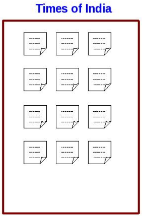
Learning Dirichlet Processes from Partially Observed Groups

Avinava Dubey[†], Indrajit Bhattacharya[‡], Mrinal Das[‡]
Tanveer Faruquie[†] and Chiranjib Bhattacharyya[‡]

† IBM India Research Lab
‡ Indian Institute of Science

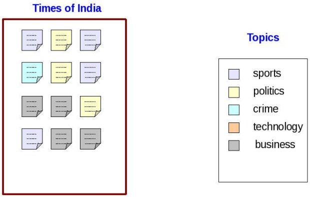
India

Motivation: Non Parametric Clustering



- Cluster articles in a single news paper according to topics
- Number of topics not known apriori

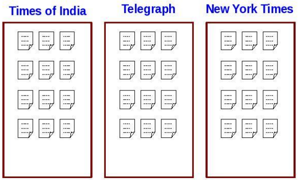
Motivation: Non Parametric Clustering



- Cluster articles in a single news paper according to topics
- Number of topics not known apriori
- Addressed by Dirichlet Process (DP)[1].

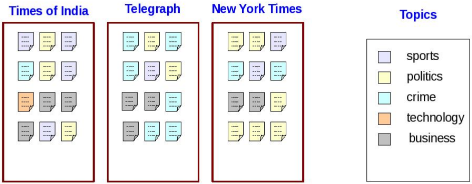
[1] C. Antoniak, 1974

Motivation: NP Clustering of Multiple Groups of Data



- Cluster articles in multiple newspapers according to topics.
- Topics shared between newspapers.
- Number of topics not known apriori.

Motivation: NP Clustering of Multiple Groups of Data

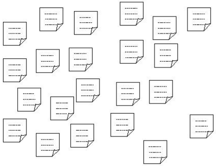


- Cluster articles in multiple newspapers according to topics.
- Topics shared between newspapers.
- Number of topics not known apriori.

- Addressed by Hierarchical Dirichlet Process(HDP)[2].

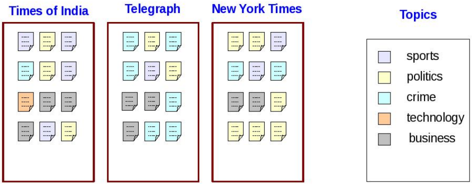
[2] Y. Teh et. al., 2006

Motivation: NP Clustering for Topics and Groups



- Association between articles and newspapers not observed.
- Cluster articles according to newspapers, as well as topics.
- Number of topics, newspapers not known apriori.

Motivation: NP Clustering for Topics and Groups

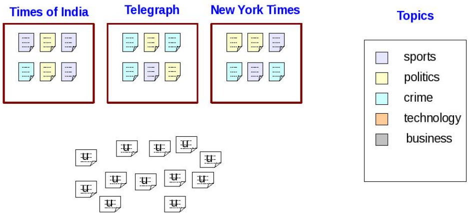


- Association between articles and newspapers not observed.
- Cluster articles according to newspapers, as well as topics.
- Number of topics, newspapers not known apriori.

- Challenging problem: Not directly addressed in literature.
- HDP-HMM[3] addresses sequential variant.

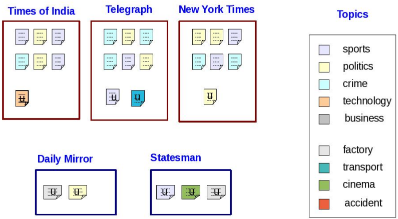
[3] E. Fox et. al., ICML 2008

Our Problem: NP Clustering with Partially Observed Groups



- One or more newspapers available with articles and topics.
- For new set of articles, determine newspaper and topic.
- Previously unseen newspapers and topics possible.

Our Problem: NP Clustering with Partially Observed Groups



- One or more newspapers available with articles and topics.
- For new set of articles, determine newspaper and topic.
- Previously unseen newspapers and topics possible.

Informal Problem Statement

- Traditional DP-based models consider *observed* groups
- We investigate DP-based models to infer *latent topics* as well as *latent groups* of *target* data items
- Conditional variant where topics and groups are observed for *source* collection

- First work studying DP-based conditional models

Related Work

Existing Models

- Dirichlet Process (DP)
- Hierarchical Dirichlet Process (HDP)
- Dependent Dirichlet Processes (DDP)[4]
- Pairwise-constrained DP (PC-DP)[5].

Variants of DP-based models with partial observations

- Sequential DP (Seq-DP)
- Partially observed DP (PO-DP)
- Partially observed HDP

[4] D. Lin et. al. NIPS 2010
[5] A. Vlachos et. al. ACL 2009

Data with Partially Observed Groups (POG Data)

Observed Group Data

$\mathcal{D}_o = \{x_i^o, \eta_i^o, z_i^o\}$

- features of data item $\{x_i^o\}$; observed
- group label $z_i^o \in \{1, \ldots, m\}$; observed
- topic label η_i^o; observed

Each observed group label: one Source

Target Data

$\mathcal{D}_u = \{x_i^u, \eta_i^u, z_i^u\}$

- x_i^u observed
- groups z_i^o, topics η_i^o not observed.

Problem Statement: Inference for POG Data

- Find posterior distribution of topic and group variables for target data items, conditioned on source data items
  - $P(\eta_1^u, z_1^u \dots \eta_n^u, z_n^u | x_1^u \dots x_n^u, \mathcal{D}_o)$
- Assume prior distribution of target topics $P(\eta_1^u \dots \eta_n^u)$ belongs to the Dirichlet Process family

- No assumption about distribution $P(\eta_1^o \dots \eta_n^o, z_i^o \dots z_n^o)$ of source groups and topics

Our Contributions

- First study of DP-based models for partially observed groups
- Propose POG-DP
- Conditional model
  - No generative assumptions for known groups
- Propose Combinatorial DP and partially observed variant
  - Finer coupling of topic selection probabilities within a group

Our Contributions

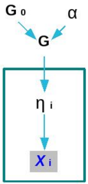
- First study of DP-based models for partially observed groups
- Propose POG-DP
- Conditional model
  - No generative assumptions for known groups
- Propose Combinatorial DP and partially observed variant
  - Finer coupling of topic selection probabilities within a group

- Developed 3 simple extensions of DP as baselines.
- Evaluated for 3 different applications.
- Outperform existing DP-based models and variants.

Background: Dirichlet Process (DP) Mixture Model

$DP(\alpha, G_0)$: scalar α, base distribution G_0.
Can be used as *non-parametric* prior for mixture models.

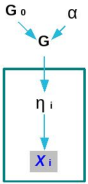
- $G \sim DP(\alpha, G_0)$
- For i^{th} data item
  - $\eta_i \sim G$
  - $x_i \sim F(\eta_i)$



Background: Dirichlet Process (DP) Mixture Model

$DP(\alpha, G_0)$: scalar α, base distribution G_0.
Can be used as *non-parametric* prior for mixture models.

- $G \sim DP(\alpha, G_0)$
- For i^{th} data item
  - $\eta_i \sim G$
  - $x_i \sim F(\eta_i)$



Conditional Distribution of n^{th} topic:

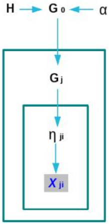
- $\eta_n \mid \eta_{1,n-1} \sim \alpha G_0 + \sum_{i=1}^{k} m_i \delta_{\phi_i}$

Works for a single group of data items

Background: Hierarchical Dirichlet Process (HDP)

One DP for each group of data items.
Coupled through shared base distribution.
Two level hierarchy of DPs.

- $G_0 \sim DP(\gamma, H)$
- For j^{th} group
  - $G_j \sim DP(\alpha, G_0)$
  - For i^{th} data item in j^{th} group
    - $\eta_{ij} \sim G_j; \quad x_{ij} \sim F(\eta_{ij})$

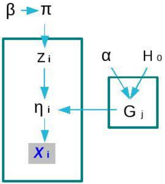


All group variables need to be observed

HDP for Unobserved Groups (UG-HDP)

- HDP generates data x_{ij} with known group memberships.
- For unobserved groups, introduce prior over groups.
- Non-parametric prior for unknown number of groups.
  - Stick-breaking prior $GEM(\beta)$

- For j^{th} group
  - $G_j \sim DP(\alpha, G_0)$
- For i^{th} data item
  - $z_i \sim mult(\pi)$
  - $\eta_i \sim G_{z_i}; x_i \sim F(\eta_i)$



Not specifically studied ; HDP-HMM : sequential variant

Modeling POG Data: Issues

- Groups z_i^u take known values $1 \dots m$ from D^o or new values
- Restrict to at most one new value $m+1$
  - Do not distinguish between different new groups

- Topics η_i^u take known values $\psi_1 \dots \psi_k$ from D^o or new values
- Restrict new topics only to new group $m+1$
  - Assume large volume of observed data D^o

Modeling POG Data: Issues II

- Base Distribution G_0 for new group G_{m+1}
- Same as the base distribution H_0 for existing groups?

$G_0 = H_0$

- Partially observed variant of UG-HDP (PO-HDP)
- Generative assumption on $G_1 \ldots G_m$ for known groups
  - Unnecessary, and possibly inappropriate
- Baseline for comparison

Modeling POG Data: Issues II

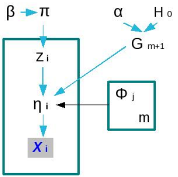
- Base Distribution G_0 for new group G_{m+1}
- Same as the base distribution H_0 for existing groups?

$G_0 \neq H_0$

- Proposed model – POG-DP
- Benefit: D_o conditionally independent of $G_1 \dots G_m$ given η^o
  - No generative assumptions on observed data D_o
- Price: Topics under G_0 distinct from existing topics η^o
  - Not restrictive for most applications

Dirichlet Processes for POG Data (POG-DP)

- For $k = 1, \ldots, m$ sources
  - $\mu^k \sim Dir(\alpha^k)$, $k = 1 \ldots m$
- For each i^{th} data item
  - $z_i^u \sim mult(\pi)$
  - $\eta_i^u \sim \sum_j \mu_j^k \delta_{\phi_j^k}$, $\quad z_i^u = k, k = 1 \ldots m$
  - $\eta_i^u \sim G_{m+1}$, $\quad z_i^u = m + 1$



- Distribution over the existing groups is parametrized
- These parameters are learned over both D_o and D_u
- PO-HDP has empirical distribution in D_o

POG-DP: Conditional Distribution

Conditional Distribution of n^{th} topic in POG-DP

$$\begin{aligned} &\eta_n^u | \eta^o, \eta_{1:n-1}^u, z_{1:n-1}^u, \alpha, \beta, H_0 \\ \sim \quad & (\beta + n_{\cdot}^{m+1}) \left(\textstyle\sum_{i=1}^{K^{m+1}} n_i^{m+1} \delta_{\phi_i^{m+1}} + \alpha^{m+1} H_0 \right) \\ & + \textstyle\sum_{k=1}^{m} (\beta + n_{\cdot}^{k}) \left(\textstyle\sum_{i=1}^{K^k} (\alpha^i + n_i^k) \delta_{\phi_i^k} \right) \end{aligned}$$

POG-DP: Conditional Distribution

Conditional Distribution of n^{th} topic in POG-DP

$$\begin{aligned} &\eta_n^u | \eta^o, \eta_{1:n-1}^u, z_{1:n-1}^u, \alpha, \beta, H_0 \\ \sim \quad & (\beta + n_{\cdot}^{m+1}) \left(\sum_{i=1}^{K^{m+1}} n_i^{m+1} \delta_{\phi_i^{m+1}} + \alpha^{m+1} H_0 \right) \\ & + \sum_{k=1}^{m} (\beta + n_{\cdot}^{k}) \left(\sum_{i=1}^{K^k} (\alpha^i + n_i^k) \delta_{\phi_i^k} \right) \end{aligned}$$

Conditional distribution of n^{th} topic in PO-HDP

$$\eta_{jn} \mid \eta_{1:j-1}, \eta_{j1} \ldots \eta_{j,n-1}; \alpha, H \sim \alpha \left(\sum_k m_k \delta_{\psi_k} + \gamma H \right) + \sum_i n_i^j \delta_{\theta_i^j}$$

- In PO-HDP, topic selection probabilities decoupled
- In POG-DP, they are coupled within each group

Coupling of Topics Selection Probabilities

- POG-DP couples all topics within a group
- Appropriate when all topics in a group form a coherent set
- Ideally, coupling only for coherent topic *subsets* in a group
- In general, involves searching over subsets; hard

- For POG data, assume *group intersections* form coherent subsets
- Introduce coupling of topics only within group intersections

Combinatorial Dirichlet Processes for POG Data (POG-CDP)

- Group Intersection $\equiv$ Combination of existing groups (sources)
- Selection probabilities for group combinations instead of individual groups
- Group label z_i^u in D^u now a binary vector
- Parametrize prior distribution over z_i^u using independence of groups

Utility:

- Coupling of topics only within group intersections
- Answer richer queries
  - Predict *combination of existing groups* for new data item

Approximate Inference using Gibbs Sampling

Collapsed Gibbs Sampling

- Repeated sample topic η_i^u and group z_i^u for each target data item from conditional distribution

Block sample (η_i^u, z_i^u) for POG-DP

- Faster convergence

Large space of possible values of z_i^u for POG-CDP

- Sample each position of vector individually
- Slower; but scales with number of sources

Evaluation – Application I : Vernacular news analysis

- Topics from news articles in English and Hindi given.
- Task is to find out topics in Bengali that corresponds to some topics in English or Hindi.
- Also find out topics reported exclusively in Bengali.

- Bengali[6], Hindi[7] and English[8] news from 01-2007 to 12-2007.
- 3000 documents over a vocabulary of 5000 words.

[6] www.anandabazar.com
[7] in.jagran.yahoo.com/epaper
[8] www.telegraphindia.com

Evaluation – Application II : Customer feedback analysis

- Finding issues from customer feedbacks.
- Given previously analyzed collection of surveys for 2 Web-service provider companies.
  - Individual feedbacks are labelled with issues in the survey.
- Task is to label feedbacks for Tele-comm company with known issues from Web-service companies or new issues exclusive to Tele-comm company.

- Tele-communication company as target, Web-service provider companies as source.
- 500 documents over 1200 words.

Evaluation – Application III : News-group analysis

- Given a collection of postings categorized into news topics.
- Task is to label new discussions with known topics or breaking discussion topics.

- 20 Newsgroups dataset.
- 14,000 postings over 5000 words.

Evaluation – Baselines

- DP mixture model, Partially Observed HDP (PO-HDP)

- Sequential DP (Seq-DP) – source items known.
- Partially observed DP (PO-DP) – source topics known.
- Pair-constrained DP (PC-DP) – pair-wise must-link and cannot-link constraints known for source.

- Merged source POG-DP (msPOG-DP) – multiple sources merged into one

Evaluation – Results

Single source

	DP	Seq DP	PO-DP	PC-DP	PO-HDP	POG-DP
VerNewsAna	0.25	0.51	**0.63**	0.59	0.59	0.71
CustFeedAna	0.26	0.26	0.33	0.31	**0.32**	0.39
NewsGrpAna	0.34	0.34	0.44	0.44	**0.45**	0.53

POG-DP outperforms other baselines.

- Modeling source is inappropriate.
- Coupling among topics in a group helps.

Evaluation – Results

Multiple sources

	PO-DP	PC-DP	PO-HDP	msPOG-DP	POG-DP	POG-CDP
VerNewsAna	0.71	0.69	**0.73**	0.81	0.81	0.86
CustFeedAna	0.33	0.31	**0.34**	0.37	0.33	0.43
NewsGrpAna	0.60	0.60	**0.63**	0.67	0.62	0.69

POG-DP outperforms other baselines.
Combinatorial model (POG-CDP) better for overlapped sources.

- Finer grained coupling of topics inside groups useful for overlapped sources.

Conclusion

- Proposed Dirichlet Process for partially observed groups.
- Models target data conditioned on the source data.
- Proposed combinatorial DP to model overlapping sources.

- Usefulness verified over various baselines on real life datasets.
- POG-DP outperforms variants of DP and HDP.
- For multiple overlapping source case, POG-CDP is the best.