Fast Search for Distance Dependent Chinese Restaurant Processes

Weiwei Feng[1], Peng Wang[1], Chuan Zhou[1], Peng Zhang[2,1], and Li Guo[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100093, China
[2]Quantum Computation and Intelligent Systems, University of Technology, Sydney (UTS), Australia
{fengweiwei,zhouchuan,guoli}@iie.ac.cn, peng860215@gmail.com

ABSTRACT

The distance dependent Chinese Restaurant Processes (dd-CRP), a nonparametric Bayesian model, can model distance sensitive data. Existing inference algorithms for dd-CRP, such as Markov Chain Monte Carlo (MCMC) and variational algorithms, are inefficient and unable to handle massive online data, because posterior distributions of dd-CRP are not marginal invariant. To solve this problem, we present a fast inference algorithm for dd-CRP based on the A-star search [3]. Experimental results show that the new search algorithm is faster than existing dd-CRP inference algorithms with comparable results.

Categories and Subject Descriptors H.2.8 [Database Management]: Database Applications - Data Mining
General Terms Theory, Algorithms, Performance
Keywords A-star search, inference, nonparametric Bayesian, distance dependent Chinese Restaurant Processes.

1. INTRODUCTION

With the fast development of online applications, such as social networks and E-commerce, new scalable data mining models that can digest massive data efficiently are urgently needed[6]. Among these models, clustering analysis is one of the key tools which has been widely used in user behavior analysis, topic modeling, outliers detection, to name a few.

The distance dependent Chinese Restaurant Processes (dd-CRP) proposed recently by Blei et. al. [2] is a new nonparametric Bayesian model for unveiling latent clusters behind non-exchangeable text data. The dd-CRP can discover distance sensitive clusters and auto-select the proper number of clusters. So it is markedly useful for modeling temporal and spatial dependent data such as news stories and user behaviors.

Existing work on inferring posterior distributions of dd-CRP relies on Gibbs sampling [4] and variational inference [1]. However, both methods are inefficient. Because dd-CRP requires to infer data-linkage instead of cluster assignments, the potential values of latent variables are large. On the other hand, the posterior of latent variables are not marginally invariant, so variational inference for dd-CRP is complex and parallel techniques are inapplicable.

To this end, we present an efficient inference method based on the A-star search [3] for dd-CRP. A-star search has been widely used in path-finding [5]. Daume et al [3] demonstrated that latent states of Bayesian models can be treated as points in a latent space. Thus, in our search-based inference, the optimal posterior of dd-CRP is taken as a special point with maximum data likelihood that can be restored by the A-star search. We use three heuristic cost functions to accelerate the searching process. We empirically demonstrate that the search-based inference can achieve comparable cluster results (*w.r.t.* data likelihood) and significantly faster than state-of-the-art methods.

2. THE A-STAR SEARCH FOR DD-CRP

2.1 Model Description

The generation process of dd-CRP is to generate links between data. Each data record can be linked either to itself or to another data record. In the former case, dd-CRP creates a new cluster. In the latter case, the linked data records (either directly or indirectly linked) are assigned to the same cluster. The probability of data linkages in dd-CRP is determined by the distance between data which can be denoted by a matrix $D = \{d_{i,j} | i, j = 1, \cdots, N\}$ and a decay function $f(d_{ij})$, as in Eq. (1),

$$p(\mathbf{c}_i = j | D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if } i \neq j \\ \alpha & \text{if } i = j \end{cases} \qquad (1)$$

where α is the concentration factor, and $\mathbf{c}_i = j$ denotes that there is a link from data record j to i. So the cluster assignments $\mathbf{z}(\mathbf{c}_{1:N})$ can be derived from $\mathbf{c}_{1:N}$, and the mixture topic model based on dd-CRP can be described as follows:

1. For each document $i \in [1, \cdots, N]$ draw assignment $\mathbf{c}_i \sim dd - CRP(\alpha, f, D)$;
2. For each cluster $k \in \{1, \cdots\}$ draw parameter $\phi_k \sim H$;
3. For each document $i \in [1, \cdots, N]$ draw $\mathbf{w}_i \sim F(\phi_{\mathbf{z}(\mathbf{c})_i})$.

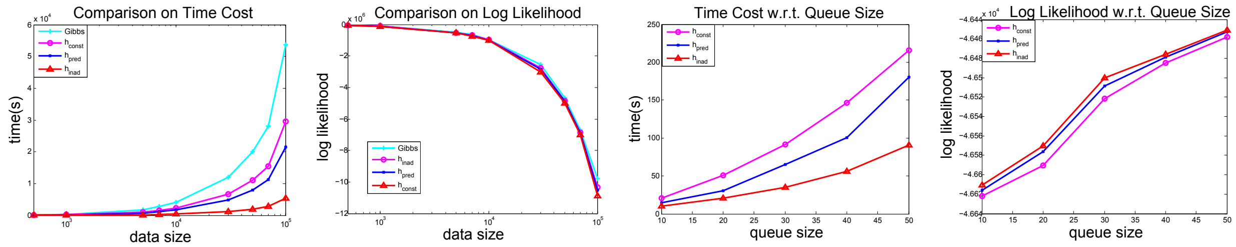
2.2 The A-star Search for inference

A-star is a best-first search. In each step, it explores the space from the point which seems to be closest to the optimal point. For dd-CRP, the state point is data linkage $\mathbf{c}_{1:N_0}$. N_0 is the number of points processed in the state. $s(\mathbf{c}_{1:N_0})$ is a knowledge-plus-heuristic cost function of the state, which measures how close the state to the optimal posterior distribution. For dd-CRP, $s(\mathbf{c}_{1:N_0})$ is the likelihood of the state, which is the sum of two functions:

- The existing state-cost function $g(\mathbf{c}_{1:N_0}, \mathbf{w}_{1:N_0})$, which is the data likelihood for states $P(\mathbf{w}_{1:N_0} | \mathbf{z}(\mathbf{c}_{1:N_0}), H)$;
- The heuristic state-cost function $h(\mathbf{c}_{1:N_0}, \mathbf{w}_{N_0+1:N})$, which is the heuristic estimate of the data likelihood $P(\mathbf{w}_{N_0+1:N} | \mathbf{z}(\mathbf{c}_{1:N_0}), H)$.

From the initial state, A-star search maintains a priority queue of states, where the priority of states is determined

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742746.



(a) (b) (c) (d)

Figure 1: Experimental results.

by $f(\mathbf{c}_{1:N_0})$. We keep the size of the queue fixed by removing the states having the lowest priority. However, due to the limited memory, the states which can potentially reach the optimal may be also pruned. Thus, the A-star search does not always guarantee the optimal posterior distribution. The A-star search inference is summarized in Algorithm 1.

Algorithm 1: A-star Search for DD-CRP.

```
Input: observations $\mathbf{w}_{1:N}$, a heuristic function $h$, queue
       size $B$, the distance matrix $D$, the decay
       function $f(\cdot)$
Output: clustering result $\mathbf{z}_{1:N}$
Initialize Queue $Q$: $Q \leftarrow [(\mathbf{c}_1 = 1)]$;
while $Q$ is not empty do
    Remove state $\mathbf{c_{1:N^0}}$ from the front of $Q$;
    if $N_0 = N$ then
        return $\mathbf{z}(\mathbf{c_{1:N}})$;
    for $\{\mathbf{c}_{N_0+1} = j | \forall f(d_{N_0+1,j}) > 0 \cup \{N_0 + 1\}\}$ do
        $\mathbf{c^{new}} = \mathbf{c_{1:N_0}} \oplus \mathbf{c}_{N_0+1}$;
        compute the score:
        $s = g(\mathbf{c}^{new}, \mathbf{w}_{1:N_0+1}) + h(\mathbf{c}^{new}, \mathbf{w}_{N_0+2:N})$;
        update queue: $Q \leftarrow (\mathbf{c}^{new}, s)$;
    if $B < \infty$ and $|Q| > B$ then
        Shrink queue: $Q \leftarrow Q_{1:B}$;
```

Here, $g(\mathbf{c^{new}}, \mathbf{w}) = P(\mathbf{w}_{1:N_0} | \mathbf{z}(\mathbf{c}_{1:N_0}), H)$ is defined as:

$$g(\mathbf{c^{new}}, \mathbf{w}) = \prod_k \int \Big(\prod_{i \in \mathbf{z}(\mathbf{c}^{new})=k} P(x_i|\phi_k) \Big) P(\phi_k|H) d\phi_k \quad (2)$$

The heuristic function $h(\cdot)$ will influence the search speed. The closer the estimation of $P(\mathbf{w}_{N_0+1:N} | \mathbf{z}(\mathbf{c}_{1:N_0}), H)$ is, the faster the algorithm will be. We propose three heuristic functions.

- **Constant heuristic function $h_{const}(\cdot)$.** We can neglect the heuristic cost by setting $h_{const}(\mathbf{c}_{1:N_0}) \equiv 1$.
- **Predictive heuristic function $h_{pred}(\cdot)$.** We can obtain a tighter heuristic function by calculating the probability distribution of unclustered data points given the clusters $\{\phi_1, \cdots, \phi_K\}$ which are derived from $z(\mathbf{c}_{1:N_0})$.

$$h_{pred}(\mathbf{c}_{1:N_0}) = \prod_{n=N_0+1}^{N} \max_{1 \leq k \leq K^0+1} P(\mathbf{z}_n = k) P(\mathbf{w}_n | \phi_k, \mathbf{w}_n = k) \quad (3)$$

- **Inadmissible heuristic function $h_{inad}(\cdot)$.** The heuristic cost is calculated by assigning each data in $\mathbf{w}_{N_0+2:N}$ a new cluster.

$$h_{inad}(\mathbf{c}_{1:N_0}) = \prod_{n=N_0+1}^{N} \int P(\mathbf{w}_n|\phi) P(\phi|H) d\phi \quad (4)$$

The estimation is even tighter. However, $h_{inad}(\cdot)$ is inadmissible[3], which implies even with infinite memory, the optimal posterior is not guaranteed.

3. EXPERIMENTS

Experimental settings. We use the synthetic data set [2] with varying data volumes. The hyper-parameters of dd-CRP are set as $\{\alpha = 1, H = 5\}$. We use window decay for $f(\cdot)$ and D is measured by time differences between data.

Results. We compare the A-star search inference (with 3 different heuristic functions) with the Gibbs sampling algorithm *w.r.t.* time cost and log likelihood. The results are depicted in Fig.1 (a)(b). Our methods are significantly faster than Gibbs sampling. For different heuristic functions, h_{inad} leads to the fastest search, while the clustering results *w.r.t.* likelihood are similar. We compare the time cost and clustering results *w.r.t.* queue size B as depicted in Fig.1 (c)(d). With the larger queue size, we can obtain better clustering results but more time cost. The results accord with the analysis in section 2.2.

4. CONCLUSIONS

We presented a fast search-based inference algorithm for dd-CRP. This inference uses the A-star search [3] to find the optimal posterior distribution. Our method is significantly faster than state-of-the-art inference algorithms and achieves comparable clustering results.

Acknowledgement. This work was supported by the NSFC (No. 61370025), and the Strategic Leading Science and Technology Projects of CAS (No.XDA06030200), 973 project (No. 2013CB329605) and Australia ARC Discovery Project (DP140102206).

5. REFERENCES

[1] S. Bartunov and D. Vetrov. Variational inference for sequential distance dependent chinese restaurant process. In *Proc. of ICML 2014*, pages 1404–1412, 2014.

[2] D. M. Blei and P. I. Frazier. Distance dependent chinese restaurant processes. *JMLR*, 12:2461–2488, 2011.

[3] H. Daumé III. Fast search for dirichlet process mixture models. *arXiv preprint arXiv:0907.1812*, 2009.

[4] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *JCGS*, 9(2):249–265, 2000.

[5] W. Zeng and R. Church. Finding shortest paths on real road networks: the case for a*. *IJGIS*, 23(4):531–543, 2009.

[6] P. Zhang, C. Zhou, P. Wang, B. J. Gao, X. Zhu, and L. Guo. E-tree: An efficient indexing structure for ensemble models on data streams. *TKDE*, 27(2):461–474, 2015.