Knowledge Transfer Using Latent Variable Models

Ayan Acharya

UT Austin, Department of ECE

July 21, 2015

Motivation & Theme

- **Motivation**
  - Labeled data is sparse in applications like document categorization and object recognition.
  - Distribution of data changes across domains or over time.
- **Theme**
  - Shared low dimensional space for transferring information across domains
  - Careful adaptation of the model parameters to fit new data

Transfer Learning

- **Transfer Learning**
  - Concurrent knowledge transfer (or multitask learning): multiple domains learnt simultaneously
  - Continual knowledge transfer (or sequential knowledge transfer): models learnt in one domain are carefully adapted to other domains

Active Learning

- only the most informative examples are queried from the unlabeled pool

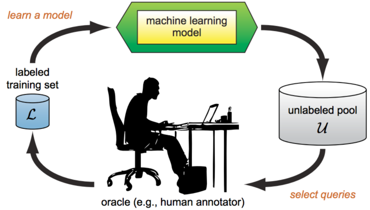
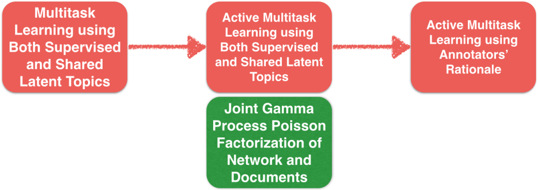


Figure: Illustration of Active Learning (Pic Courtesy: Burr Settles)

Section Outline



- Multitask Learning Using Both Supervised and Latent Shared Topics (ECML 2013)
- Active Multitask Learning Using Both Supervised and Latent Shared Topics (NIPS13 Topic Model Workshop, SDM 2014)
- Active Multitask Learning with Annotator's Rationale
- Joint Modeling of Network and Documents using Gamma Process Poisson Factorization (KDD SRS Workshop 2015, ECML 2015)

Multitask Learning Using Both Supervised and Latent Shared Topics
(ECML 2013)

Problem Setting

- In training corpus each document/image belongs to a known class and has a set of attributes (supervised topics).
- aYahoo – **Classes**: carriage, centaur, bag, building, donkey, goat, jetski, monkey, mug, statue, wolf, and zebra; **Attributes**: “has head”, “has wheel”, “has torso” and 61 others
- ACM Conf. – **Classes**: ICML, KDD, SIGIR, WWW, ISPD, DAC; **Attributes**: keywords
- Train models using words, supervised topics and class labels, and classify completely unlabeled test data (no supervised topic or class label)

Class: Carriage

Attributes:
“has wheel?” Yes.
“has wood?” Yes.

Doubly Supervised Laten Dirichlet Allocation (DSLDA)

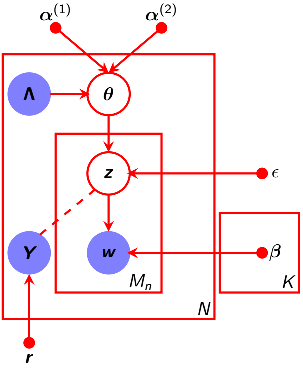


Figure: DSLDA – Supervision at both topic and category level

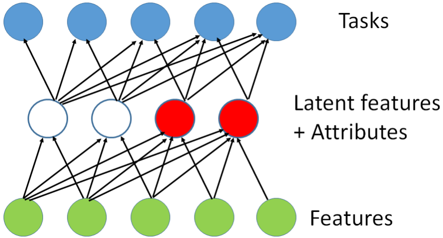
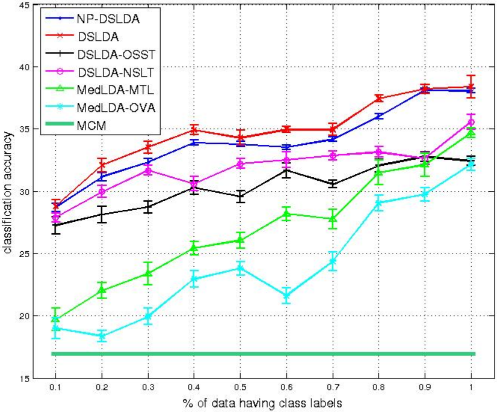


Figure: Visual Representation

- Variational EM used for inference and learning

Multitask Learning Results: aYahoo



- observation: multitask learning method with latent and supervised topics performs better compared to other methods

Active Multitask Learning Using Both Supervised and Latent Shared Topics
(NIPS13 Topic Model Workshop, SDM 2014)

Problem Setting

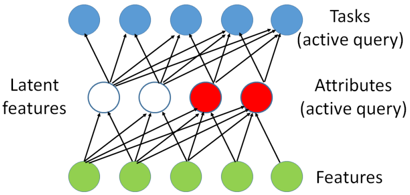
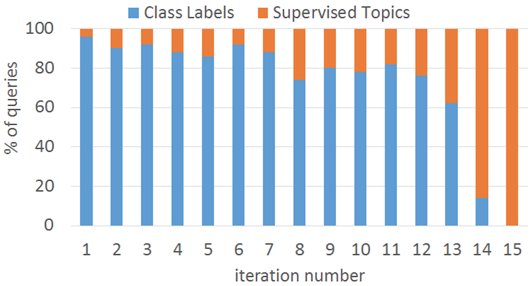


Figure: Visual Representation of Active Doubly Supervised Latent Dirichlet Allocation (Act-DSLDA)

- An active MTL framework that can use and query over both attributes and class labels
- Active learning measure: expected error reduction
- Batch mode: variational EM, online SVM
- Active selection mode: incremental EM, online SVM

Active Multitask Learning Results: ACM Conf. Query Distribution



- observation: more category labels (*e.g.* KDD, ICML, ISPD) queried in the initial phase, more attributes (keywords) queried later on

Active Multitask Learning Using Annotators' Rationale

Problem Setting

- An active multitask learning framework that can query over attributes, class labels and their rationales

Results for Active Multitask Learning with Rationale: ACM Conf.

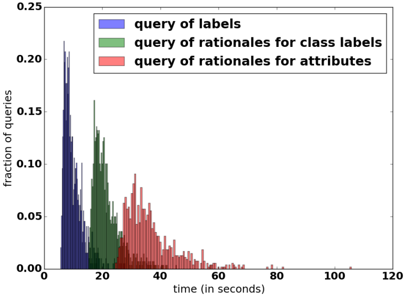


Figure: Query Distribution

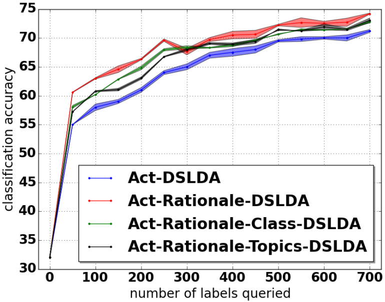


Figure: Learning Curve

- observation: active learning method with rationales and supervised topics performs much better compared to baselines

Active Rationale Results: ACM Conf.

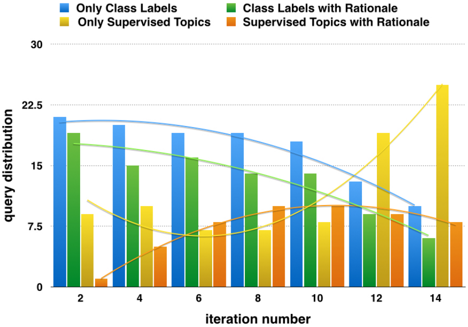
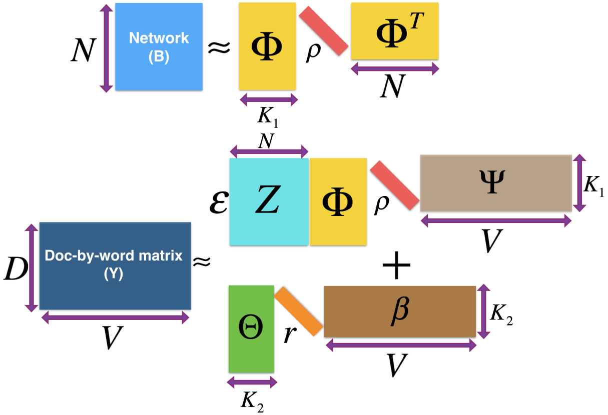


Figure: Query Distribution: ACM Conf.

- observation: more labels with rationales queried in the initial phase

Gamma Process Poisson Factorization for Joint Modeling of Network and Documents
(ECML 2015)

GPPF for Joint Network and Topic Modeling (J-GPPF)

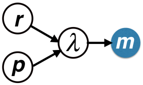


Characteristics of J-GPPF

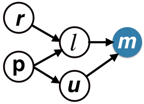
- Poisson factorization: $Y_{dw} \sim \text{Pois}(\langle \boldsymbol{\theta}_d, \boldsymbol{\beta}_w \rangle)$, samples latent counts corresponding to non-zeros only
- Joint Poisson factorization for imputing a graph
- Hierarchy of Gamma priors for less sensitivity towards initialization
- Non-parametric modeling with closed form inference updates

Negative Binomial Distribution (NB)

- Number of heads seen until r number of tails occurs while tossing a biased coin with probability of head p (or, number of successes before r failures in successive Bernoulli trials): $m \sim \mathrm{NB}(r, p)$
- $m \sim \mathrm{Poisson}(\lambda), \lambda \sim \mathrm{Gam}(r, p)$ – Gamma-Poisson Construction
- $m \sim \sum_{t=1}^{\ell} u_t,\ u_t \sim \mathrm{Log}(p),\ \ell \sim \mathrm{Poisson}(-r \log(1-p))$ – Compound Poisson Construction



Gamma-Poisson Construction



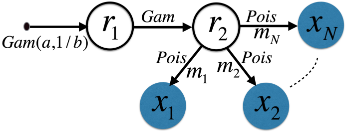
Compound Poisson Construction

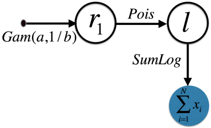
Figure: Constructions of Negative Binomial Distribution

Lemma

If $m \sim NB(r, p)$ is represented under its compound Poisson representation, then the conditional posterior of ℓ given m and r is given by $(\ell|m, r) \sim CRT(m, r)$, which can be generated via $\ell = \sum_{n=1}^{m} z_n, z_n \sim Bernoulli(r/(n-1+r))$.

Inference of Shape Parameter of Gamma Distribution





- $x_i \sim \mathrm{Pois}(m_i r_2)\ \forall i \in \{1, 2, \cdots, N\}$, $r_2 \sim \mathrm{Gam}(r_1, 1/d)$, $r_1 \sim \mathrm{Gam}(a, 1/b)$.

Lemma

If $x_i \sim Pois(m_i r_2)\ \forall i$, $r_2 \sim Gam(r_1, 1/d)$, $r_1 \sim Gam(a, 1/b)$, then $(r_1|-) \sim Gam(a + \ell, 1/(b - log(1 - p)))$ where $(\ell|\{x_i\}_i, r_1) \sim CRT(\sum_i x_i, r_1)$, $p = \sum_i m_i/(d + \sum_i m_i)$.

J-GPPF Results: Real-world Data

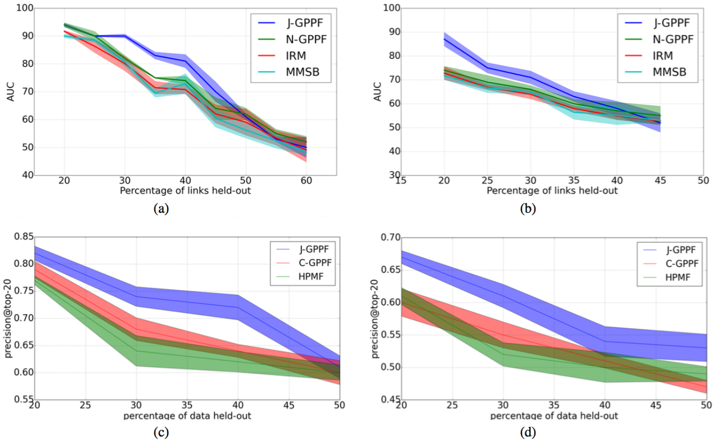
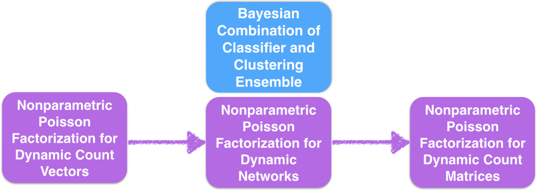


Figure: (a) AUC on NIPS, (b) AUC on Twitter, (c) MAP on NIPS, (d) MAP on Twitter

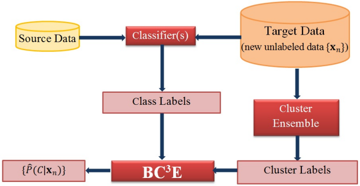
Section Outline



- Bayesian Combination of Classification and Clustering Ensembles (SDM 2013)
- Nonparametric Dynamic Models
  - Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices (AISTATS 2015)
  - Nonparametric Dynamic Relational Model (KDD MiLeTs Workshop 2015)
  - Nonparametric Dynamic Count Matrix Factorization

Bayesian Combination of Classifier and Clustering Ensemble
(SDM 2013)

Bayesian Combination of Classifier and Clustering Ensemble



	$\mathbf{w}_1^{(1)}$	$\mathbf{w}_2^{(1)}$	$\cdots$	$\mathbf{w}_{r_1}^{(1)}$
$\mathbf{x}_1$	2	3	$\cdots$	1
$\mathbf{x}_2$	1	3	$\cdots$	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{x}_N$	2	3	$\cdots$	3

Table: From Classifiers

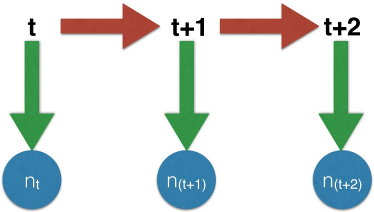
	$\mathbf{w}_1^{(2)}$	$\mathbf{w}_2^{(2)}$	$\cdots$	$\mathbf{w}_{r_2}^{(2)}$
$\mathbf{x}_1$	4	5	$\cdots$	4
$\mathbf{x}_2$	2	4	$\cdots$	4
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{x}_N$	2	4	$\cdots$	2

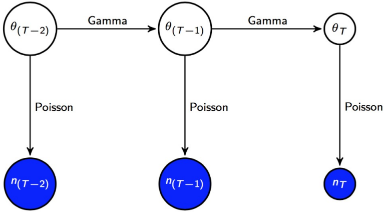
Table: From Clusterings

- Prior Work – C^{3}E: An Optimization Framework for Combining Ensembles of Classifiers and Clusterers with Applications to Nontransductive Semisupervised Learning and Transfer Learning (Acharya *et. al.*, 2014), Appeared in ACM Transaction on KDD

Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices
(AISTATS 2015)

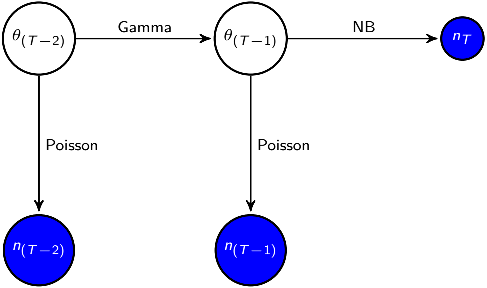
Gamma Poisson Autoregressive Model





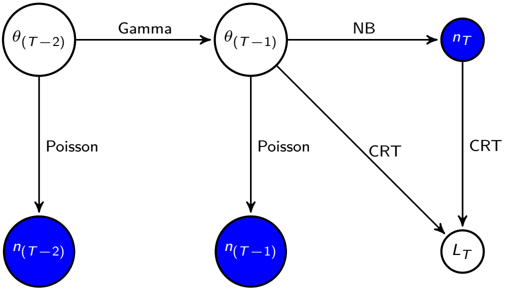
- $\theta_t \sim \text{Gam}(\theta_{(t-1)}, 1/c)$, $n_t \sim \text{Pois}(\theta_t)$.
- Gamma-Gamma construction breaks conjugacy

Inference in Gamma Poisson Autoregressive Model



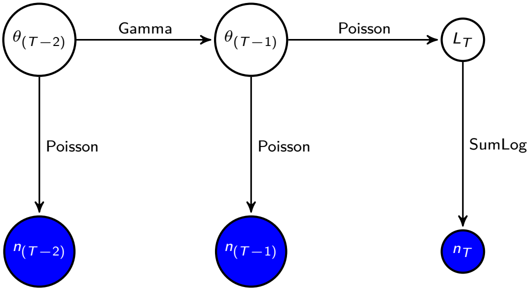
- use Gamma-Poisson construction of NB
- $n_T \sim \text{NB}(\theta_{(T-1)}, 1/(c+1))$.

Inference in Gamma Poisson Autoregressive Model



- $n_T \sim \mathrm{NB}(\theta_{(T-1)}, 1/(c+1))$. Augment $L_T \sim \mathrm{CRT}(n_T, \theta_{(T-1)})$.

Inference in Gamma Poisson Autoregressive Model



- use compound poisson construction of NB
- $n_T \sim \sum_{t=1}^{L_T} \mathrm{Log}(1/(c+1)), L_T \sim \mathrm{Poisson}(\theta_{(T-1)} \log((c+1)/c)).$
- Gamma-Poisson construction facilitates closed form Gibbs sampling.

Gibbs Sampling in Gamma Poisson Autoregressive Model

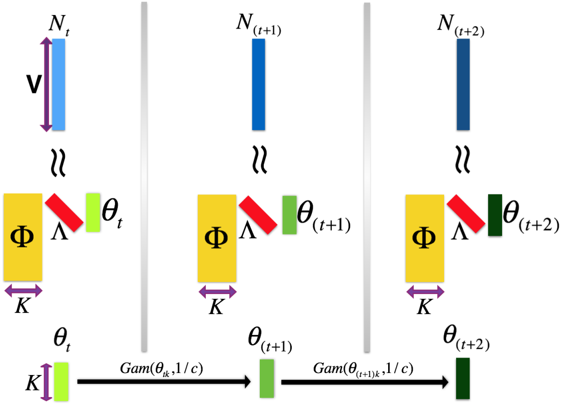
Backward sampling of augmented variables from $t = T$ to 1,

$$L_t \sim \mathrm{CRT}(n_t, \theta_{(t-1)}).$$

Forward sampling of latent rates for $t = 1$ to T,

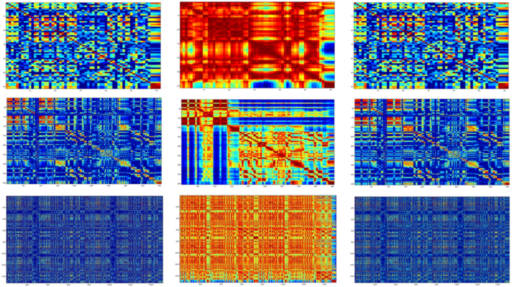
$$\theta_t \sim \mathrm{Gam}(\theta_{(t-1)} + n'_t, p_t),$$
$$p_t = 1/(1 + c - \log(p_{(t-1)})), \;\; n'_t = n_t + L_{(t+1)}.$$

Gamma Process Dynamic Poisson Factor Analysis (GPDPFA)



- $n_{wt} = \sum_k n_{wtk}$, $n_{wtk} \sim \text{Pois}(\lambda_k \phi_{wk} \theta_{tk})$.
- $\lambda_k \sim \text{Gam}(r_0/K, 1/c)$, $\phi_k \sim \text{Dir}(\eta_1, \cdots, \eta_V)$, $\theta_{tk} \sim \text{Gam}(\theta_{(t-1)k}, 1/c_t)$.

Results from Gamma Process Dynamic Poisson Factor Analysis



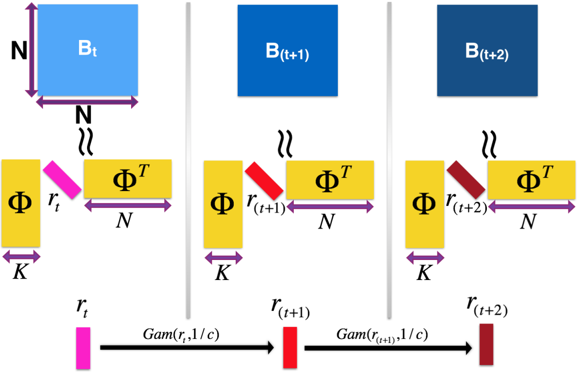
(a) (b) (c)

Figure: (a) Correlation of original vectors, (b) Correlation in the latent space, (c) Correlation between original and derived vectors

Data	Model	MP	MR	PP
STU	GP-DPFA	**0.2230**±0.0009	0.1976±0.0004	**0.1891**±0.0028
	DRFM	0.2171±0.0025	**0.1978**±0.0014	0.1773±0.0104
	Baseline	0.1018±0.0216	0.1329±0.0173	0.0612±0.0328
Conf.	GP-DPFA	0.3020±0.0004	**0.2681**±0.0003	**0.2412**±0.0004
	DRFM	**0.3023**±0.0005	0.2566±0.0006	0.2410±0.0006
	Baseline	0.1241±0.0194	0.1107±0.0131	0.1014±0.0370

Nonparametric Dynamic Relational Model
(KDD MiLeTs Workshop 2015)

Gamma Process Poisson Factorization for Dynamic Network Modeling (D-NGPPF)



- $b_{tnm} = I_{\{x_{tnm} \geq 1\}}$, $x_{tnm} = \sum_k x_{tnmk}$, $x_{tnmk} \sim \text{Pois}(r_{tk}\phi_{nk}\phi_{mk})$.
- $r_{tk} \sim \text{Gam}(r_{(t-1)k}/K, 1/c)$, $c \sim \text{Gam}(g_0, 1/h_0)$, $r_{0k} \sim \text{Gam}(\gamma_0, 1/f_0)$.
- $\phi_k \sim \prod_{n=1}^{N} \text{Gam}(a_0, 1/c_n)$, $c_n \sim \text{Gam}(c_0, 1/d_0)$.

Results from Dynamic Network Modeling: Synthetic Data

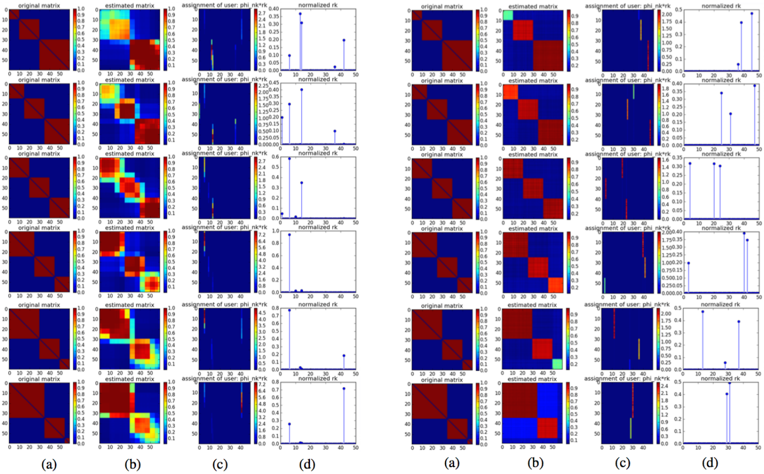


Figure: Results from dynamic model (left) and non-dynamic model (right)

Results from Dynamic Network Modeling: Real-world Data

- DSBM: Dynamic stochastic block model
- N-GPPF: Gamma Process Poisson factorization for networks
- MMSB: Mixed membership stochastic block model

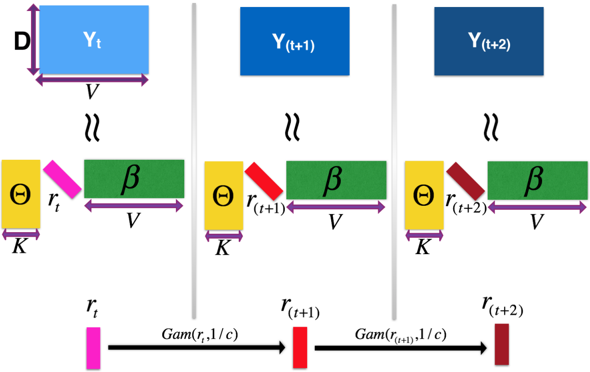
Dataset	D-NGPPF	DSBM	N-GPPF	MMSB
NIPS	$\mathbf{0.797} \pm 0.016$	0.780 ± 0.010	0.766 ± 0.012	0.740 ± 0.009
DBLP	$\mathbf{0.836} \pm 0.013$	0.810 ± 0.013	0.756 ± 0.020	0.749 ± 0.014
Infocom	$\mathbf{0.907} \pm 0.008$	0.901 ± 0.006	0.856 ± 0.011	0.831 ± 0.006

Figure: AUC Results

Method	D-NGPPF	DSBM	N-GPPF	MMSB
Complexity	$O((S+N+T)K)$	$O(N^2KT)$	$O((S+N)KT)$	$O(N^2KT)$

Nonparametric Dynamic Count Matrix Factorization

Gamma Process Poisson Factorization for Dynamic Count Matrix Factorization (D-CGPPF)



- $y_{tdw} = \sum_k y_{tdwk}$, $y_{tdwk} \sim \text{Pois}(r_{tk}\theta_{dk}\beta_{wk})$.
- $r_{tk} \sim \text{Gam}(r_{(t-1)k}/K, 1/c)$, $\theta_k \sim \prod_{d=1}^{D} \text{Gam}(a_0, 1/c_d)$, $\beta_k \sim \prod_{w=1}^{V} \text{Gam}(b_0, 1/c_w)$.

Results from Dynamic Count Matrix Factorization

- BPTF: Bayesian probabilistic tensor factorization
- C-GPPF: Gamma Process Poisson factorization for modeling count matrix

Dataset	D-CGPPF	BPTF	C-GPPF
Movielens100K	**0.597** $\pm$ 0.023	0.512 $\pm$ 0.010	0.238 $\pm$ 0.047
Movielens1M	**0.641** $\pm$ 0.010	0.632 $\pm$ 0.008	0.521 $\pm$ 0.019
Netflix	**0.490** $\pm$ 0.008	0.418 $\pm$ 0.002	0.251 $\pm$ 0.039

Figure: Precision@top-50%

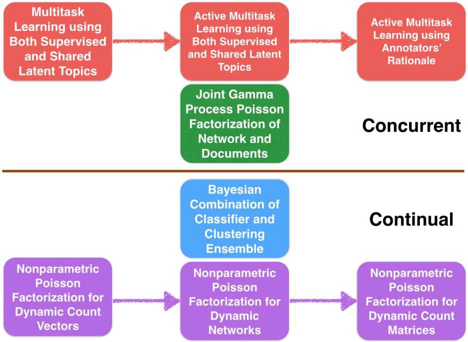
Dataset	D-CGPPF	BPTF	C-GPPF
Movielens100K	**0.714** $\pm$ 0.016	0.703 $\pm$ 0.010	0.455 $\pm$ 0.012
Movielens1M	0.721 $\pm$ 0.013	**0.725** $\pm$ 0.013	0.585 $\pm$ 0.020
Netflix	**0.613** $\pm$ 0.007	0.592 $\pm$ 0.011	0.451 $\pm$ 0.018

Figure: NDCG@top-50%

Method	D-CGPPF	BPTF	C-GPPF
Complexity	$O((S+D+V+T)K)$	$O(DVK^2+(D+V+T)K^3)$	$O((S+D+V)KT)$

Conclusion and Future Works

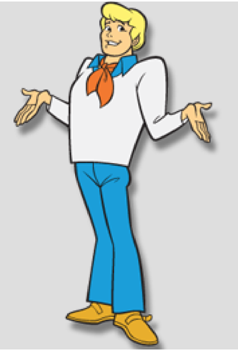
Conclusion:



Future Works:

- Dynamic Topic Model
- Dynamic Tensor Factorization for analysis of EHR data
- Distributed Poisson Factorization

Questions?



Publications

1. **Acharya, Ayan**, Teffer, Dean, Zhou, Mingyuan, and Ghosh, Joydeep, Network Discovery and Recommendation via Joint Network and Topic Modeling, KDD Workshop on Social Recommender Systems, 2015. [.pdf]
2. **Acharya, Ayan**, Saha, Avijit, Zhou, Mingyuan, Ghosh, Joydeep, and Teffer, Dean, Nonparametric Dynamic Network Model, KDD Workshop on Mining and Learning from Time Series, 2015. [.pdf]
3. **Acharya, Ayan**, Ghosh, Joydeep, and Zhou, Mingyuan, Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices, Proc. of AISTATS, 2015. [.pdf]
4. Coletta, Luiz Fernando, Ponti, Moacir, Hruschka, Eduardo R., **Acharya, Ayan**, and Ghosh, Joydeep, Combining Clustering and Active Learning for the Detection and Learning of New Image Classes, International Journal of Image and Vision Computing (submitted), 2015. [.pdf]
5. **Acharya, Ayan**, Teffer, Dean, Henderson, Jette, Tyler, Marcus, Zhou, Mingyuan, and Ghosh, Joydeep, Gamma Process Poisson Factorization for Joint Modeling of Network and Documents, ECML, 2015. [.pdf]
6. Ghosh, Joydeep and **Acharya, Ayan**, A Survey of Consensus Clustering, Appearing in Handbook of Cluster Analysis, 2015. [.pdf]
7. Coletta, Luiz F. S., Hruschka, Eduardo R., **Acharya, Ayan**, and Ghosh, Joydeep, Using metaheuristics to optimize the combination of classifier and cluster ensembles, Appearing in Integrated Computer-Aided Engineering, 2015. [.pdf]
8. **Acharya, Ayan**, Mooney, Raymond J., and Ghosh, Joydeep, Active Multitask Learning Using Both Latent and Supervised Shared Topics, Appearing in Pattern Recognition: from Classical to Modern Approaches, pp., 2015. [.pdf]
9. **Acharya, Ayan**, Hruschka, Eduardo R., Ghosh, Joydeep, and Acharyya, Sreangsu, An Optimization Framework for Combining Ensembles of Classifiers and Clusterers with Applications to Non-transductive Semi-Supervised Learning and Transfer Learning, In ACM Transactions on Knowledge Discovery from Data, September, 2014 [.pdf].

Publications

10. Coletta, Luiz Fernando, Hruschka, Eduardo R., **Acharya, Ayan**, and Ghosh, Joydeep, A Differential Evolution Algorithm to Optimize the Combination of Classifier and Cluster Ensembles, International Journal of Bio-Inspired Computation, 2014.
11. **Acharya, Ayan**, Mooney, Raymond J., and Ghosh, Joydeep, Active Multitask Learning Using Both Latent and Supervised Shared Topics, Proceedings of the 2014 SIAM International Conference on Data Mining, pp.190-198, 2014.
12. **Acharya, Ayan**, Hruschka, Eduardo R., Ghosh, Joydeep, Sarwar, Badrul, and Ruvini, Jean-David, Probabilistic Combination of Classifier and Cluster Ensembles for Non-transductive Learning, SDM, 2013 [.pdf].
13. Gunasekar, Suriya, **Acharya, Ayan**, Gaur, Neeraj, and Ghosh, Joydeep, Noisy Matrix Completion Using Alternating Minimization, ECML PKDD, Part II, LNAI 8189, pp.194-209, 2013 [.pdf].
14. **Acharya, Ayan**, Rawal, Aditya, Mooney, Raymond J., and Hruschka, Eduardo R., Using Both Supervised and Latent Shared Topics for Multitask Learning, ECML PKDD, Part II, LNAI 8189, pp.369-384, 2013 [.pdf].
15. Ghosh, Joydeep and **Acharya, Ayan**, Cluster Ensembles: Theory and Applications, in Data Clustering: Algorithms and Applications, 2013 [.pdf].
16. **Acharya, Ayan**, Mooney, Raymond J., Ghosh, Joydeep, Active Multitask Learning Using Doubly Supervised Latent Dirichlet Allocation, NIPS Topic Model Workshop, 2013 [.pdf].
17. Ghosh, Joydeep and **Acharya, Ayan**, A Survey of Consensus Clustering, Appearing in Handbook of Cluster Analysis, 2013 [.pdf].
18. Coletta, Luiz Fernando, Hruschka, Eduardo R., **Acharya, Ayan**, and Ghosh, Joydeep, Towards the Use of Metaheuristics for Optimizing the Combination of Classifier and Cluster Ensembles, Appearing in 11th Brazilian Congress (CBIC) on Computational Intelligence, 2013, [.pdf].
19. **Acharya, Ayan**, Hruschka, Eduardo R., Ghosh, Joydeep, and Acharyya, Sreangsu, Transfer Learning with Cluster Ensembles, Journal of Machine Learning Research - Proceedings Track, 27 , pp.123-132, 2012 [.pdf].

Publications

20. **Acharya, Ayan**, Lee, Jangwon, and Chen, An, Real Time Car Detection and Tracking in Mobile Devices, IEEE International Conference on Connected Vehicles and Expo, 2012 [.pdf].
21. Ghosh, Joydeep and **Acharya, Ayan**, Cluster ensembles, Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery, 1 (4) , pp.305-315, 2011 [.pdf].
22. **Acharya, Ayan**, Hruschka, Eduardo R., Ghosh, Joydeep, and Acharyya, Sreangsu, C^3E: A Framework for Combining Ensembles of Classifiers and Clusterers, MCS, pp.269-278, 2011 [.pdf].
23. **Acharya, Ayan**, Hruschka, Eduardo R., and Ghosh, Joydeep, A Privacy-Aware Bayesian Approach for Combining Classifier and Cluster Ensembles, SocialCom/PASSAT, pp.1169-1172, 2011 [.pdf].

Baselines: Multitask learning experiments

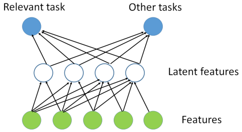


Figure: MedLDA-OVA

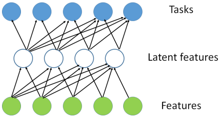


Figure: MedLDA-MTL

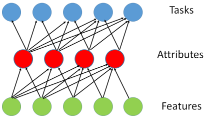


Figure: DSLDA-OSST

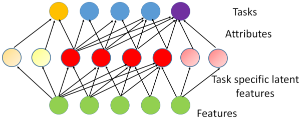


Figure: DSLDA-NSLT

Baselines: Active multitask learning experiments

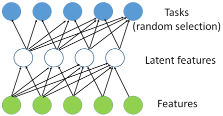


Figure: Random MedLDA-MTL (R-MedLDA-MTL)

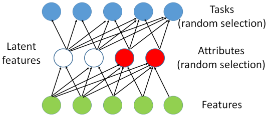


Figure: Random DSLDA (R-DSLDA)

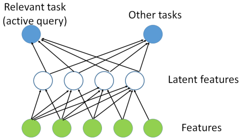


Figure: Active MedLDA-OVA (Act-MedLDA-OVA)

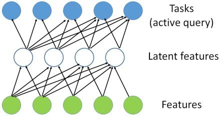
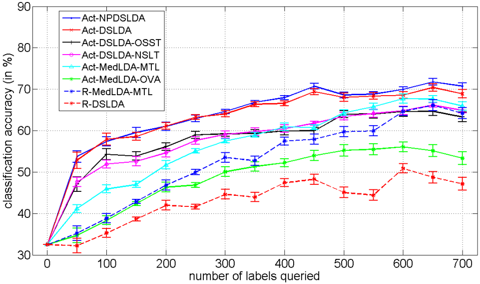


Figure: Active MedLDA-MTL (Act-MedLDA-MTL)

Active multitask learning results: ACM Conf. learning curves



- observation: active learning method with both latent and supervised topics performs much better than other baselines which do not use active learning and/or two different sets of topics

Gamma Process (GP)

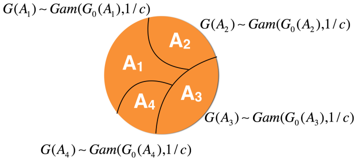


Figure: Illustration of Gamma Process

- The Gamma Process $G \sim \Gamma\mathrm{P}(G_0, c)$ is a completely random measure defined on the product space $\mathbb{R}_+ \times \Omega$ with concentration parameter c and a finite and continuous base measure G_0 over a complete separable metric space Ω, such that $G(A_i) \sim \mathrm{Gam}(G_0(A_i), 1/c)$ are independent gamma random variables for disjoint partition $\{A_i\}_i$ of Ω.

Gamma Process (GP)

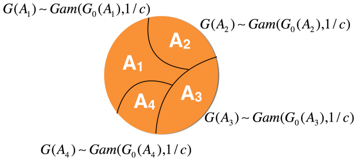


Figure: Illustration of Gamma Process

- The Gamma Process $G \sim \Gamma\mathrm{P}(G_0, c)$ is a completely random measure defined on the product space $\mathbb{R}_+ \times \Omega$ with concentration parameter c and a finite and continuous base measure G_0 over a complete separable metric space Ω, such that $G(A_i) \sim \mathrm{Gam}(G_0(A_i), 1/c)$ are independent gamma random variables for disjoint partition $\{A_i\}_i$ of Ω.
- $G = \sum_{k=1}^{\infty} r_k \delta_{\omega_k}, (r_k, \omega_k) \overset{iid}{\sim} r^{-1} e^{-cr} dr G_0(d\omega)$.

Gamma Process (GP)

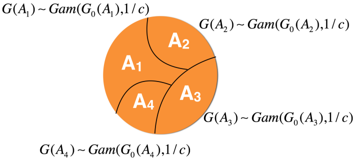


Figure: Illustration of Gamma Process

- The Gamma Process $G \sim \Gamma P(G_0, c)$ is a completely random measure defined on the product space $\mathbb{R}_+ \times \Omega$ with concentration parameter c and a finite and continuous base measure G_0 over a complete separable metric space Ω, such that $G(A_i) \sim \mathrm{Gam}(G_0(A_i), 1/c)$ are independent gamma random variables for disjoint partition $\{A_i\}_i$ of Ω.
- $G = \sum_{k=1}^{\infty} r_k \delta_{\omega_k}, (r_k, \omega_k) \overset{iid}{\sim} r^{-1} e^{-cr} dr G_0(d\omega)$.
- Finite approximation of ΓP:

$$G = \sum_{k=1}^{K} r_k \delta_{\omega_k}, (r_k, \omega_k) \overset{iid}{\sim} r^{(\gamma_0/K-1)} e^{-cr} dr G_0(d\omega), \ \gamma_0 = G_0(\omega).$$

Chinese Restaurant Table Distribution (CRT)

- Chinese Restaurant Process: occupy an empty table w.p. γ_0 or occupy a table w.p. proportional to the number of customers in that table
- m : number of data points (number of customers)
- K : number of distinct atoms (number of tables)

$$\Pr(K = l|m, \gamma_0) = \frac{\Gamma(\gamma_0)}{\Gamma(m+\gamma_0)}|s(m,l)|\gamma_0^l, \quad l = 0, 1, \cdots, m,$$

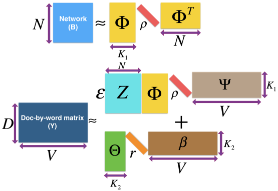
where, $s(m, l)$ is the Stirling number of the first kind

Figure: Illustration of Chinese Restaurant Table Distribution

Lemma

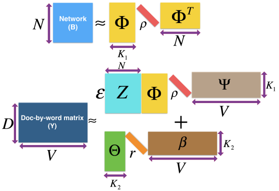
If $m \sim NB(r, p)$ is represented under its compound Poisson representation, then the conditional posterior of ℓ given m and r is given by $(\ell|m, r) \sim CRT(m, r)$, which can be generated via $\ell = \sum_{n=1}^{m} z_n, z_n \sim Bernoulli(r/(n-1+r))$.

GPPF for Joint Network and Topic Modeling (J-GPPF)



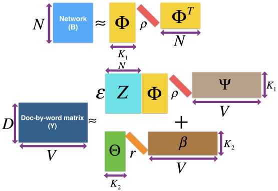
- $b_{nm} = I_{\{x_{nm} \geq 1\}}, x_{nm} \sim \text{Pois}(\sum_{k_B=1}^{K_1} \rho_{k_B} \phi_{nk_B} \phi_{mk_B})$, $\rho_{k_B} \sim \text{Gam}(\gamma_B / K_B, 1/c_B)$, $\phi_{k_B} \sim \prod_{n=1}^{N} \text{Gam}(a_B, 1/\sigma_n)$.

GPPF for Joint Network and Topic Modeling (J-GPPF)



- $b_{nm} = I_{\{x_{nm} \geq 1\}}, x_{nm} \sim \mathrm{Pois}(\sum_{k_{\boldsymbol{B}}=1}^{K_1} \rho_{k_{\boldsymbol{B}}} \phi_{nk_{\boldsymbol{B}}} \phi_{mk_{\boldsymbol{B}}})$, $\rho_{k_{\boldsymbol{B}}} \sim \mathrm{Gam}(\gamma_{\boldsymbol{B}}/K_{\boldsymbol{B}}, 1/c_{\boldsymbol{B}})$, $\phi_{k_{\boldsymbol{B}}} \sim \prod_{n=1}^{N} \mathrm{Gam}(a_{\boldsymbol{B}}, 1/\sigma_n)$.
- $y_{dw} \sim \mathrm{Pois}(\sum_{k_{\boldsymbol{Y}}=1}^{K_2} r_{k_{\boldsymbol{Y}}} \theta_{dk_{\boldsymbol{Y}}} \beta_{wk_{\boldsymbol{Y}}} + \epsilon \sum_{k_{\boldsymbol{B}}=1}^{K_1} \rho_{k_{\boldsymbol{B}}} (\sum_n Z_{nd} \phi_{nk_{\boldsymbol{B}}}) \psi_{wk_{\boldsymbol{B}}})$,
- $r_{k_{\boldsymbol{Y}}} \sim \mathrm{Gam}(\gamma_{\boldsymbol{Y}}/K_{\boldsymbol{Y}}, 1/c_{\boldsymbol{Y}})$, $\theta_{k_{\boldsymbol{Y}}} \sim \prod_{d=1}^{D} \mathrm{Gam}(a_{\boldsymbol{Y}}, 1/\varsigma_d)$, $\beta_{k_{\boldsymbol{Y}}} \sim \prod_{w=1}^{V} \mathrm{Gam}(\xi_{\boldsymbol{Y}}, 1/\eta_w)$, $\psi_{k_{\boldsymbol{B}}} \sim \prod_{w=1}^{V} \mathrm{Gam}(\xi_{\boldsymbol{B}}, 1/\zeta_w)$, $\epsilon \sim \mathrm{Gam}(f_0, 1/g_0)$.

GPPF for Joint Network and Topic Modeling (J-GPPF)



- $b_{nm} = I_{\{x_{nm} \geq 1\}}, x_{nm} \sim \text{Pois}(\sum_{k_B=1}^{K_1} \rho_{k_B} \phi_{nk_B} \phi_{mk_B})$, $\rho_{k_B} \sim \text{Gam}(\gamma_B / K_B, 1/c_B)$, $\phi_{k_B} \sim \prod_{n=1}^{N} \text{Gam}(a_B, 1/\sigma_n)$.
- $y_{dw} \sim \text{Pois}(\sum_{k_Y=1}^{K_2} r_{k_Y} \theta_{dk_Y} \beta_{wk_Y} + \epsilon \sum_{k_B=1}^{K_1} \rho_{k_B} (\sum_n Z_{nd} \phi_{nk_B}) \psi_{wk_B})$,
- $r_{k_Y} \sim \text{Gam}(\gamma_Y / K_Y, 1/c_Y)$, $\theta_{k_Y} \sim \prod_{d=1}^{D} \text{Gam}(a_Y, 1/\varsigma_d)$, $\beta_{k_Y} \sim \prod_{w=1}^{V} \text{Gam}(\xi_Y, 1/\eta_w)$, $\psi_{k_B} \sim \prod_{w=1}^{V} \text{Gam}(\xi_B, 1/\zeta_w)$, $\epsilon \sim \text{Gam}(f_0, 1/g_0)$.
- $\gamma_B \sim \text{Gam}(e_B, 1/f_B)$, $\gamma_Y \sim \text{Gam}(e_Y, 1/f_Y)$.

BC^{3}E: Problem Setting

	$\mathbf{w}_1^{(1)}$	$\mathbf{w}_2^{(1)}$	$\cdots$	$\mathbf{w}_{r_1}^{(1)}$
$\mathbf{x}_1$	2	3	$\cdots$	1
$\mathbf{x}_2$	1	3	$\cdots$	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{x}_N$	2	3	$\cdots$	3

Table: From Classifiers

	$\mathbf{w}_1^{(2)}$	$\mathbf{w}_2^{(2)}$	$\cdots$	$\mathbf{w}_{r_2}^{(2)}$
$\mathbf{x}_1$	4	5	$\cdots$	4
$\mathbf{x}_2$	2	4	$\cdots$	4
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{x}_N$	2	4	$\cdots$	2

Table: From Clusterings

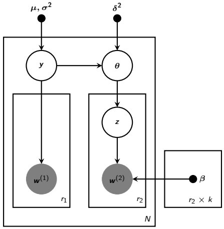
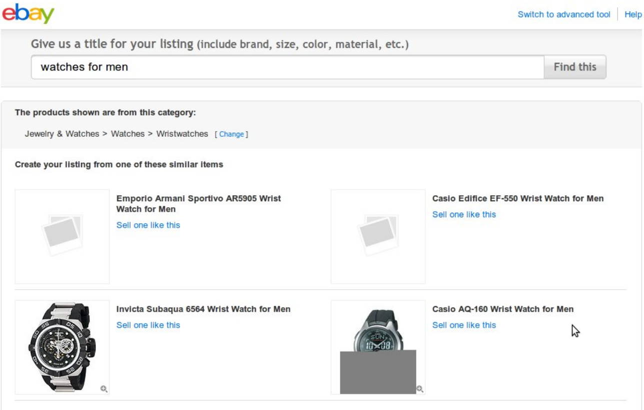


Figure: Graphical Model of **BC^{3}E**

Dataset from eBay Inc.

39 top level nodes called *meta-categories* and 20K+ bottom level nodes called *leaf categories*.



Transfer learning on text data from eBay Inc.

Group ID	$\|\mathcal{X}\|$	k-NN	BGCM	LWE	C^3E-Ideal	BC^3E
42	1299	64.90	73.78 ($\pm$ 0.94)	76.86 ($\pm$ 1.01)	83.99 ($\pm$ 0.41)	83.68 ($\pm$ 1.09)
84	611	63.67	69.23 ($\pm$ 0.17)	75.24 ($\pm$ 0.26)	81.18 ($\pm$ 0.16)	76.27 ($\pm$ 1.31)
86	2381	77.66	84.33 ($\pm$ 2.74)	83.29 ($\pm$ 1.02)	92.78 ($\pm$ 0.35)	87.20 ($\pm$ 0.91)
67	789	72.75	72.75 ($\pm$ 0.07)	78.03 ($\pm$ 0.72)	82.64 ($\pm$ 0.82)	81.75 ($\pm$ 1.37)
52	1076	76.95	77.01 ($\pm$ 1.18)	77.49 ($\pm$ 1.41)	88.38 ($\pm$ 0.22)	85.04 ($\pm$ 2.14)
99	827	84.04	85.12 ($\pm$ 0.52)	86.90 ($\pm$ 0.92)	91.54 ($\pm$ 0.27)	91.17 ($\pm$ 0.82)
48	3445	86.33	86.19 ($\pm$ 0.25)	90.38 ($\pm$ 1.03)	92.71 ($\pm$ 0.31)	92.71 ($\pm$ 1.16)
94	440	79.32	81.08 ($\pm$ 0.73)	82.52 ($\pm$ 0.83)	85.45 ($\pm$ 0.09)	85.45 ($\pm$ 0.79)
35	4907	82.41	82.10 ($\pm$ 0.37)	85.08 ($\pm$ 1.39)	88.16 ($\pm$ 0.17)	88.22 ($\pm$ 1.21)
45	1952	74.80	73.12 ($\pm$ 0.81)	73.64 ($\pm$ 1.68)	84.32 ($\pm$ 0.23)	77.97 ($\pm$ 0.47)

Table: Performance of **BC^{3}E** on text classification data — Avg. Accuracies $\pm$(Standard Deviations).