Unsupervised Learning Of Finite Mixture Models With Deterministic Annealing For Large-scale Data Analysis

Thesis Defense, January 12, 2012

Student: Jong Youl Choi
Advisor: Geoffrey Fox
School of Informatics and Computing
Pervasive Technology Institute
Indiana University
Outline

I. Finite Mixture Model (FMM)
II. Model Fitting with Deterministic Annealing (DA)
III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)
IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)
V. Conclusion And Future Work
I. Finite Mixture Model (FMM)

II. Model Fitting with Deterministic Annealing (DA)

III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)

IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)

V. Conclusion And Future Work
Learning from observed data
Inferring a data generating process
Essential structure, abstract, summary, …
Finite Mixture Model (FMM)

- **Component Model**
  - A mixture of simple distributions (components)
  - Hidden or latent components
  - Serve as abstract or summary of data

- **Generative Model**
  - Simulate observed random sample

- **Convenient and flexible**

- **Model fitting is hard**
  - Too many parameters
Two Finite Mixture Models

- Traditional model
- Component \( \approx \) Cluster
- Clustering, Gaussian Mixture (GM), GTM, …

\[
P(x_i | \Omega, \pi) = \sum_{k=1}^{K} \pi_k P(x_i | \omega_k)
\]
\[
\sum_{k=1}^{K} \pi_k = 1
\]

- Factor model
- Component \( \approx \) Generator
- PLSA

\[
P(x_i | \Omega, \Psi) = \sum_{k=1}^{K} \psi_{ik} P(x_i | \omega_k)
\]
\[
\sum_{k=1}^{K} \psi_{ik} = 1
\]
Model Fitting

- **Bayes’ Theorem**
  \[
P(\theta_i | X) = \frac{P(X | \theta_i)P(\theta_i)}{P(X)}
  \]
  - $\theta_i$: Parameters
  - $X$: Observations
  - $P(\theta_i)$: Prior
  - $P(X | \theta_i)$: Likelihood

- **Maximum Likelihood Estimator (MLE)**
  - Used to find the most plausible $\theta$, given $X$
  - Maximize likelihood or log-likelihood
    $\Rightarrow$ Optimization problem
Expectation-Maximization (EM) Algorithm

- Problems in MLE
  - Observation $X$ is often not complete
  - Latent (hidden) variable $Z$ exists
  - Hard to explore whole parameter space

- EM algorithm
  - Random initialization $\theta^{old}$
  - E-step : Expectation $P(Z \mid X, \theta^{old})$
  - M-step : Maximize (log-)likelihood
  - Repeat E-,M-step until converge.
Motivation

- **Local optimum problem**
  - Easily trap in the local optimum
  - Sensitive to initial conditions or parameters
  - High-variance solution
  - SA, GA, …

- **Overfitting problem**
  - Poor generalization quality
  - Directly related with predicting power
  - Early stopping, cross validation, …
Contributions

- Solve FMM with DA
  - Avoid local optimum problem
  - Avoid overfitting problem
- Present DA applications
  - GTM with DA (DA-GTM)
  - PLSA with DA (DA-PLSA)
- Experimental results
  - Data visualization
  - Text mining
I. Finite Mixture Model (FMM)

II. Model Fitting with Deterministic Annealing (DA)

III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)

IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)

V. Conclusion And Future Work
Deterministic Annealing (DA)

- **Optimization**
  - Gradually lowering numeric temperature
  - No stochastic process

- **Local optimum avoidance**
  - Tracing the global solution by changing level of smoothness
  - Smoothed $\rightarrow$ bumpy

- **Principle of Maximum Entropy**
  - A solution with maximum entropy
  - Minimize the free energy $F$ of log-likelihood
  - Eventually, we will have maximized log-likelihood
Finite Mixture Model with EM

E-step

M-step

Update Log-Likelihood

Converged

Yes

No

No
Finite Mixture Model with DA

- **Set Temp High**
- **Update Temperature**
  - **E-step**
  - **M-step**
  - **Update Free Energy**
  - **Converged**

- **Last?**
  - **No**
  - **Yes**

- **Annealing (High → Low)**
- **High temperature**
  - Soft (or fuzzy) association
  - Smooth cost function
- **Low temperature**
  - Hard association
  - Bumpy cost function
  - Revealing full complexity

- Minimize free energy
- Free energy \( F = f(\text{Temp, Entropy of Log-Likelihood}) \)
Free Energy for Finite Mixture Model

Free Energy

\[ F = D - TS \]
\[ = -T \sum_{n=1}^{N} \ln Z_n \]
- D: expected cost \(<d_{nk}>\)
- S: Shannon entropy
- T: computational temperature
- \(Z_n\): partition function

General form for Finite Mixture Model

\[ F_{FMM} = -T \sum_{n=1}^{N} \log \sum_{k=1}^{K} \left\{ c(n, k) p(x_n|y_k) \right\} \frac{1}{T} \]
- Cost function: \(d_{nk} = - \log p(x_n|y_k)\)
I. Finite Mixture Model (FMM)

II. Model Fitting with Deterministic Annealing (DA)

III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)

IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)

V. Conclusion And Future Work
Dimension Reduction

- Simplification, feature selection/extraction, visualization, etc.
- Preserve the original data’s information as much as possible in lower dimension

PubChem Data (166 dimensions)
Generative Topographic Mapping

▸ An algorithm for dimension reduction
  – Find an optimal K latent variables in a latent space
  – \( f \) is a non-linear mappings
  – Strict Gaussian mixture model
  – EM model fitting
▸ DA optimization can improve the fitting process
Advantages of GTM

- Computational complexity is $O(KN)$, where
  - $N$ is the number of data points
  - $K$ is the number of latent variables or clusters. $K << N$
- Efficient, compared with MDS which is $O(N^2)$
- Produce more separable map (right) than PCA (left)
Free Energy for GTM

\[
F_{FMFM} = -T \sum_{n=1}^{N} \log \sum_{k=1}^{K} \left\{ c(n, k) \ p(x_n | y_k) \right\}^{\frac{1}{T}}
\]

- **GTM Model Setting**
  - **Strict Gaussian assumption**
    \[
p(x_n | y_k) = \mathcal{N}(x_n | \mu_k, \sigma_k)
\]
  - **Constant mixing weight**
    \[
c(n, k) = \frac{1}{K}
\]
# GTM with Deterministic Annealing

<table>
<thead>
<tr>
<th>Optimization</th>
<th>EM-GTM (Traditional method)</th>
<th>DA-GTM (New algorithm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective Function</td>
<td>Maximize log-likelihood $L$</td>
<td>Minimize free energy $F$</td>
</tr>
<tr>
<td></td>
<td>$\sum_{n=1}^{N} \ln \left{ \frac{1}{K} \sum_{k=1}^{K} p(\mathbf{x}_n</td>
<td>\mathbf{y}_k) \right}$</td>
</tr>
</tbody>
</table>

When $T = 1$, $L = -F$.

## Pros & Cons

<table>
<thead>
<tr>
<th>EM-GTM</th>
<th>DA-GTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very sensitive to parameters</td>
<td>Less sensitive to poor parameters</td>
</tr>
<tr>
<td>Trapped in local optima</td>
<td>Avoid local optimum</td>
</tr>
<tr>
<td>Faster</td>
<td>Require more computational time</td>
</tr>
</tbody>
</table>
Cooling Schedules

- Traditional method: static cooling schedule
- Adaptive cooling, a dynamic cooling schedule
  - Able to adjust the problem on the fly
  - Move to a temperature at which $F$ may change
Phase Transition

- Discrete behavior of DA
  - In some temperatures, the free energy is stable.
  - At a specific temperature, start to explode, which is known as critical temperature $T_c$.

- Critical temperature $T_c$
  - Free energy $F$ is drastically changing at $T_c$.
  - Second derivative test: Hessian matrix loose its positive definiteness at $T_c$.
  - $\det(H) = 0$ at $T_c$, where

$$H = \begin{bmatrix}
H_{11} & \cdots & H_{1K} \\
\vdots & & \vdots \\
H_{K1} & \cdots & H_{KK}
\end{bmatrix}$$

$$H_{kk} = \frac{\partial^2 F}{\partial y_k \partial y_k^T} \quad H_{kk'} = \frac{\partial^2 F}{\partial y_k \partial y_{k'}^T}$$
DA-GTM with Adaptive Cooling

Progress of log-likelihood

Adaptive changes in cooling schedule

Oil flow data (1000 points with 12 Dimensions)
DA-GTM Result

Start Temperature

Log Likelihood (llh)

N/A

5.0

7.0

9.0

Type

EM

Adaptive

Exp−A (α = 0.95)

Exp−B (α = 0.99)

Start Temperature (1\st \ T_c = 4.64)

Oil flow data (1000 points with 12 Dimensions)
Conclusion

- GTM with Deterministic Annealing (DA-GTM)
  - Overcome short-comes of traditional EM method
  - Avoid local optimum
  - Robust against poor initial parameters
- Phase-transitions in DA-GTM
  - Use Hessian matrix for detection
  - Eigenvalue computation
- Adaptive cooling schedule
  - New convergence approach
  - Dynamically determine next convergence point
I. Finite Mixture Model (FMM)

II. Model Fitting with Deterministic Annealing (DA)

III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)

IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)

V. Conclusion And Future Work
Corpus Analysis

- **Polysems**
  - A word with multiple meanings
  - E.g., ‘thread’

- **Synonyms**
  - Different words that have similar meaning, a topic
  - E.g., ‘car’ and ‘automotive’
Probabilistic Latent Semantic Analysis (PLSA)

- **Topic model**
  - Assume latent K topics generating words
  - Each document is a mixture of K topics

- **FMM Type-2**
  - The original proposal used EM for model fitting
An Example of DA-PLSA

(AP: 2,246 documents & 10,473 words)

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>percent</td>
<td>stock</td>
<td>soviet</td>
<td>bush</td>
<td>percent</td>
</tr>
<tr>
<td>million</td>
<td>market</td>
<td>gorbachev</td>
<td>dukakis</td>
<td>computer</td>
</tr>
<tr>
<td>year</td>
<td>index</td>
<td>party</td>
<td>percent</td>
<td>aids</td>
</tr>
<tr>
<td>sales</td>
<td>million</td>
<td>i</td>
<td>i</td>
<td>year</td>
</tr>
<tr>
<td>billion</td>
<td>percent</td>
<td>president</td>
<td>jackson</td>
<td>new</td>
</tr>
<tr>
<td>new</td>
<td>stocks</td>
<td>union</td>
<td>campaign</td>
<td>drug</td>
</tr>
<tr>
<td>company</td>
<td>trading</td>
<td>gorbachevs</td>
<td>poll</td>
<td>virus</td>
</tr>
<tr>
<td>last</td>
<td>shares</td>
<td>government</td>
<td>president</td>
<td>futures</td>
</tr>
<tr>
<td>corp</td>
<td>new</td>
<td>new</td>
<td>new</td>
<td>people</td>
</tr>
<tr>
<td>share</td>
<td>exchange</td>
<td>news</td>
<td>israel</td>
<td>two</td>
</tr>
</tbody>
</table>

Top 10 list of the best words of the AP news dataset for 30 topics. Processed by DA-PLSA and shown only 5 topics among 30 topics
Overfitting Problem

- Predictive power
  - Maintain good performance on unseen data
  - A generalized model is preferable
Free Energy for PLSA

\[ F_{FMM} = -T \sum_{n=1}^{N} \log \sum_{k=1}^{K} \{ c(n, k) \, p(x_n \mid y_k) \} \frac{1}{T} \]

- **PLSA Model Setting**
  - Use Multinomial distribution
    \[ p(x_n \mid y_k) = \text{Multi}(x_n \mid \theta_k) \]
  - Flexible mixing weight
    \[ c(n, k) = \psi_{nk} \]
Overfitting Avoidance in DA

- DA can control smoothness
  - Smoothed solution at high temperature
  - Getting specific as annealing
  - Early stopping to get a smoothed (general) model

- Stop condition
  - Use a V-fold cross validation method
  - Measure total perplexity, sum of log-likelihood of both training set and testing set

\[
Total\ Perplexity = a \cdot \mathcal{L}_{PLSA}(X_{\text{training}}, \Theta, \Psi) + b \cdot \mathcal{L}_{PLSA}(X_{\text{testing}}, \Theta, \Psi)
\]

- Tempered-EM, proposed by Hofmann (the original author of PLSA), but annealing is done in a reversed way
Annealing in DA-PLSA

Changes of Log-Likelihood

-700
-800
-900
-1000

Temperature

100 50 10 5 1

Training Set
Mix B
Mix C
Testing Set

Annealing progresses from high temp to low temp

Improved fitting quality with training set during annealing

Over-fitting at Temp=1

Early-stop temperatures depending on schemes:
A (a=0.0, b=1.0)
B (a=0.5, b=0.5)
C (a=0.9, b=0.1)
D (a=1.0, b=0.0)
Predicting Power in DA-PLSA

Log of word probabilities of AP data (100 topics for 10,473 words)

Early stop (Temp = 49.98)

Over-fitting (Temp = 1.0)
AP data with DA-PLSA

(AP: 2,246 documents & 10,473 words)
NIPS data with DA-PLSA

(NIPS: 1,500 doc & 12,419 words)
DA-PLSA with DA-GTM

Corpus (Set of documents) → DA-PLSA → Corpus in K-dimension → DA-GTM → Embedded Corpus in 3D

AP data with 500 topics

- 93 (99)
- 331 (61)
- 406 (19)
- 424 (194)
- 435 (174)
- 445 (146)
- 492 (130)
In the previous picture, we found among 500 topics:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>lately</td>
<td>lately</td>
<td>mandate</td>
<td>mandate</td>
<td>mandate</td>
<td>plunging</td>
</tr>
<tr>
<td>oferrell</td>
<td>oferrell</td>
<td>kuwarts</td>
<td>kuwarts</td>
<td>lately</td>
<td>referred</td>
</tr>
<tr>
<td>mandate</td>
<td>mandate</td>
<td>cardboard</td>
<td>cardboard</td>
<td>ACK</td>
<td>informal</td>
</tr>
<tr>
<td>ACK</td>
<td>cardboard</td>
<td>commuter</td>
<td>commuter</td>
<td>fcc</td>
<td>Anticommu.</td>
</tr>
<tr>
<td>fcc</td>
<td>exam</td>
<td>ACK</td>
<td>lately</td>
<td>ACK</td>
<td>origin</td>
</tr>
<tr>
<td>cardboard</td>
<td>commuter</td>
<td>fcc</td>
<td>Ack</td>
<td>commuter</td>
<td>details</td>
</tr>
<tr>
<td>commuter</td>
<td>exam</td>
<td>lately</td>
<td>exam</td>
<td>oferrell</td>
<td>relieve</td>
</tr>
<tr>
<td>exam</td>
<td>commuter</td>
<td>fabrics</td>
<td>fabrics</td>
<td>fabrics</td>
<td>psychologist</td>
</tr>
<tr>
<td>kuwaits</td>
<td>fabrics</td>
<td>oferrell</td>
<td>fabrics</td>
<td>fabrics</td>
<td>lately</td>
</tr>
<tr>
<td>fabrics</td>
<td>coroon</td>
<td></td>
<td></td>
<td></td>
<td>thatcher</td>
</tr>
</tbody>
</table>

ACK : acknowledges
Anticommu. : anticommunist
# EM vs. DA-\{GTM, PLSA\}

<table>
<thead>
<tr>
<th>Optimization</th>
<th>EM</th>
<th>DA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective Functions</strong></td>
<td>Maximize log-likelihood $L$</td>
<td>Minimize free energy $F$</td>
</tr>
</tbody>
</table>
| GTM | \[
\sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{k=1}^{K} p(x_n | y_k) \right\}
\] | \[-T \sum_{n=1}^{N} \ln \left\{ \left( \frac{1}{K} \right)^{\frac{1}{T}} \sum_{k=1}^{K} p(x_n | y_k)^{\frac{1}{T}} \right\} \] |
| PLSA | \[
\sum_{n=1}^{N} \sum_{k=1}^{K} \psi_{nk} \text{Multi}(x_n | y_k)
\] | \[-T \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \psi_{nk} \text{Multi}(x_n | y_k)^{\frac{1}{T}} \] |

Note: When $T = 1$, $L = -F$. This implies EM can be treated as a special case in DA

<table>
<thead>
<tr>
<th>Pros &amp; Cons</th>
<th>EM</th>
<th>DA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Very sensitive</strong></td>
<td><strong>Less sensitive to an initial condition</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Trapped in local optima</strong></td>
<td><strong>Find global optimum</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Faster</strong></td>
<td><strong>Require more computational time</strong></td>
<td></td>
</tr>
</tbody>
</table>

Jong Youl Choi (Jan 12, 2012)
I. Finite Mixture Model (FMM)
II. Model Fitting with Deterministic Annealing (DA)
III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)
IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)
V. Conclusion And Future Work
Conclusion

- Finite Mixture Model (FMM) problems
  - FMM-1 and FMM-2
  - Maximize log-likelihood for model fitting (MLE)
  - Traditional solutions use EM

- Solve FMMs with DA
  - Avoid local optimum problem
  - Find generalized (smoothed) solution

- Enhance and develop two data mining algorithms
  - DA-GTM
  - DA-PLSA
Future Work

- Determine number of components
  - Help to choose the right number of clusters, topics, the number of lower dimension, …
  - Bayesian model selection, minimum description length (MDL), Bayesian information criteria (BIC), …
  - Need to develop in a DA framework

- Quality study for DA-PLSA
  - Comparison with LDA
  - Precision and recall measurements

- Performance study for data-intensive analysis
  - MPI, MapReduce, PGAS, …
Related Publications


Thank you!!

Question?

Email me at jychoi@cs.indiana.edu