Distance dependent Maximum Margin Dirichlet Process Mixture

Quan Nguyen[1], Mikko Lauri[1], and Simone Frintrop[1]

Department of Informatics, University of Hamburg, Germany
{nguyen,lauri,frintrop}@informatik.uni-hamburg.de

Abstract. We propose distance dependent maximum margin Dirichlet Process Mixture (STANDPM), a nonparametric Bayesian clustering model that combines distance-based priors with the discriminatively learned likelihood of the Maximum Margin Dirichlet Process Mixture. STANDPM generalizes the distance-based prior introduced in the distance dependent Chinese Restaurant Process for non-sequential distances and allows modeling of complex dependencies between data points and clusters. The generalized distance-based prior is formulated as an abstract similarity measurement between a data point and a cluster. Empirical results show that the STANDPM model with abstract similarity achieves state-of-the-art performances on a number of challenging clustering datasets.

Keywords: Dirichlet Process Mixture models · Chinese Restaurant Process · Gibbs sampling · Probabilistic clustering · Uncertainty modelling.

1 Introduction and Related Work

Cluster analysis is an unsupervised technique that allows organizing a dataset into groups of similar data points. It has been widely applied to application domains such as data mining [King, 2014] and image segmentation [Achanta et al., 2012]. Due to the absence of prior information, a number of assumptions on the characteristics of the clusters, such as the number of clusters or the data generation process, are often needed to facilitate the clustering process. Reducing the number of such assumptions and explicitly modeling the uncertainty about the clusters have been one of the main research goals of this field. Several nonparametric Bayesian methods have been proposed to model the uncertainty about the number of clusters [Zhu et al., 2011], the generative model [Neal, 2000] or the hierarchy of the clusters [Heller and Ghahramani, 2005]. One particular class of methods that has remained popular in the last two decades is the Dirichlet process mixture (DPM) models. The DPM models can be constructed using the Chinese Restaurant Process (CRP) [Aldous, 1985] which specifies a prior distribution on the structure of the clusters. Viewing clusters as tables and data points as customers in a restaurant, CRP can be intuitively described as follows: a number of customers sequentially enter a restaurant, and each of them chooses a table to sit at. If a customer sits at a table, he is said to have a *link*

from himself to the table. Each customer can choose to link to an existing table with a probability proportional to the number of customers already sitting at that table, or he can start a new table with a probability proportional to a scaling parameter. The CRP prior works well when the dependencies between customers have no impact on the table configurations. A more complex model for dealing with complex temporal and spatial dependence between data points is the distance dependent Chinese Restaurant Process (ddCRP) [Blei and Frazier, 2011]. In ddCRP, a customer is linked to a customer instead of a table. In other words, a customer decides on another customer to sit with instead of a table to sit at. The prior probability of linking two customers depends on their pairwise distances.

DPM models often rely on Markov Chain Monte Carlo sampling techniques such as Gibbs sampling to approximate the posterior distribution of clusters. To keep the posterior inference tractable, DPM models are often limited by a number of assumptions such as conjugate prior. It is also challenging for these models to estimate the component parameters of the mixtures, especially for mixtures of high dimensional data [Chen et al., 2016]. One recent model that attempts to overcome these limitations of CRP-based DPM models is the Maximum Margin Dirichlet Process Mixture (MMDPM) model [Chen et al., 2016]. The MMDPM model turns the Gibbs sampling process into an online learning process that efficiently learns the component parameters of the clusters from high dimensional data. The learning process of MMDPM requires each cluster to be represented explicitly by a vector of component parameters. In contrast, ddCRP-based DPM models require an implicit representation of the clusters so that the clusters can be merged together during the Gibbs sampling process [Blei and Frazier, 2011]. Consequently, it has remained an open question of how to combine a distance-based prior and a discriminatively learned likelihood.

This paper proposes diSTANce dependent maximum margin Dirichlet Process Mixture (STANDPM) model, a nonparametric Bayesian clustering model that combines the usage of pairwise distances in the ddCRP prior and the learning process for the likelihood in the MMDPM. The central idea of STANDPM is to establish the link from a data point to a cluster via the links from that data point to the existing data points of that cluster. This technique allows STANDPM to use the pairwise distances between data points in the prior and avoid merging clusters during Gibbs sampling. The probability of linking a data point to a cluster in STANDPM depends on a similarity measurement between a data point and a cluster, denoted as *abstract similarity*. The term "abstract" refers to the fact that this similarity measurement is generally not a true distance metric. Figure 1 illustrates an example of the establishment of the links from a data point to two clusters during Gibbs sampling. The abstract similarity allows integrating domain knowledge through the prior function rather than direct regularization of the posterior inference process as in [Chen et al., 2014]. Our work here also differs from the distance-based priors in [Dahl, 2008] because the prior distribution in our model does not require normalization of the customer assignment probabilities. We focus on the integration of pairwise

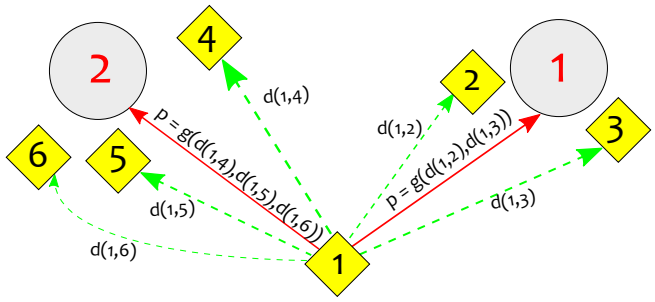

Fig. 1. Visualization of the probabilistic clustering process. The customers (data points) and tables (clusters) are illustrated by squares and circles, respectively. Customer 1 is choosing a table to sit. Customer 2 and 3 are sitting at table 1. Customer 4, 5, 6 are sitting at table 2. The prior probability $p(.)$ (solid arrows) of linking customer 1 to a table is proportional to the abstract similarity $g(.)$ between customer 1 and that table. The abstract similarity is a function of the distances (dashed arrows) from customer 1 to all existing customers of each table.

distances into the posterior inference using Gibbs sampling instead of variational Bayesian techniques [Zhu et al., 2011]. This paper has two main contributions:

1) A novel, theoretically justified prior function which allows designing DPM models that have the distance-based clustering effects of ddCRP prior without sacrificing the efficient maximum margin likelihood of MMDPM.

2) New state-of-the-art clustering performances for Bayesian methods on a number of large, high-dimensional datasets without being given the number of true clusters *a priori*.

2 Unsupervised clustering with DPM models

We consider the following problem: Let a set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ of N data points in $\mathbb{R}^D$ generated from an unknown number of mixtures in a mixture model and a pairwise distance function $d(i,j)$ measuring the dissimilarity between every pair of two data points be given. Estimate the number of mixtures and generate their corresponding clusters of data points.

The DPM model provides a Bayesian framework for modelling the posterior distribution over all possible clusterings of a dataset. In this framework, the data points can be seen as samples generated from K mixtures in a Dirichlet process mixture model $\mathrm{DP}(G_0, \alpha)$ with symmetric DP prior [Chen et al., 2016]:

$$\begin{aligned} &\boldsymbol{\pi}|\alpha \sim Dir(\alpha/K, \ldots, \alpha/K) \\ &z_i|\boldsymbol{\pi} \sim Discrete(\pi_1, \ldots, \pi_K) \\ &\boldsymbol{\theta}_k \sim G_0 \\ &\mathbf{x}_i|z_i, \{\boldsymbol{\theta}_k\}_{k=1}^K \sim p(\mathbf{x}_i|\boldsymbol{\theta}_{z_i}) \end{aligned} \tag{1}$$

where G_0 is the base distribution, α is the concentration parameter, $\boldsymbol{\pi}$ is the mixture weights, z_i is the cluster indicator and $\boldsymbol{\theta}_k$ is the parameter vector of cluster k.

While the exact computation of the posterior distribution is often intractable, it can be estimated by the samples generated using Gibbs sampling [Neal, 2000]. Data point $\mathbf{x}_i$ belongs to cluster k if there is a link $z_i = k$ between them. Denote the vector of all such links from all data points by $\mathbf{z}$. Gibbs sampling samples a cluster assignment of each data point from a conditional posterior distribution over z_i, keeping all other variables fixed. The conditional posterior in DPM is

$$\begin{aligned} &p(z_i = k|\mathbf{z}_{-i}, \mathbf{x}_i, \{\boldsymbol{\theta}_k\}_{k=1}^K, \alpha) \\ &= p(z_i = k|\mathbf{z}_{-i}, \{\boldsymbol{\theta}_k\}_{k=1}^K, \alpha) p(\mathbf{x}_i|z_i = k, \{\boldsymbol{\theta}_k\}_{k=1}^K) \\ &= p(z_i = k|\mathbf{z}_{-i}, \alpha) p(\mathbf{x}_i|\boldsymbol{\theta}_k) \end{aligned} \tag{2}$$

where $\mathbf{z}_{-i}$ is the link vector $\mathbf{z}$ excluding the ith element. Above, $p(z_i = k|\mathbf{z}_{-i}, \alpha)$ is the prior and $p(\mathbf{x}_i|\boldsymbol{\theta}_k)$ is the likelihood of a data point $\mathbf{x}_i$.

2.1 The distance-dependent CRP prior

Distance-based priors for DPM models have been studied extensively for their advantages in modelling complex relationships between data points. The popular ddCRP [Blei and Frazier, 2011] uses the pairwise distances $d(i, j)$ to model the prior probability of linking two data points i and j. Note that ddCRP does not model the links from data points to clusters. Given a DPM model with ddCRP prior, when using Gibbs sampling for posterior inference, the action of merging two clusters is inevitable: any new link between two data points in two clusters will merge two clusters together, resulting in a larger cluster containing all data points from the two original clusters. Merging two clusters is possible in ddCRP because the components of the new cluster can be computed directly from all the data points in the two original clusters without taking into account the components of the two original clusters.

We denote a link from data point i to data point j by $c_i = j$. Let $\mathbf{c}$ be the vector of size N of all such links. The ddCRP prior probability of establishing a link $c_i = j$ is

$$p(c_i = j|\mathbf{c}_{-i}, \alpha) \propto f(d(i, j)) \tag{3}$$

where f is a decay function for controlling how distances affect the distribution over clusterings [Blei and Frazier, 2011] and $\mathbf{c}_{-i}$ is the vector $\mathbf{c}$ excluding assignment of data point i.

2.2 The maximum margin likelihood

Computing the likelihood for high dimensional data has been a long standing problem because of the intractability in computing the normalizing constants and

updating the parameters of the likelihood distribution [Blei and Jordan, 2005]. The recent MMDPM model [Chen et al., 2016] solves this problem by replacing the generative model in DPM with a discriminative SVM classifier for learning the cluster components. Whenever a data point is assigned to or removed from a cluster, the cluster components are updated with respect to that data point only instead of re-computing the components from all data points in the cluster. The MMDPM model uses the vector of components $\boldsymbol{\theta}_k$ as an explicit representation for cluster k. The likelihood of linking a data point i to cluster k is

$$p(\mathbf{x}_i|\boldsymbol{\theta}_k) \propto \exp(\mathbf{x}_i^T \boldsymbol{\theta}_k - \lambda \left\|\boldsymbol{\theta}_k\right\|^2), \tag{4}$$

where λ is a regularization hyper-parameter to avoid trivial clustering results and control the separation between clusters. The MMDPM uses the standard CRP prior which relies only on the size of the clusters:

$$p(z_i = k|\mathbf{z}_{-i}, \alpha) = \frac{n_{-i,k}}{Z} \tag{5}$$

where $n_{-i,k}$ is the number of data points in cluster k excluding data point i, $Z = N - 1 + \alpha$ is the normalization factor. When $n_{-i,k} = 0$, $p(z_i = k|\mathbf{z}_{-i}, \alpha) = \alpha/Z$. In contrast to ddCRP, the CRP prior links data points directly to clusters.

2.3 Combining the ddCRP prior and the MMDPM likelihood

We focus on the question of using the prior in Equation (3) for the conditional posterior in Equation (2). During each iteration of the Gibbs sampling, the maximum margin online learning process in MMDPM takes each data point and its sampled cluster as a training example. In this training example, the data point is an input and the sampled cluster is the expected output. The component parameters $\{\boldsymbol{\theta}_k\}_{k=1}^K$ are updated only if the signed margin from the data point to its sampled cluster is not maximal among all clusters.

If the ddCRP prior were to be used directly, whenever two clusters are merged, the margins from the data points in the two clusters to all other clusters would have to be computed to simulate the process of linking all these data points to a new cluster. This would add a significant computational cost of $O(N^2 \times K \times D)$ to each iteration of the Gibbs sampling. Furthermore, this merging operation has a reverse effect to the learning process which tries to separate the clusters as much as possible. Since the clusters in MMDPM are explicitly represented and separately learned, merging them directly would be impractical. In this paper, we tackle this challenge by adapting the prior function to the max-margin likelihood learning process.

3 Distance dependent MMDPM

We present the STANDPM model as an extension of the MMDPM model by replacing the CRP prior in MMDPM with an abstract similarity function. The STANDPM model resolves the challenge of combining the ddCRP prior and

the MMDPM likelihood outlined in Subsection 2.3. We show that the abstract similarity generalizes the clustering effects in CRP and ddCRP priors.

3.1 Abstract Similarity in STANDPM

Let $\{\mathbf{x}\}^k$ denote the subset of data points currently linked to cluster k. In our new STANDPM model, the prior term in Equation (2) is expressed as a function g of the data point i and $\{\mathbf{x}\}^k$:

$$p(z_i = k|\mathbf{z}_{-i}, \alpha) \propto g(\mathbf{x}_i, \{\mathbf{x}\}^k)\,, \tag{6}$$

where $g\colon \mathbb{R}^D \times \mathcal{X} \to \mathbb{R}^+$ is a non-negative function, $\mathcal{X}$ is a set of points in $\mathbb{R}^D$. If the cluster is new, this prior is set to $\frac{\alpha}{N}$. When $g(\mathbf{x}_i, \{\mathbf{x}\}^k)$ is a function of the pairwise distances, the link z_i is considered being drawn from a distance dependent prior.

We keep the likelihood in our model the same as in MMDPM. Substituting the prior in Equation (6) and the likelihood in Equation (4) into Equation (2), we obtain the general form of the conditional posterior in STANDPM as

$$\begin{aligned} p(z_i = k|\mathbf{z}_{-i},&\mathbf{x}_i, \{\boldsymbol{\theta}_k\}_{k=1}^K, \alpha) \\ &\propto g(\mathbf{x}_i, \{\mathbf{x}\}^k) \exp(\mathbf{x}_i^T \boldsymbol{\theta}_k - \lambda \left\|\boldsymbol{\theta}_k\right\|^2)\,. \end{aligned} \tag{7}$$

The normalization factor for the prior distribution is

$$Z_i = \frac{\alpha}{N} + \sum_{k=1}^{K} g(\mathbf{x}_i, \{\mathbf{x}\}^k)\,. \tag{8}$$

Note that the normalization changes for different $\mathbf{x}_i$. The hyper-parameter α controls the probability of generating a new cluster. In practice, this normalization factor does not need to be computed since in each iteration of the Gibbs sampling, each data point is only processed once.

The function g can be selected so that if the data point i is similar to the set of data points in cluster k, then $g(\mathbf{x}_i, \{\mathbf{x}\}^k)$ will be large and vice versa. In this setting, g can be can be any non-negative function that expresses a similarity measurement from a data point to a cluster of data points, enabling the model to leverage possible structural information of a collection of data points. The prior function of this STANDPM model is defined entirely by the distances between data points and clusters.

Abstract similarity based on the nearest data point in a cluster: One possible choice for the abstract similarity is a function of the minimum distances from a data point to the members of a cluster. This setting brings the neighborhood effect in which only a small and diverse set of the nearest neighbors of a data point influences its decision on which cluster to join, thereby encouraging

clusters of different sizes instead of always favoring large clusters as the CRP prior.

This similarity measurement from a data point to a cluster is defined as

$$g_{\max}(\mathbf{x}_i, \{\mathbf{x}\}^k) = f(\min_{\mathbf{x}_j \in \{\mathbf{x}\}^k} d(i,j)) \tag{9}$$

The subscript max reflects the fact that since f is a decreasing function, $f(\min_{\mathbf{x}_j \in \{\mathbf{x}\}^k} d(i,j)) = \max_{\mathbf{x}_j \in \{\mathbf{x}\}^k} f(d(i,j))$. We denote an STANDPM model with this $g_{\max}$ function *STANDPM-max*.

Abstract similarity based on all data points in a cluster: This setting directly integrates the clustering effect of the ddCRP prior into MMDPM. In this case, the abstract similarity is a summation over all the links from a data point to the members of a cluster:

$$g_{sum}(\mathbf{x}_i, \{\mathbf{x}\}^k) = \sum_{\mathbf{x}_j \in \{\mathbf{x}\}^k} f(d(i,j)) \tag{10}$$

An STANDPM model with the abstract similarity represented by g_{sum} function is denoted as *STANDPM-sum*.

3.2 Abstract similarity generalizes ddCRP and CRP priors

The CRP and ddCRP priors can both be constructed from the abstract similarity g_{sum}. Recall that in each iteration of the Gibbs sampling, each data point i is processed once to select a cluster k among K existing clusters for linking.

Proposition 1 *Let a DPM model with the CRP prior, a uniform distance function $d(i,j) = \ln(Z)$ and an exponential decay function $f(d) = \exp(-d)$, where Z is the normalization factor in Equation (5), be given. Then for the data point i and cluster k being processed in an iteration of the Gibbs sampling, $g_{sum}(\mathbf{x}_i, \{\mathbf{x}\}^k)$ is equal to the CRP prior probability of i linking to k.*

This proposition is a straightforward result of applying the given distance and decay functions to Equation (10) which yields Equation (5).

Proposition 2 *Let a DPM model with the ddCRP prior, any distance function $d(i,j)$ and decay function $f(d)$ be given. Then for the data point i and cluster k being processed in an iteration of the Gibbs sampling, $g_{sum}(\mathbf{x}_i, \{\mathbf{x}\}^k)$ is proportional to the ddCRP prior probability of i linking to k.*

Proof. Data point i will be linked to cluster k if it is linked to any data point already in that cluster. The probability of i linking to k is the sum of probabilities of links from i to any j in $\{\mathbf{x}\}^k$:

$$
\begin{aligned}
p(z_i = k|\mathbf{z}_{-i}, \alpha) &= \sum_{\mathbf{x}_j \in \{\mathbf{x}\}^k} p(c_i = j|\mathbf{c}_{-i}, \alpha) \\
&\propto \sum_{\mathbf{x}_j \in \{\mathbf{x}\}^k} f(d(i,j)) \\
&= g_{sum}(\mathbf{x}_i, \{\mathbf{x}\}^k),
\end{aligned} \tag{11}
$$

where the proportionality is due to Equation (3).

Proposition 2 shows that the abstract similarity g_{sum} exhibits the same clustering effect as the ddCRP prior: data points whose distances are small will have high prior probability of being in the same cluster.

3.3 Time Complexity Analysis

The time complexity of each iteration in the Gibbs sampling of the proposed model depends on the specific implementation of the abstract similarity g. For g_{max} and g_{sum}, the time complexity of each iteration is $O(N^2 + N \times D^2 \times K)$, larger than the $O(N \times D^2 \times K)$ complexity of MMDPM. In practice, the speed of convergence depends strongly on the speed of reduction in the number of clusters K, so an MMDPM model converging to a large number of clusters might still be slower than a STANDPM model converging to a small number of clusters. In addition, the additional $O(N^2)$ is still orders of magnitude lower than the additional $O(N^2 \times K \times D)$ that would have been required if merging clusters was used. In general, the significance of the abstract similarity is in its ability to model a larger range of clustering effects and to integrate domain-specific constraints naturally for cluster analysis, rather than to reduce time complexity of the inference of a DPM model.

4 Experiments

We study the effectiveness of the two abstract similarity functions $g_{\max}$ and g_{sum} by comparing the two models STANDPM-max and STANDPM-sum with a number of baseline methods on several low and high dimensional datasets.

Baselines: the STANDPM is compared directly to three baseline models: ddCRP, DPM and MMDPM. We also include a number of other state-of-the-art clustering methods: Bayesian Nonparametric Kmeans (BN-Kmeans) [Kulis and Jordan, 2012], Gaussian Mixture Model (GMM) [McLachlan et al., 2003], Spectral clustering, DP Variable Clustering (DPVC) Palla et al. [2012]. Because BN-Kmeans, Spectral and GMM require a pre-defined number of clusters, they **do not** directly solve the problem of interest which involves estimating the number of clusters. In the experiments, they are given the number of ground truth clusters and their performances serve only as upper bounds for their respective approaches. When possible, some results of the baseline methods are taken from [Chen et al., 2016].

Datasets: The models are evaluated on 10 datasets shown in Table 1, including three synthetic datasets Aggregation, Jain and Flame (available online[1]); four small datasets from the UCI Machine Learning Repository: Wdbc, Glass, Iris, Wine [2]; the MNIST dataset [3]; two large datasets Reuters21578 [Cai et al., 2005] and 20 Newsgroup[4]. All datasets are pre-processed and normalized similar to [Chen et al., 2016] for a fair comparison. The Aggregation dataset is generated from 7 Gaussian mixtures and thus the performance of the DPM model is the upper bound for all other methods [Chen et al., 2016]. The ddCRP model is not tested on Reuters21578 and 20 Newsgroup since it does not scale to these large datasets, taking more than 12 hours to complete a Gibbs sampling iteration.

Evaluation measure: The clustering performance is measured in F-score [Rijsbergen, 1979], V-score [Rosenberg and Hirschberg, 2007] (equivalent to the Normalized Mutual Information metric Becker [2011]) and Adjusted Rand Index (ARI). F-score belongs to the class of pair-matching metrics which favors a high number of pairs of data points that are in the same clusters in two clustering results [Amigó et al., 2009]. On the other hand, V-score belongs to the class of entropy-based metrics which measure homogeneity and completeness of clusters.

Table 1. Number of data points N, dimensionality D and number of classes of all datasets.

Dataset	N	D	#Classes
Jain	373	2	2
Aggregation	788	2	7
Flame	240	2	2
Wdbc	569	30	2
Glass	214	9	6
Iris	150	4	3
Wine	178	13	3
MNIST	2000	784	10
Reuters21578	8293	18933	65
20 Newsgroup	10000	61188	20

Hyper-parameters setting: The Euclidean distance metric is used for the computation of the pairwise distances in the prior function. The decay function has the form $f(d) = \exp(\frac{-d}{\gamma})$ where γ is a decay factor, $\min_{i,j} d(i,j) \leq \gamma \leq \max_{i,j} d(i,j)$. The generative model in ddCRP is specified by the Normal-Inverse-Wishart distribution. Similar to [Chen et al., 2016], we find that the hyper-parameter λ has the highest impact on the number of clusters. A random search on the range from 0.001 to 5 is used to tune λ. The number of burn-in iterations in all experiments is $T = 100$. After the burn-in phase, 5 clusterings are sampled and their average score is reported. All experiments are run on a computer with Intel(R) Core(TM) i7-5930K CPU 3.50GHz and 32 GB of RAM.

Results and Analysis

Small low and high dimensional datasets: The clustering performance on the first 8 small datasets are shown in Tables 2 and 3. The STANDPM models

[1] http://cs.joensuu.fi/sipu/datasets/
[2] http://archive.ics.uci.edu/ml
[3] http://yann.lecun.com/exdb/mnist/
[4] http://qwone.com/ jason/20Newsgroups/

Table 2. Results by **F-score on the first 8 small datasets** of STANDPM and baseline methods. Bold numbers indicate the best scores for each data set. Underlined numbers indicate that the scores of STANDPM are better than or equal to MMDPM.

	ddCRP	BN-Kmeans	DPM	MMDPM	STANDPM-max(ours)	STANDPM-sum(ours)
Jain	0.37	0.57	0.59	0.73	0.76	**0.87**
Aggregation	0.59	0.77	**0.91**	0.79	0.84	0.83
Flame	0.63	0.627	0.43	0.6	0.73	**0.74**
Wdbc	0.61	0.80	0.43	0.85	0.85	**0.87**
Glass	0.39	0.49	0.49	0.51	**0.54**	0.52
Iris	0.75	0.74	0.71	0.75	0.75	**0.76**
Wine	0.33	0.87	0.36	0.68	0.88	**0.90**
MNIST	0.21	0.239	0.18	**0.368**	0.320	0.275

Table 3. Results by **V-score on the first 8 small datasets** of STANDPM and baseline methods. Bold numbers indicate the best scores for each data set. Underlined numbers indicate that the scores of STANDPM are better than or equal to MMDPM.

	ddCRP	BN-Kmeans	DPM	MMDPM	STANDPM-max(ours)	STANDPM-sum(ours)
Jain	0.43	0.46	0.46	0.41	0.47	**0.56**
Aggregation	0.81	0.76	**0.90**	0.75	0.86	0.85
Flame	0.52	0.53	0.41	0.29	**0.63**	0.61
Wdbc	0.003	0.51	0.26	0.57	0.59	**0.62**
Glass	0.03	0.30	0.43	0.38	**0.46**	0.42
Iris	**0.73**	0.71	0.66	0.68	**0.73**	0.67
Wine	0.03	0.78	0.42	0.48	0.81	**0.85**
MNIST	0.12	0.262	0.06	0.389	**0.505**	0.45

consistently outperform ddCRP and MMDPM in both F-score and V-score on all datasets. The clustering results on the synthetic Aggregation dataset are shown in Figure 2. It can be observed that the separation between two rightmost clusters are captured almost perfectly by STANDPM, while MMDPM fails to do so, suggesting that the abstract similarity helps improving the max-margin likelihood learning process. In the task of character clustering on MNIST dataset, while slightly staying behind in F-score, STANDPM models outperform MMDPM by large margins in V-score. Without access to the true number of clusters, they also outperforms BN-Kmeans on all datasets.

Document clustering: The Reuters21578 dataset consists of 8293 documents of 65 categories. This dataset is challenging because the distribution of data points into classes is highly unbalanced, with 44.77% of the data in one single class and nearly 80% of the data in the five largest classes. As a result, methods assuming a CRP prior with "richer get richer" effects such as DPM and MMDPM are advantageous over other methods. This advantage is reflected in Table 4 where DPM and MMDPM outperform other methods in F-score and ARI. However, both STANDPM models have higher V-score than MMDPM, indicating that clusters generated by STANDPM have higher degree of homogeneity and completeness simultaneously. We conjecture that the abstract simi-

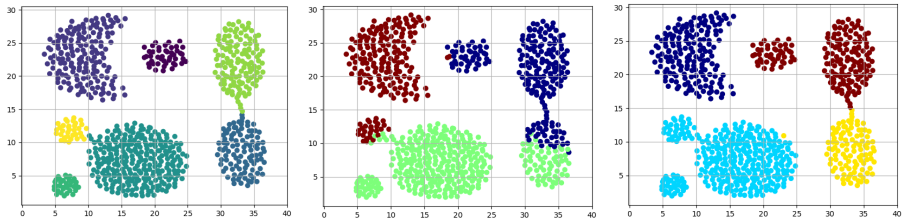

Fig. 2. Visualization of the ground truth (left), clustering output by MMDPM (middle, F-score=0.79, V-score=0.75) and clustering output by STANDPM-max (right, F-score=0.84, V-score=0.86) for the Aggregation dataset. Notice the data in the two rightmost classes are directly connected and yet the STANDPM is able to find an almost perfect separation between the two classes.

Table 4. Experimental results on **the Reuters dataset** of STANDPM in comparison with baseline methods. GMM and spectral clustering are provided with the ground truth number of clusters.

	GMM	Spectral	DPM	DPVC	MMDPM	STANDPM-max(ours)	STANDPM-sum(ours)
F-score	0.173	0.09	0.484	0.32	**0.507**	0.335	0.379
V-score	0.464	0.432	**0.472**	0.395	0.335	<u>0.372</u>	<u>0.434</u>
ARI	0.123	0.062	0.383	0.211	**0.416**	0.265	0.273

larity could not capture the imbalance in this dataset as effectively as the CRP prior because the Euclidean distances become less effective for high-dimensional and sparse features [Aggarwal et al., 2001]. The same observation can be seen in [Chen et al., 2016] where the deterministic K-means with the true number of clusters failed to compete with MMDPM in F-score and ARI.

News article clustering: To further demonstrate scalability, we perform an additional experiment on the 20 Newsgroup dataset with 18846 documents of 61188 vocabularies. We extract 10000 documents and project them onto 250 most frequent words similar to [Chen et al., 2016]. On this large dataset, the STANDPM-max model outperforms MMDPM by a large margin and achieves the best result in V-score; the STANDPM-sum model achieves the best result in F-score, tie with MMDPM. This shows that STANDPM scales up to large datasets. The results are given in Table 5.

Number of generated clusters: Figure 3 shows the number of clusters generated by STANDPM and MMDPM in comparison to the ground truth. Since the abstract similarity is less dependent on the size of the clusters, STANDPM tends to generate more clusters than MMDPM. In some cases, this leads to scattered clusters of small sizes. Since the summation in g_{sum} implies that large clusters tend to have higher prior probability, it often generates more clusters g_{max}.

Table 5. Experimental results on **the 20 Newsgroup dataset** of STANDPM in comparison with baseline methods. GMM and spectral clustering are provided with the ground truth number of clusters.

	GMM	Spectral	DPM	DPVC	MMDPM	STANDPM-max(ours)	STANDPM-sum(ours)
F-score	0.088	0.095	0.094	0.09	**0.10**	0.077	**0.10**
V-score	0.104	0.061	0.049	0.02	0.066	**0.18**	0.061

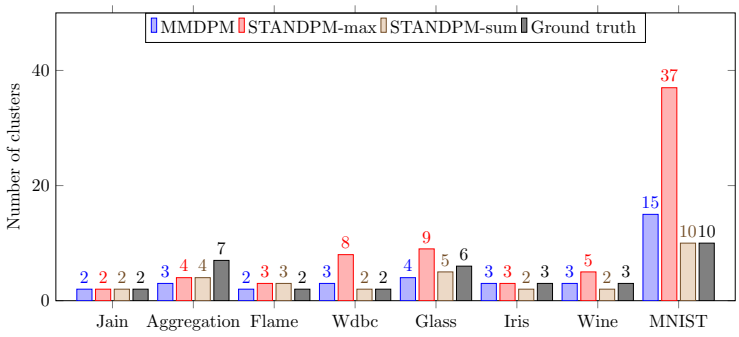

Fig. 3. Number of clusters generated by MMDPM and STANDPM for different datasets.

5 Conclusion

This paper introduces STANDPM model, an efficient and scalable solution to the challenge of integrating of the distance-based ddCRP prior into the max-margin discriminatively learned DPM model. The distances are integrated via the abstract similarity measurement between a data point and a cluster. The abstract similarity can be flexibly chosen so that either only a subset of nearest neighbors of a data point (g_{max}) or the whole set of data points (g_{sum}) contributes to the computation of the prior. We formally show that the abstract similarity can generalize the clustering effects of ddCRP prior and there exists efficient Gibbs sampling inference for its DPM models. Experimental results show that STANDPM models achieve state-of-the-art clustering performance on challenging real datasets without access to the true number of clusters.

References

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, November 2012. ISSN 0162-8828.

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory — ICDT 2001*, pages 420–434. Springer Berlin Heidelberg, 2001.

D. J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII*, pages 1–198. Springer Berlin Heidelberg, 1985.

E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, Aug 2009. ISSN 1573-7659.

Hila Becker. *Identification and Characterization of Events inSocial Media*. PhD thesis, Columbia University, 2011.

D. M. Blei and P. I. Frazier. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461–2488, November 2011.

D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 2005.

D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12):1624–1637, December 2005.

C. Chen, J. Zhu, and X. Zhang. Robust bayesian max-margin clustering. In *Advances in Neural Information Processing Systems 27*, pages 532–540. 2014.

G. Chen, H. Zhang, and C. Xiong. Maximum margin dirichlet process mixtures for clustering. In *AAAI Conference on Artificial Intelligence*, 2016.

D. Dahl. Distance-Based Probability Distribution for Set Partitions with Applications to Bayesian Nonparametrics. In *JSM*, 2008.

K. A. Heller and Z. Ghahramani. Bayesian hierarchical clustering. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 297–304, New York, NY, USA, 2005. ACM.

R. S. King. *Cluster Analysis and Data Mining: An Introduction*. Mercury Learning and Information, USA, 2014.

B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, pages 1131–1138, USA, 2012. Omnipress.

G. J. McLachlan, S. K. Ng, and D. Peel. On clustering by mixture models. In *Exploratory Data Analysis in Empirical Research*, pages 141–148, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

K. Palla, D. A. Knowles, and Z. Ghahramani. A nonparametric variable clustering model. In *Advances in Neural Information Processing Systems 25*, pages 2987–2995. 2012.

C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979. ISBN 0408709294.

A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, 2007.

J. Zhu, N. Chen, and E. Xing. Infinite svm: a dirichlet process mixture of large-margin kernel machines. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 617–624, New York, NY, USA, June 2011. ACM.