Quest Journals
Journal of Software Engineering and Simulation
Volume 6 ~ Issue 5 (2020) pp: 09-16
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

A Dirichlet Process Mixture Model for Multi-view Clustering

Hongtao Wang and Hongmei Wang
[1](North China Electric Power University)
Corresponding Author: Hongmei Wang

ABSTRACT: *Multi-view clustering is an important task in machine learning domain. It aims to partition data into clusters across multiple views of representations. Traditional approaches always concatenate features together to perform single-view clustering algorithms, which may raise a fundamental problem: the relationships of different view representations are overlooked. In recent years, many alternative approaches have been proposed to make use of these heterogeneous multi-view information. However, they often suffer from the assumption that there are the same feature generation distributions for all views, which may be not true in practice. In this paper, we propose a novel non-parametric Bayesian generative model, Multi-View data via a Dirichlet Process Mixture Model, namely MVDPMM, to address this issue. By incorporating dirichlet process, the proposed model can automatically form a cluster structure across all views, without selecting the number of clusters. Furthermore, we assume there is a latent parameter space shared by all views, and data in each view are derived through a projection matrix. We devise an effective inference algorithm using the collapsed Gibbs sampling to solve the proposed model. The model is evaluated on both synthetic and real-world datasets with different settings, and the experimental results demonstrate the effectiveness of our model in contrast with some competitive baselines.*
KEYWORDS: *Multi-view; Clustering; Dirichlet Process*

R*eceived 14 December,, 2020;* ***A****ccepted 28 December, 2020 © The author(s) 2020.*
P*ublished with open access at* **www.questjournals.org**

I. INTRODUCTION

The profusion of big data brings not only more volumes, but also more attributes and aspects of data. Since data are increasingly collected from diverse sources, many datasets are formed by heterogeneous representations for same instances. This kind of data is usually called multi-view data, where multi-view means a number of feature sets can be organized to form multiple views or representations from data. For example, the news can be displayed by different sources with different languages; the web pages always contain different kinds of con- tents such as texts, images and links. These multi-view data not only provides useful knowledge for applications but sheds light on understanding the intrinsic structure of data generation.

Performing clustering on multi-view data is an extraordinarily important task in multi-view machine learning. Multi-view clustering aims to partition data into clusters across multiple view representations. Although features can be fused together to perform single-view clustering algorithms, this roadmap raises a fundamental problem: the relationships of different view representations are overlooked. A number of alternative approaches have been proposed to make use of these heterogeneous information, which can be classified into three categories: subspace methods to obtain a feature subspace representation and perform clustering on them; non-negative matrix factorization (NMF) based methods which seek common latent factors among views; graph based methods which generate a fusion graph to perform spectral clustering algorithms.

These prominent solutions demonstrate the promise of heterogeneous information, but suffer from several limitations: (1) the difficulty of model selection is often bypassed when the number of clusters is fixed; (2) they are sensitive to outliers which could reduce the performance when deeming outliers as normal instances; and (3) they often assume the same feature generation distributions for all views, which may be not true in practice.

To address the above questions, in this paper, we propose a novel non- parametric bayesian approach for multi-view clustering. Specifically, we model Multi-View data via a Dirichlet Process Mixture Model, namely MVDPMM. By incorporating dirichlet process, the proposed model can automatically form a cluster structure across all views, without selecting the number of clusters. The noise and outliers generated by different distributions can effectively lower the impact of normal instances. Furthermore, we assume there is a latent parameter space shared by all views, and subspace in each view are derived through a projection matrix. We devise an effective inference algorithm to solve the proposed model. In contrast with some competitive baselines, our approach can achieve a better clustering performance on several multi-view benchmarks. We summarize the major contributions as follows:

- We formalize the multi-view clustering problem in a non-parametric way, and propose a generative approach based on dirichlet process mixture model. Our approach is more interpretable than others since it models data in a probabilistic generative perspective.
- With our model, features in each view are generated by diverse distributions from a parameter subspace which is projected from a common parameter space. The advantage is that complementary information across views con- tributes to a consistent clustering result.
- We employ a Gibbs Sampling method to effectively solve the model inference. The performance of the proposed approach is then evaluated through extensive experiments on several real-world benchmark datasets.

The rest of the paper is organized as follows. In section II we review related work briefly on multi-view clustering. We then formally present the model specification and propose the inference algorithm to discover clusters in Section III. In Section IV, we evaluate the proposed model using several real-world datasets, and report the performance results with discussions. We finally conclude the paper in Section V.

II. RELATED WORKS

Multi-view data clustering has been ubiquitous in recent years since multiple view often provides coherent or complimentary information. Studying multi-view Data is with great challenges compared to single-view data due to different generation mechanisms under different views. In this section, we review the related models on multi-view clustering methods.

The most popular research on multi-view data is multi-view clustering or co-clustering. Multi-view clustering aims to divide objects into clusters based on multiple views of entities. One class of methods is to utilize traditional single- view clustering algorithms. For example, the straightforward methods incorporate attributes of multiple views into the classical clustering process directly [1]. In contrast, late integration method [4] derives results from each individual view and then fuses them base on consensus. [2] projects multi-view data into a common lower dimensional subspace and then applies k-means to get the partitions. Another cate- gory of multi-view clustering is relying on matrix factorization technique (MF). [3] proposed a non-negative matrix factorization (NMF) based approach that gives compatible clustering solutions across multiple views. To find out the key subset of features in each view that are associated with the clusters, [5] simultaneously decomposes multiple data matrices into sparse row and columns vectors.

Spectral clustering [6], [7] is a popular clustering technique that has shown good performance on arbitrary shaped data. It is extensively studied as a significant framework by many literatures on multi-view datasets. [8] proposed a random walk based solution to derive a global graph cut over the single graphs in other views. In [9], the authors applied the idea of co-training and proposed a Co-Training multi-view Spectral Clustering (CTMSC), where the new graph similarity of one view is limited by solving the eigenvectors of the Laplacian of other views. [10] added a co-regularization term to enforce the eigenvectors of diverse views having high pairwise similarities, and proposed a method called Co-Regularized multi-view Spectral Clustering (CRMSC). [11] presented a novel method named by MVSC-CEV, to computing the common eigenvectors of the Laplacian matrices derived from the similarity matrices of the input views. To address the possible dependencies among views, [12] adopts the brainstorming process to compensate the biases caused by information sharing between multiple views with dependent opinions, and finally a compromise opinion is merged.

III. THE PROPOSED MODEL

3.1 A Brief Review of DPMM

Dirichlet Process Mixture Model (DPMM) is the most successful application to the Dirichlet Process. It can be considered as an infinite extension of Finite Mixture Model (FMM) which assumes data is generated from a mixture of components or distributions. Figure 1 shows the graphical description of DPMM in stick-breaking representations.

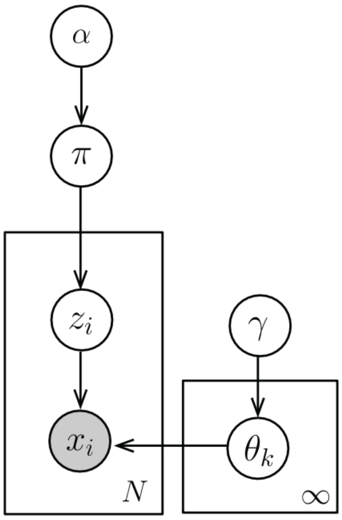

Figure 1. Graphical Model Representation of DPMM.

Suppose we have N observations $x_1, x_2, \cdots, x_N$. The generative process which is equivalent to the graph is as follows:

$$\begin{aligned} &\boldsymbol{\pi}|\alpha \sim GEM(1,\alpha) \\ &z_i|\boldsymbol{\pi} \sim Discreate(\boldsymbol{\pi}) \\ &\theta_k|\gamma \sim G_0(\gamma) \\ &x_i|z_i, \{\theta_k\}_{k=1}^{\infty} \sim F(\theta_{z_i}). \end{aligned} \tag{1}$$

In this model, $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ is the mixture weight of components $\sum_{k=1}^{\infty} = 1$. It can be constructed from a stick-breaking process, named after Griffith, Engen, and McCloskey (GEM). z_i denotes the cluster indicator for instance x_i and can be sampled by a discrete distribution given π. The cluster parameter θ_k are sampled from a shared prior distribution $G_0(\gamma)$ parameterized by γ. Finally, each instance x_i is generated from a distribution $F(\cdot)$ given the parameter θ_{zi} .

If we integrating the mixture weight π out then the DPMM is transformed to the Chinese Restaurant Process (CRP), which explicitly illustrates the conditional distribution of cluster indicators as follows:

$$p(z_i = k|\boldsymbol{z}_{-i}, \alpha) = \begin{cases} \frac{N_{k,-i}}{N+\alpha-1} & \text{for cluster } k \\ \frac{\alpha}{N+\alpha-1} & \text{for new cluster.} \end{cases} \tag{2}$$

where z_{-i} denotes the cluster indicators except z_i, and $N_{k,-i}$ denotes the number of instances in cluster k excluding the i-th instance.

The above model can be Inferred by either Gibbs Sampling or Variational Inference methods. We can clearly see that DPMM can effectively cluster data without requiring us to specify the number of clusters.

3.2 Model Specification

We now start to study the problem multi-view clustering. Given a multi-view dataset with N instances, suppose each instance has D views denoted by $x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(D)})$. Multi-view clustering problem is to partition these instances into several clusters. We admit the general assumption as previous researches that instances have consistent cluster structure with all views. The difficulty in the probabilistic generative perspective lies in that data in different views could be generated by diverse distributions with totally different parameter space.

To address this challenge, we assume that there is a common parameter space shared by all views, and each datum is implicitly sampled from a latent measure parameterized by θ_k in this shared space. In order to generate $x_i^{(j)}$ in view j, there exists a projection matrix $W^{(j)}$ that transform θ_k to parameter $\theta_k^{(j)}$. Note that $\theta_k^{(j)}$ may be in a subspace of the original shared space, or be in a totally different parameter space. Then instances in view j are generated by view-specific distribution with projected parameters. The graphical representation of our MVDPMM is shown in Figure 2.

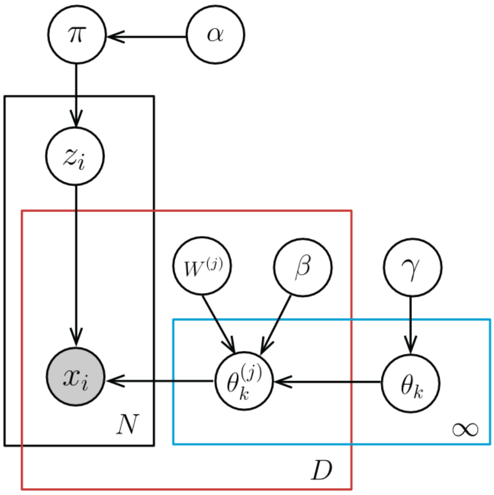

Figure 2. Graphical Model Representation of MVDPMM.

The generative process of MVDPMM is as follows:

$$
\begin{aligned}
&\boldsymbol{\pi}|\alpha \sim GEM(1,\alpha) \\
&z_i|\boldsymbol{\pi} \sim Discreate(\boldsymbol{\pi}) \\
&\theta_k|\gamma \sim G_0(\gamma) \\
&\theta_k^{(j)}|W^{(j)},\theta_k,\beta \sim \mathcal{N}\left(W^{(j)}\theta_k,\beta^{-1}\boldsymbol{I}\right) \\
&x_i^{(j)}|z_i,\{\theta_k^{(j)}\}_{k=1}^{\infty} \sim F_j(\theta_{z_i}^{(j)}).
\end{aligned}
\tag{3}
$$

where $N(\cdot)$ denotes the Gaussian distribution, $\boldsymbol{I}$ is the identity matrix, and $F_j(\cdot)$ denotes the specific distributions in view *j*.

There are three advantages of our model. Firstly, we can clearly see that the generation of π, z_i and θ_k in MVDPMM are consistent with DPMM. This ensures the infinite mixture nature of our model to be a successful generalization of Dirichlet Process, and endow MVDPMM the ability of automatically select the number of clusters. Another novelty of MVDPMM is that it can utilize the heterogeneous information in multiple views in a robust matter. To achieve this, the view-specific parameter $\theta_k^{(j)}$ is generated by a Gaussian distribution conditioned on $W^{(j)}$, θ_k, and β. The mean $W^{(j)}\theta_k$ converts cluster parameters in common space to the parameter in view *j*. It is more practical to allow noise and outliers added in the process of projection. Thus we adopt β as precision parameter to control the robustness of our model. The third virtue is that each instance xi can be generated from diverse distributions $F_j(\cdot)$ with different projected parameters $\theta_{z_i}^{(j)}$ for the same cluster z_i.

Discussions on $W^{(j)}$, The projection matrix $W^{(j)}$ in our model can either be a static parameter fixed beforehand, or a latent variable. For the simplest example, let the common space C be expanded by the concatenation of parameters in all views. Thus C is the direct sum of parameter subspaces satisfying $dim(\mathcal{C}) = \sum_{j=1}^{D} dim(V^{(j)})$, where *dim*() demotes the dimension of space and $V^{(j)}$ denotes parameter subspace. In this case, $W^{(j)}$can be defined as $(\boldsymbol{O}, \cdots, \boldsymbol{O}, \boldsymbol{I}, \boldsymbol{O}, \cdots, \boldsymbol{O})$ with dimension $dim(V^{(j)})\times dim(C)$, where $\boldsymbol{I}$ is identity matrix corresponding to the *j*-th component. Then θ(j) can be directly extracted *k* from θ_k by $W^{(j)}\theta_k$ without considering noise. If we treat $W^{(j)}$ as a latent variable, it needs to be estimated from data by maximum likelihood method or bayesian treatment. The details of estimating $W^{(j)}$ are in the next subsection.

3.2 Model Inference

The goal of model inference for MVDPMM is to learn the posterior distribution of latent variables, especially z_i. The complete joint likelihood of data and latent variables can be written as:

$$
\begin{aligned}
&p(\mathbf{z},\boldsymbol{\pi},\mathbf{x},\boldsymbol{W},\{\theta_k^{(j)}\}_{k=1\cdots\infty}^{j=1\cdots D},\{\theta_k\}_{k=1\cdots\infty} \mid \alpha,\beta,\gamma) \\
&= \prod_i^N \prod_j^D F_j(x_i^{(j)}|\theta_{z_i}^{(j)}) Discrete(z_i|\boldsymbol{\pi}) GEM(\boldsymbol{\pi}|1,\alpha) \\
&\quad \times \mathcal{N}(\theta_{z_i}^{(j)}|W^{(j)}\theta_k,\beta^{-1}\boldsymbol{I}) P(W^{(j)}) G_0(\theta_k|\gamma).
\end{aligned}
\tag{4}
$$

Since directly solving the posterior distribution of latent variables from Eq.(4) is difficult, we adopt the Gibbs Sampling [15] method. Gibbs Sampling is one of Markov chain Monte Carlo (MCMC) sampling algorithms for statistical inference. We breakdown the inference procedure into four steps in each iteration as follows.

Step 1: Sampling cluster indicator variables z.

According to the conditional independence principle, the posterior of cluster indicator variable z_i can be given by

$$\begin{aligned} p(z_i = k|\sim) &= p(z_i = k \,|\mathbf{z}_{-i}, \boldsymbol{\pi}, \mathbf{x}, \boldsymbol{W}, \{\theta_k^{(j)}\}_{k=1\cdots\infty}^{j=1\cdots D}, \{\theta_k\}_{k=1\cdots\infty}, \alpha, \beta, \gamma) \\ &\propto p(z_i = k \,|\boldsymbol{\pi}) \prod_{j=1}^{D} p(x_i^{(j)}|\theta_k^{(j)}) = \pi_k \prod_{j=1}^{D} F_j(x_i^{(j)}|\theta_k^{(j)}). \end{aligned} \tag{5}$$

where ~ denotes all variables except z_i, and z_{-i} denotes the set of cluster indicator variables excluding z_i.

Next a straightforward method is to sample mixture weight π from GEM distribution. To induce a more efficient inference procedure, an alternative way is integrating π out. By incorporating both (2) and (5), we can derive the following posterior of z_i using Bayesian rule and Markov property:

$$\begin{aligned} p(z_i = k|\sim) &= p(z_i = k \,|\mathbf{z}_{-i}, \mathbf{x}, \{\theta_k^{(j)}\}_{k=1\cdots\infty}^{j=1\cdots D}, \alpha) \\ &\propto \begin{cases} \frac{N_{k,-i}}{N+\alpha-1} \prod_{j=1}^{D} F_j(x_i^{(j)}|\theta_k^{(j)}) & \text{for cluster } k \\ \frac{\alpha}{N+\alpha-1} \prod_{j=1}^{D} F_j(x_i^{(j)}|\theta_k^{(j)}) & \text{for new cluster.} \end{cases} \end{aligned} \tag{6}$$

According to Eq.(6), we could sample the cluster indicator z_i for x_i through fixing other cluster indicators except z_i. Then in this step, we could alternatively sample the cluster indicators for all instances.

Step 2: Sampling cluster parameter $\theta_k^{(j)}$ for each view.

$$\begin{aligned} p(\theta_k^{(j)}|\sim) &\propto p(\{x_i^{(j)}\}_{i=1\cdots N} \mid z_i, \theta_k^{(j)}) \cdot p(\theta_k^{(j)} \mid W^{(j)}, \theta_k, \beta) \\ &\propto \prod_{i=1}^{N} p(x_i^{(j)}|\theta_{z_i=k}^{(j)}) \cdot \mathcal{N}(\theta_k^{(j)} \mid W^{(j)}\theta_k, \beta^{-1}\boldsymbol{I}). \end{aligned} \tag{7}$$

Step 3: Sampling shared cluster parameter θ_k.

$$\begin{aligned} p(\theta_k|\sim) &= p(\theta_k \mid \theta_k^{(j)}, W^{(j)}, \beta, \gamma) \\ &\propto \prod_{j=1}^{D} p(\theta_k^{(j)} \mid \theta_k, W^{(j)}, \beta) \cdot p(\theta_k|\gamma) \\ &\propto \prod_{j=1}^{D} \mathcal{N}(\theta_k^{(j)} \mid W^{(j)}\theta_k, \beta^{-1}\boldsymbol{I}) \cdot G_0(\gamma). \end{aligned} \tag{8}$$

Step 4: Estimating and Updating projection matrices W.

Given $\theta_k^{(j)}$ and θ_k, we can update projection matrix $W^{(j)}$ via a bayesian view:

$$p(W^{(j)}|\theta_k^{(j)}, \theta_k, \beta) = p(\theta_k^{(j)} \mid W^{(j)}, \theta_k, \beta) \cdot p(W^{(j)}). \tag{9}$$

Note that $p(\theta_k^{(j)} \mid W^{(j)}, \theta_k, \beta) = \mathcal{N}(\theta_k^{(j)} \mid W^{(j)}\theta_k, \beta^{-1}\boldsymbol{I})$ and let the prior distribution $p(W^{(j)}) = 1$ for simplicity. Then $W^{(j)}$ can be estimated by maximizing the posterior distribution and we can the following results:

$$W^{(j)} = (\theta_k\theta_k^T)^{-1}\theta_k^{(j)}\theta_k^T. \tag{10}$$

IV. PERFORMANCE EVALUATION

In this section, we conduct experiments on both synthetic and real-world multi-view datasets. First, we introduce the experimental settings as well as baseline methods. Then we evaluate the effectiveness and performance of the proposed FMSC approach against the baseline methods.

4.1. Experiment Settings.

To evaluate the performance of the proposed approach, we adopt five multi-view datasets including both synthetic and real-world datasets.

Synthetic data is generated from two views to four views with two clusters. Similar to the settings in [9] and [10], we randomly sample 1000 instances (n = 1000) for each view via 2-dimentional Gaussian mixture model. The means and covariance of all clusters in different views are defined in Table 1.

We also adopt five popular real-world datasets: UCI digits, Reuters, BBC, BBCSport and Yale datasets. The datasets are introduced as follows:

- UCI digits. This dataset is created from handwritten numerals from 0 to 9, and each instance forms an image with 15 × 16 pixels. It is available at the UCI repository.
- Reuters. Reuters multilingual corpus is a set of news articles written in five languages: English, French, German, Italian and Spanish. We sample 10, 000 instances into our experiments from the original corpus that contains 18,758 articles. For each view, we use a word dictionary as features and compute the term frequencies as feature values.
- BBC and BBCSport datasets. Both of the two datasets are multi-view news articles from the BBC. BBC contains five possible topics (business, entertainment, politics, sport and technology), and BBCSport focuses on five sports topics (athletics, cricket, football, rugby and tennis). We also adopt term frequencies as feature values.
- Yale. Yale is an image dataset that consists of 165 face images of 15 classes in pixel. Each class has 11 images.

We summarize some characters of the above datasets in Table1. Note that the numbers in column '#Feature' are the sum of features in all views.

Table 1. Description of real-world datasets.

Dataset	#Sample	#Feature	#View	#Class
UCI digits	2000	649	6	10
Reuters	10000	2500	5	6
BBCSport	544	6386	2	5
BBC	685	18491	4	5
Yale	165	14150	3	15

To validate the effectiveness of our MVDPMM approach, we run our approach on the above-mentioned datasets and compare the results against three baseline clustering methods: Concatenation, Co-training, and Co-reg. Concatenation is a straightforward approach that features of each view are concatenated to perform a standard spectral clustering. Co-training [9] alternatively performs spectral clustering on each view by incorporating other view's eigenvector matrices as constraints. Co-reg refers to Co-Regularized multi-view spectral clustering approach proposed by [10]. Our work is based on Co-reg and extends it to the federated scenario.

We measure the quality of clustering results on several metrics: Normalized Mutual Information (NMI), Adjusted mutual information (AMI), Adjusted Rand index (ARI), Completeness (COM) and Homogeneity (HOM). NMI and AMI are two different normalized versions of Mutual Information (MI), which quantifies how much the estimated clustering is informative about the true clustering [13]. The ARI metric is similar to the clustering accuracy and measures the degree of agreement between the estimated clustering and the true clustering. COM and HOM are two important conditional entropy based metrics [14]. COM measures the ratio of the member of a given class that is assigned to the same cluster, while HOM measures the ratio of instances of a single class pertaining to a single cluster.

In order to evaluate the execution time of the proposed approach, we also introduce Running Time in seconds as a metric. For all the above metrics, a higher value means the corresponding method has a better performance. We repeat 10 times and report the mean values for each experiment. All approaches are run on an Intel i9-9900K 3.6 GHz with 64GB of RAM, using single-threaded processes.

4.2. Results.

In this subsection, we report the experiment results of our MVDPMM approach on datasets compared with baselines. The aim is to verify whether it can receive as considerable clustering performance as baseline approaches. We report the results as follows.

Table 2 demonstrates the clustering results on synthetic datasets. We can easily see that the overall performance of four-views dataset is better than that of three-views, and the overall performance of three-views dataset is better than that of two-views. The results suggest that more views can bring improvements on the clustering performance. For different approaches, the performance of MVDPMM and Co-reg is superior than Co-training and Concatenation. Our MVDPMM approach has almost as the same performance as Co-reg.

Table 2. Clustering results on synthetic datasets.

Dataset	Method	NMI	AMI	ARI	COM	HOM
Two views	Concatenation	0.719	0.715	0.772	0.723	0.715
	Co-training	0.571	0.569	0.661	0.574	0.569
	Co-reg	0.744	0.742	0.81	0.746	0.742
	MVDPMM	0.748	0.743	0.820	0.749	0.745
Three views	Concatenation	0.863	0.862	0.913	0.863	0.862
	Co-training	0.802	0.802	0.876	0.803	0.802
	Co-reg	0.872	0.872	0.924	0.873	0.872
	MVDPMM	0.876	0.876	0.929	0.877	0.876
Four views	Concatenation	0.872	0.872	0.923	0.873	0.872
	Co-training	0.824	0.823	0.894	0.824	0.824
	Co-reg	0.887	0.886	0.937	0.887	0.886
	MVDPMM	0.893	0.892	0.945	0.893	0.893

Table 3 shows the multi-view clustering results on five real-world datasets. For UCI digits and Reuters, our MVDPMM approach has the best performance than other baselines. For on BBCSports, BBC, and Yale dataset, the performance of our approach is as almost the same as Co-training and Co-reg methods on all metrics. So, we can safely say that our MVDPMM is a good clustering approach.

Table 3. Clustering results on real-world datasets.

Dataset	Method	NMI	AMI	ARI	COM	HOM
UCI digits	Concatenation	0.724	0.706	0.598	0.74	0.709
	Co-training	0.841	0.839	0.826	0.841	0.84
	Co-reg	0.85	0.848	0.832	0.851	0.849
	MVDPMM	0.856	0.854	0.846	0.855	0.852
Reuters	Concatenation	0.397	0.309	0.177	0.504	0.313
	Co-training	0.397	0.375	0.282	0.416	0.378
	Co-reg	0.396	0.369	0.273	0.42	0.373
	MVDPMM	0.403	0.382	0.287	0.426	0.387
BBCSports	Concatenation	0.61	0.598	0.584	0.619	0.602
	Co-training	0.708	0.693	0.684	0.696	0.721
	Co-reg	0.618	0.611	0.589	0.615	0.621
	MVDPMM	0.712	0.703	0.686	0.670	0.722
BBC	Concatenation	0.677	0.665	0.703	0.687	0.668
	Co-training	0.738	0.728	0.751	0.73	0.746
	Co-reg	0.738	0.728	0.768	0.731	0.745
	MVDPMM	0.742	0.732	0.766	0.731	0.745
Yale	Concatenation	0.618	0.483	0.38	0.623	0.613
	Co-training	0.726	0.628	0.544	0.731	0.722
	Co-reg	0.726	0.629	0.54	0.73	0.723
	MVDPMM	0.735	0.632	0.548	0.731	0.725

V. CONCLUSION

In this paper, we propose a non-parametric bayesian model, namely, MVDPMM, for multi-view clustering. By incorporating dirichlet process, the proposed model can automatically form a cluster structure across all views, without selecting the number of clusters. The noise and outliers generated by different distributions can effectively lower the impact of normal instances. Furthermore, we assume there is a latent parameter space shared by all views, and subspace in each view are derived through a projection matrix. We devise an effective inference algorithm to solve the proposed model. Experimental results show that our approach can achieve better clustering performance on several multi-view benchmarks.

REFERENCES

[1]. Steffen Bickel and Tobias Scheffer. Multi-view clustering. In Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), pages 19–26, 2004.

[2]. Kamalika Chaudhuri, Sham M. Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, pages 129–136, 2009.

[3]. Jing Gao, Jiawei Han, Jialu Liu, and Chi Wang. Multi-view clustering via joint nonnegative matrix factorization. In Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA., pages 252–260, 2013.

[4]. Eric Bruno and St´ephane Marchand-Maillet. Multiview clustering: a late fusion approach us- ing latent models. In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 736–737, 2009.

[5]. Jiangwen Sun, Jin Lu, Tingyang Xu, and Jinbo Bi. Multi-view sparse co-clustering via proxi- mal alternating linearized minimization. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, pages 757–766, 2015.

[6]. A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, Vancouver, British Columbia, Canada, December 3- 8, 2001*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, pp. 849–856.

[7]. U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[8]. D. Zhou and C. J. C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proceedings of the Twenty-Fourth International Conference on Machine Learning, Corvallis, Oregon, USA, June 20-24, 2007*, Z. Ghahramani, Ed., pp. 1159–1166.

[9]. A. Kumar and H. D. III, "A co-training approach for multi-view spectral clustering," in *Proceedings of the 28th International Conference on Ma- chine Learning, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 393–400.

[10]. A. Kumar, P. Rai, and H. D. III, "Co-regularized multi-view spectral clustering," in *Proceedings of the 25th Annual Conference on Neural Information Processing Systems, Granada, Spain, December 12-14 ,2011*, pp. 1413–1421.

[11]. S. Kanaan-Izquierdo, A. Ziyatdinov, and A. Perera-Lluna, "Multiview and multifeature spectral clustering using common eigenvectors," *Pattern Recognit. Lett.*, vol. 102, pp. 30–36, 2018.

[12]. J. W. Son, J. Jeon, A. Lee, and S. Kim, "Spectral clustering with brain- storming process for multi-view data," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, February 4-9, 2017*, pp. 2548–2554.

[13]. R. Boutalbi, L. Labiod, and M. Nadif, "Co-clustering from tensor data," in *Proceedings of the 23rd Pacific-Asia Conference, Macau, China, April 14-17, 2019*, pp. 370–383.

[14]. 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, Prague, Czech Republic, June 28-30, 2007*, pp. 410–420.

[15]. ndrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. Bayesian data analysis, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.