRP model, we use Beta DP mixture model to estimate the intensity of each failure process then clustering with CRP, and for k-means and hierarchical we use median value. We compared the mean value of in-cluter-distance and between-cluster-distance between our DTW-ddCRP model with baseline measure showing in Fig. 3.

From our experiments we found out that our proposed DTW-ddCRP model has similar performance with the Beta DP mixture model and other two traditional clustering methods. However, our is much more efficiency, especially when

the number of dataset is larger and process itself is longer, and we do not need to give any cluster number or clustering restriction as k-means and hierarchical do.

5 Conclusions

In this paper, we proposed a DTW-ddCRP model that can discover the latent intensity pattern of water mains' failure events and realize clustering of the water mains. The proposed model has a distance dependent Bayesian nonparametric prior over NHPP, and DTW similarity measure to reflect relationship between customer to customer, while the assignment is governed by a ddCRP clustering measure. With such construction, the water mains with similar burst pattern can be clustered together with a much higher efficiency. Besides, we do not need to preset the number of clusters for water mains, which is difficult in unsupervised learning, especially for real world data. Instead, our proposed model can automatically generate the cluster number from the provided data.

The empirical study shows expecting outcome, suggesting that our model can well discover the latent intensity pattern of water mains' failure burst process and the obtained result can be use to make accurate prediction in the certain domain, indicating those water mains that needs to be checked and reduce the burst risk.

For the future work, speed improving methods for the distance learning part [16] can be add to improve on clustering efficiency. Meanwhile we will test how to using Bayesian clustering method for continuous sequences, maybe the soft-DTW [5] measure can be added to implement this idea.

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. ACM (2009)
2. Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y.: Time-series clustering-a decade review. Inform. Syst. **53**, 16–38 (2015)
3. Blei, D.M., Frazier, P.I.: Distance dependent chinese restaurant processes. J. Mach. Learn. Res. **12**(Aug), 2461–2488 (2011)
4. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. J. ACM (JACM) **57**(2), 7 (2010)
5. Cuturi, M., Blondel, M.: Soft-DTW: a differentiable loss function for time-series. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 894–903. JMLR.org (2017)
6. Geler, Z., Kurbalija, V., Radovanović, M., Ivanović, M.: Impact of the Sakoe-Chiba band on the DTW time series distance measure for kNN classification. In: Buchmann, R., Kifor, C.V., Yu, J. (eds.) KSEM 2014. LNCS (LNAI), vol. 8793, pp. 105–114. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12096-6_10
7. 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)

8. Han, J., Pei, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, Amsterdam (2011)
9. Hardiman, S.J., Bercot, N., Bouchaud, J.P.: Critical reflexivity in financial markets: a hawkes process analysis. Eur. Phys. J. B **86**(10), 442 (2013)
10. Hrycej, T.: Gibbs sampling in bayesian networks. Artif. Intell. **46**(3), 351–363 (1990)
11. Kottas, A.: Dirichlet process mixtures of beta distributions, with applications to density and intensity estimation. In: Workshop on Learning with Nonparametric Bayesian Methods, 23rd International Conference on Machine Learning (ICML) (2006)
12. Liao, T.W.: Clustering of time series data–a survey. Pattern Recogn. **38**(11), 1857–1874 (2005)
13. Mueen, A., Keogh, E.: Extracting optimal performance from dynamic time warping. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2129–2130. ACM (2016)
14. Niekum, S.: A brief introduction to Bayesian nonparametric methods for clustering and time series analysis. Technical report CMU-RI-TR-15-02, Robotics Institute, Carnegie Mellon University (2015)
15. Petitjean, F., Forestier, G., Webb, G.I., Nicholson, A.E., Chen, Y., Keogh, E.: Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm. Knowl. Inf. Syst. **47**(1), 1–26 (2016)
16. Rakthanmanon, T., et al.: Searching and mining trillions of time series subsequences under dynamic time warping. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 262–270. ACM (2012)
17. Robert, C., Casella, G.: Monte Carlo Statistical Methods. Springer, Heidelberg (2013)
18. Wolff, R.W.: Poisson arrivals see time averages. Oper. Res. **30**(2), 223–231 (1982)
19. Zhang, B., Zhang, L., Guo, T., Wang, Y., Chen, F.: Simultaneous urban region function discovery and popularity estimation via an infinite urbanization process model. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2692–2700. ACM (2018)