Learning Global Models Based on Distributed Data Abstractions

Xiaofeng Zhang and William K. Cheung

Department of Computer Science Hong Kong Baptist University Kowloon Tong, Hong Kong {xfzhang,william}@comp.hkbu.edu.hk

Abstract

Due to the increasing demand of massive and distributed data analysis, achieving highly accurate global data analysis results with local data privacy preserved becomes an increasingly important research issue. In this paper, we propose to adopt a model-based method (Gaussian mixture model) for local data abstraction and aggregate the local model parameters for learning global models. To support global model learning based on solely local GMM parameters instead of virtual data generated from the aggregated local model, a novel EM-like algorithm is derived. Experiments have been performed using synthetic datasets and the proposed method was demonstrated to be able to achieve the global model accuracy comparable to that of using the data regeneration approach at a much lower computational cost.

1 Introduction

Most machine learning and data mining algorithms work with a rather basic assumption that all the data can first be pooled together in a centralized data repository. Recently, there exist a growing number of cases that the data have to be physically distributed due to some practical constraints, such as privacy concerns. The need to tackle them calls for the recent research interest in distributed data mining.

A common methodology for distributed data mining is to first perform local data analysis and then combine the results to form a global one [Kargupta et al., 2000]. One major limitation is that local analysis may result in loss of information for global pattern discovery. In [Zhang et al., 2004], a model parameter exchange approach was proposed to improve the global model accuracy, which however is not a generic methodology which can easily be generalized to different types of global models. An alternative is to adopt a flexible enough model for local data abstraction and have the subsequent analysis based on the local model parameters, which is what we adopt in this paper.

As inspired by [Merugu and Ghosh, 2003], Gaussian mixture model (GMM) [Bishop, 1995] is chosen for the local data abstraction. Via different parameter settings, GMM can interpolate a set of data with more detailed information as well

as a rough ensemble illustration. Thus, by sharing only local model parameters for global analysis, the degree of data privacy preservation and the bandwidth requirement in a distributed environment can readily be controlled. In this paper, an EM-like algorithm is proposed for learning a global GMM directly from the aggregated version of the local abstractions.

2 Problem Formulation

2.1 Local Data Abstraction

Let $t_i \in \Re^d$ denote an observed data instance, θ_l denote the parameters of the l^{th} local model (GMM) as the abstraction of the l^{th} source. The probability density function of the l^{th} local model $p_{local}(t_i|\theta_l)$ with K_l components is given as,

$$\begin{aligned} p_{local}(t_i|\theta_l) &= \sum_{j=1}^{K_l} \alpha_{jl} p_j(t_i|\theta_{lj}) & \sum_{j=1}^{K_l} \alpha_{jl} = 1 \\ p_j(t_i|\theta_{lj}) &= (2\pi)^{-\frac{d}{2}} |\Sigma_{lj}|^{-\frac{1}{2}} \exp\{-\frac{1}{2} (t_i - \mu_{lj})^T \Sigma_{lj}^{-1} (t_i - \mu_{lj})\}. \end{aligned}$$

L sets of local GMMs' parameters $\{\theta_1, \theta_2, ..., \theta_L\}$ obtained by performing local model-based density estimation (e.g., the EM algorithm) are sent to a global server to learn a global data model.

2.2 Learning Global Model Parameters

We start the derivation of our learning algorithm by considering first the conventional setting of the Expectation Maximization (EM) algorithm. Let R_{ik} denote the estimated indicator for the k^{th} component of the global model generating the i^{th} data instance.

Assume R_{lk} is now a new indicator for the l^{th} local component with its underlying data to be generated by the k^{th} global component and \overline{R}_{ik} denotes the average estimate of the probability that data instances from the l^{th} local component are generated by the k^{th} global component. By approximating R_{ik} as

$$R_{ik} \approx \overline{R}_{ik} = \frac{\sum_{i \in l^{th} source} R_{ik}}{N_l} \tag{1}$$

where N_l denotes the number of data from the l^{th} source, the

¹Note that we abuse the index "l" to refer to one local component of the aggregated local model.

new M-steps become

$$\mu_{k} = \frac{\sum_{l=1}^{L} R_{lk} \mu_{l}}{\sum_{l=1}^{L} R_{lk}} \quad \alpha_{k} = \frac{1}{L} \sum_{l=1}^{L} R_{lk}$$

$$\Sigma_{k} = \frac{\sum_{l=1}^{L} R_{lk} (\Sigma_{l} + \mu_{l} \mu_{l}^{T})}{\sum_{l=1}^{L} R_{lk}} - \mu_{k} \mu_{k}^{T}.$$
(2)

KL-divergence is adopted to compute R_{lk} (E-step), resulting in

$$R_{lk} = \frac{\exp\{-D(p_{global}(t|\phi_k)||p_{local}(t|\theta_l))\}}{\sum_{i=1}^{M} \exp\{-D(p_{global}(t|\phi_i)||p_{local}(t|\theta_l))\}}$$
(3)

where D(P||Q) denotes the KL-divergence between the two probabilistic models P and Q, which can be derived as

$$\ln \frac{|\Sigma_l|^{\frac{1}{2}}}{|\Sigma_k|^{\frac{1}{2}}} + \frac{1}{2} (trace(\Sigma_l^{-1}\Sigma_k) + (\mu_k - \mu_l)^T \Sigma_l^{-1} (\mu_l - \mu_k) - 1).$$

3 Experiments

Experiments have been performed using synthetic data sets for performance evaluation. Four data sets (as ground truth) were generated using known mixture models with 4, 5, 6 and 7 components, respectively. The data instances were then randomly partitioned as four sources and local GMMs were learned via the conventional EM algorithm for each source. Two global models were trained, one using the proposed method and one using the conventional data regeneration approach. As shown in the Table 1, the proposed method can achieve a speedup factor ranging from 1.5 to 257 with highly comparable model accuracy.

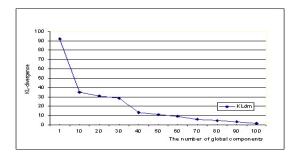
Table 1: Comparing the global models learned using the proposed method and the data regeneration method (gx denotes the true data model with x components).

g4 2: 0.09s/1.5 -1470.2/+2 3.7/-0.03 0.4	3
3: 0.22s/1.8 -1169.4/-70.9 1.0/-0.03 0.1	2
g5 2: 1.33s/33 -1850.6/-17 3.2/-0.02 0.2	1
3: 8.11s/162 -1898.5/-67 4.1/+0.04 1.3	1
4: 10.3s/257 -1788.4/+1 3.3/+0.02 3.6	6

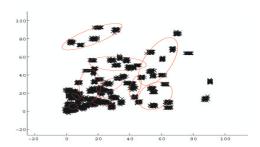
The proposed approach has also been applied to a more sizeable dataset generated using a GMM with 100 components. Global models with different number of components were learned and their corresponding KL-divergence values when compared with the true model are shown in Figure 1 (a). The divergence value decreased rapidly when the number of components increased from 1 to 10 and then started to saturate. One could determine the optimal number of global components by setting a threshold on the KL-divergence value as cut-off. Figure 1 (b) shows the result with 10 global components identified.

4 Conclusion and Future Plan

In this paper, we proposed a novel EM-like algorithm for learning a global model directly from the aggregated local model parameters with promising experimental results on synthetic datasets. Currently, we are extending the obtained results to learn more sophisticated global models. Furthermore, we are investigating how active negotiation between



(a) The KL-divergence of the model learned using the proposed method with different numbers of components of the global model.



(b) GMM with 10 components learned based on 100 local Gaussian components for the aggregated local model.

Figure 1: Experimental results with a sizeable dataset.

the global mediator and the local data sources for compromising individual local data sources' privacy levels can be adopted to improve the global model accuracy.

Acknowledgement

This work has been partially supported by RGC Central Allocation Group Research Grant (HKBU 2/03/C).

References

[Bishop, 1995] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[Kargupta et al., 2000] H. Kargupta, B. Park, D. Hershberger, and E. Johnson. Collective Data Mining: A New Perspective Towards Distributed Data Mining. In Advances in Distributed and Parallel Knowledge Discovery. MIT/AAAI Press, 2000.

[Merugu and Ghosh, 2003] S. Merugu and J. Ghosh. Privacy-preserving Distributed Clustering using Generative Models. In *Proceedings of IEEE International Conference on Data Mining*, 2003.

[Zhang et al., 2004] Xiaofeng Zhang, C.M. Lam, and William K. Cheung. Mining local data sources for learning global cluster models via local model exchange. *The IEEE Intelligent Informatics Bulletin*, 4(2), 2004.