Parallel and Distributed Systems Term Project Proposal

Parallel iterative weighted sampling based Expectation Maximization for Gaussian Mixture Models

Ran Pang

The purpose of this project is to parallel the first stage algorithm in the Scalable Weighted Iterative Sampling for Flow Cytometry Clustering (SWIFT) algorithm. SWIFT is developed to cluster data generated by Flow Cytometry (FC) technique for rapid multivariate analysis and functional discrimination of cells.[1] As a Expectation Maximization (EM) based algorithm, SWIFT is more efficient in dealing with huge data and finding small clusters, which is important in FC data analysis.

SWIFT consists of two stages.[1] At the first stage, the parameters of a Gaussian Mixture Model (GMM) are determined based on the FC data. In the second stage, overlapping Gaussian clusters are merged. This project is to parallel the first stage algorithm, i.e. iterative weighted sampling based EM for GMMs, which is described in the left box below. The EM algorithm for GMMs, as applied in step 2 and step 5.4, is described in the right box below.[2] Possible steps to parallel is in red.

This project is going to build the parallel program in C/C++ running in a single machine to achieve a balance between portability and performance. Since the FC data set is normally not too huge (20 to 30 of million rows and around 20 columns), a distributed system solution (such as Spark) is put as a backup plan for the single machine program. An OpenMP program should be a good start.

Input: $\mathbf{X}$, k, n, p
$\mathbf{X}$: sequence of n data vectors $\{\mathbf{X}^{(i)}\}_{i=1}^{n}$
k: number of Gaussian mixture components
n: sample size
p: number of components to fix at a time
Output: θ: parameters of Gaussian Mixture Model

1. Obtain set $\mathbf{S}$ of n random samples drawn from $\mathbf{X}$.
2. Estimate parameters θ_S using EM on $\mathbf{S}$.
3. Estimate posterior probabilities $\gamma^{(i)}_j$ via an E-step on $\mathbf{X}$ using parameters θ_S.
4. Let $\mathbf{F}$ be the set of Gaussian components whose parameters have been fixed. Initialize $\mathbf{F} \leftarrow \varnothing$.
5. **repeat**
   - 5.1. Determine $\mathbf{F_1}$ = {The p most populous Gaussian components $\notin \mathbf{F}$} for the current model θ_S.
   - 5.2. Fix the parameters of components $\in \mathbf{F_1}$. Set $\mathbf{F} \leftarrow \mathbf{F} \cup \mathbf{F_1}$.
   - 5.3. Resample a set of n points $\mathbf{S}$ from $\mathbf{X}$ with a weighted distribution where each point is selected with probability $(1 - \sum_{l \in \mathbf{F}} \gamma^{(i)}_l)$.
   - 5.4. Apply EM on $\mathbf{S}$ where the parameters of components $\in \mathbf{F}$ are not updated.
   - 5.5. Normalize the mixing probabilities π_j where $j \notin \mathbf{F}$, computed in the M step to $(1 - \sum_{l \in \mathbf{F}} \gamma^{(i)}_l)$.
   - 5.6. Perform a single E-step on $\mathbf{X}$ to recalculate the posteriors $\gamma^{(i)}_j$.
6. **until** all the components are fixed.
7. $\theta \leftarrow$ parameters of all the components $\in \mathbf{F}$.

Input: $\mathbf{X}$, k, n, p
$\mathbf{X}$: sequence of n data vectors $\{\mathbf{X}^{(i)}\}_{i=1}^{n}$
k: number of Gaussian mixture components
n: sample size
Output: μ_k, Σ_k, and mixing coefficients π_k: parameters of Gaussian Mixture Model

1. Initialize the means μ_k, Σ_k, and mixing coefficients π_k, and evaluate the initial value of the log likelihood $\ln p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \ln\{\sum_{k=1}^{n} \pi_k N(x_i \mid \mu_k, \Sigma_k)\}$
2. **repeat**
   - 2.1. E step. Evaluate the responsibilities using the current parameter values
     $\gamma(z_{nk}) = \pi_k N(x_i \mid \mu_k, \Sigma_k) \,/\, \sum_{j=1}^{K} \pi_j N(x_n \mid \mu_j, \Sigma_j)$
   - 2.2. M step. Re-estimate the parameters using the current responsibilities
     $\mu_k^{new} = (1/N_k) \sum_{i=1}^{n} \gamma(z_{ik}) x_i$
     $\Sigma_k^{new} = (1/N_k) \sum_{i=1}^{n} \gamma(z_{ik})(x_i - \mu_k^{new})^T (x_i - \mu_k^{new})$
     $\pi_k^{new} = (N_k/N)$
     where $N_k = \sum_{i=1}^{n} \gamma(z_{ik})$
   - 2.3. Evaluate the initial value of the log likelihood $\ln p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \ln\{\sum_{k=1}^{n} \pi_k N(x_i \mid \mu_k, \Sigma_k)\}$
3. **until** the log likelihood or the parameters converge

[1]
IFTEKHAR NAIM, SUPRAKASH DATTA, GAURAV SHARMA, JAMES S. CAVENAUGH, TIM R. MOSMANN; "SWIFT: SCALABLE WEIGHTED ITERATIVE SAMPLING FOR FLOW CYTOMETRY CLUSTERING"; *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING (ICASSP)*; MARCH 2010, PP. 509–512.

[2]
BISHOP, CHRISTOPHER M. *PATTERN RECOGNITION AND MACHINE LEARNING*. SECTION 9.2 VOL. 4. NO. 4. NEW YORK: SPRINGER, 2006.