Research on Variational Bayesian-based Data Classification Algorithm

Hong Mei Zhang, Wen Qian Zhang and Peng Xie

Abstract. With the rapid development of Internet technology, the size and complexity of the database are continually growing, the traditional classification method can no longer meet the demand of the classification of complex data. As for such problems, a data classification algorithm based on variational bayesian is proposed. The algorithm introduce variational approximation theory on the basis of traditional bayesian inference, combined with the thought of maximum expected algorithm, utilizing the mean field theory in the statistical physics, and take gaussian mixture model as an experiment simulation. The experimental results show that randomly generated data can be seen clearly mixtured by three groups of gaussian model after 382 iterations, lower bound of likelihood function rises with the increase of iteration number, curve becomes flat as expected after 350 iterations, and getting mean value and the inverse of covariance matrix close to the real data in the range of allowable error, realizing its classification processing. Under the requirement of high precision, its calculation speed is faster, calculation efficiency is higher, which can in accordance with actual engineering application background.

Keywords: Variational Bayesian, Classification Algorithm, EM Algorithm.

Corresponding author: Wen-qian Zhang, College of Equipment Management and Safety Engineering, Air force Engineering University, Xi'an Shaanxi 710051, China, wenqian_z@163.com
Hong-Mei Zhang, Peng Xie, College of Science, Air force Engineering University, Xi'an Shaanxi 710051, China; zhm_plum@163.com, xpf68@163.com
1 Preface

With rapid development of internet technology, it becomes increasingly easy to collect data which leads to larger scales, more complexity and higher dimension of database. It is key problem in date mining of how to extract connotative and unknown useful data from large database\(^1\).

It is an important aspect in classification studying and research, aimed in studying classifier according to labeled training samples, which relative algorithm achievement is extremely rich. The typical classification algorithms are decision tree decision algorithm, Bayesian classification algorithm, traditional classification algorithm based on association rule, supporting vector machine classification algorithm. Of them decision tree decision algorithm can well handle noise data, but just for small-scale training sample sets; Bayesian classification algorithm performs with higher precision, fast and low error rate, but less accurate of classification; Traditional classification algorithm based on association rule performs well but restricted in hardware memory; Supporting vector machine classification algorithm has high accuracy and less complexity, but slow speed, so each algorithm has its advantages and defects\(^2\). In view of the existing problems in above algorithms performing complex data, we proposed variational bayesian-based data classification algorithm that realizing complex data classification performing fast and accurately.

2 Variational Bayesian Theory

Variational Bayesian is one kind method of Bayesian estimation, which introduce variational approximation theory on the basis of traditional bayesian inference and EM algorithm\(^3\). This method widely applied in approximation estimation of statistical model with hidden variables, including in parameter estimation of stochastic system, the blind source separation, and speech enhance field\(^4\).

In statistical model, known about a set of observation data \(D\), then determined posterior distribution \(P(Z|D)\) of which parameters and latent variables \(Z = \{Z_1, \ldots, Z_n\}\). Considering about high dimensional complexity of Posterior probability \(P(Z|D)\), we use one simple model \(Q(Z)\) to replace it, that is \(P(Z|D) \approx Q(Z)\). For the distance between \(Q(Z)\) and \(P(Z|D)\), we use relative entropy\(^5\), that is KL divergence to describe:
\[
\begin{align*}
\log P(D) &= D_{KL}(Q \| P) - \sum_{Z} Q(Z) \log \frac{Q(Z)}{P(Z,D)} = D_{KL}(Q \| P) + L(Q) \\
\text{(1)}
\end{align*}
\]

To minimize the distance between the two distributions, and meanwhile keep Edge logarithmic likelihood function unchanged, the lower bound \( L(Q) \) must reach the maximum.

Except that, considering algorithm complexity, we need to reduce \( L(Q) \) algorithm complexity through the value of analysis. For description easier, we collectively call variational free energy\(^6\) both of Posterior \( P(Z | D) \) approximation and calculated lower bound \( L(Q) \):

\[
L(Q) = \sum_{Z} Q(Z) \log P(Z,D) - \sum_{Z} Q(Z) \log Q(Z) = \mathbb{E}_{q} [\log P(Z,D)] + H(Q) \\
\text{(2)}
\]

For convenient expression, we define first term on right side as energy, and next term as information entropy.

For solving approximate simplification form of high-dimensional complex integral function, the physical branch of statistics has given the theory—average field theory. Simply speaking: in a known macro system, the results of partial interaction between individuals affecting macro level will make stable and fixed among them. Then we make posterior conditions independent of the assumptions, that is:

\[
\forall i, p(Z | D) = p(Z_i | D) p(Z_{\sim i} | D) \\
\text{(3)}
\]

We transform expression into a functional problem, and seek extreme value under the constraint condition, the expression is as follows:

\[
P(Z | D) \approx \forall i, \frac{\partial}{\partial Q_i(Z_i)} \left[ -D_{KL}(Q_i(Z_i) \| Q_i'(Z_i)) - \lambda_i \left( \mathbb{E}_{Q_i}(Z_i) dZ_i - 1 \right) \right] = 0 \\
\text{(4)}
\]

Obviously, \( L(D) \) reached the maximum value in the case of KL divergence is 0, the expression is as follows:

\[
Q_i(Z_i) = Q_i'(Z_i) = \frac{1}{C} \exp \left\{ \ln P(Z_i, Z_{\sim i}, D) \right\}_{Q_i(Z_i)} \\
C = \mathbb{E}_{Q_i(Z_i)} \left[ \exp \left\{ \ln P(Z_i, Z_{\sim i}, D) \right\} dZ_i \right], \quad Q(Z_i) \text{ is the logarithmic desired joint probability density function}^{[7]} \\
\text{(5)}
\]
3 Application Examples Stimulation

Take the mixed Gaussian model\cite{8} in high dimensional data as an example, assuming that the data $D = \{x_n\}_{n=1}^N$ of observed database get from independent samples of the above mixed distribution, the likelihood function is:

$$P(D|\pi, \mu, T) = \prod_{n=1}^N \left| \sum_{i=1}^M \pi_i N(x_n | \mu_i, T) \right|$$

(6)

Mixed coefficient $\pi$, $0 < \pi_i < 1$ and $\sum_{i=1}^M \pi_i = 1$, Mean Value $\mu$, Inverse covariance matrix $T$

Hidden variable $s = \{s_{in}\}_{i=1, n=1}^{M,N}$ is introduced to describe the affiliation between each data and each Gaussian component, for each data $x_n$, $s_n$:

If $x_n$ is the $j$th component of Gaussian model, that is $s_n = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

(7)

The dependency between the observed variables and the non-observed variables can be shown in Figure 1:

[Diagram showing dependency between observable variable and unobservable variable]

FIGURE 1. Dependency between observable variable and unobservable variable.

We can get the joint distribution and the prior distribution of all random variables, and for operation of maximize the edge, we consider the maximization of its lower bound. Denote $\theta = \{\mu, T, s\}$, and then use the inequality Jensen, we can get:

$$\ln P(D|\pi) = \ln \left[ Q(\theta) \frac{P(D, \theta|\pi)}{Q(\theta)} \right] d\theta \geq \int Q(\theta) \ln \frac{P(D, \theta|\pi)}{Q(\theta)} d\theta = L(Q)$$

(8)

For $Q_j = Q_j(\theta_j)$, we broke $L(Q)$ into two part of inclusive $Q_j$ and inclusive $Q_j$ and optimize about $Q_j$

$$L(Q) = \left\{ Q_j \left[ \ln P(D|\pi) \prod_{i \neq j} Q_i d\theta_i \right] d\theta_j - \int Q_j \ln Q_j d\theta_j + \text{const} \right\}$$

(9)
\[
\ln \hat{P}(D, \theta | \pi) = \langle \ln P(D, \theta | \pi) \rangle_{\pi} + \text{const}
\]
(10)
therefore \( L(Q) \) take the maximum when \( Q_j = \hat{P}(D, \theta | \pi) \), that is
\[
\ln Q_j'(\theta_j) = \ln \hat{P}(D, \theta_j | \pi) = \langle \ln P(D, \theta | \pi) \rangle_{\pi} + \text{const}
\]
(11)
So the optimal variable posterior should meet the following expression:
\[
Q_j'(\theta_j) = \frac{\exp \langle \ln P(D, \theta | \pi) \rangle_{\pi}}{\int \exp \langle \ln P(D, \theta | \pi) \rangle_{\pi} d\theta_j}
\]
(12)

3.1 The Description of Algorithm

(1) The optimal variational priori of \( \{s, \mu, T\} \) can be calculated by using formulation (11).

(2) \( \pi \) is gotten by using the Lagrangian multiplier method for the variational lower bound, only consider the terms \( \pi_i \) in the lower bound of the variance, by reusing restrictions \( \sum_{i=1}^{M} \pi_i = 1 \), the objective function of the following form can be constructed:
\[
L(\pi, \lambda) = \sum_{i=1}^{M} \sum_{s=1}^{N} \langle s \rangle \ln \pi_i + \lambda(\sum_{i=1}^{M} \pi_i - 1)
\]
(13)
To solve the following two equations:
\[
\begin{align*}
\frac{\partial L}{\partial \pi_j} &= \sum_{s=1}^{N} \langle s \rangle \frac{1}{\pi_j} + \lambda = 0 \\
\sum_{i=1}^{M} \pi_i &= 1
\end{align*}
\]
(14)

(3) \( M \) is automatically determined by subtracting the smaller component of the mixed coefficient.

(4) The calculation of the lower bound of variance
\[
L(Q) = \{\ln P(D|\mu, T, s)\}_{\mu, T, s} + \{\ln P(s)\}_{s} + \{\ln P(\mu)\}_{\mu} + \{\ln P(T)\}_{T} - \{\ln Q_i(s)\}_{s} - \{\ln Q_i(\mu)\}_{\mu} - \{\ln Q_i(T)\}_{T}
\]
(15)

(5) The Gaussian mixture model algorithm diagram is shown in Figure 2.
4 The Results and Analysis of Experiment
4.1 The Results of Experiment

A simulation model of 3-Gaussian mixture model (same covariance).

| TABLE 1. Mean value and covariance matrix under the real and simulated data. |
|------------------|------------------|------------------|------------------|
| Mean value       | covariance matrix |
| Real             | [0, -2] [0, 0] [0, 2] | \[\begin{pmatrix} 2 & 0 \\ 0 & 0.2 \end{pmatrix} \] |
| simulation       | [-0.06, -0.06]   | \[\begin{pmatrix} 2.06 & -0.01 \\ -0.01 & 0.19 \end{pmatrix}, \begin{pmatrix} 2.26 & 0.07 \\ 0.07 & 2.26 \end{pmatrix}, \begin{pmatrix} 1.96 & 0.10 \\ 0.10 & 2.26 \end{pmatrix} \] |

4.2 The Evaluation of Experiment

In this experiment, we first randomly generate data as Figure 4, and then through the simulation can be seen from Figure 4 that the elliptical results can be seen clearly mixed by three groups of gaussian model after 382 iterations, by the center of the mean value and the radius of inverse covariance matrix, and within
the allowable range of error, close to the real data. It is ideal for simulation results. Except that, in figure 5 lower bound of likelihood function rises with the increase of iteration number, curve becomes flat as expected after 350 iterations. The real data and mean value and the inverse of covariance matrix of simulated data is shown in table 1 and table 2: known about a set of observed data D, we determined the unknowns in the Gaussian mixture model including the hidden variables \( s \) by the variational bayesian method, the mean of the Gaussian components, inverse covariance matrices \( \{ \mu, T \} \) the mixing coefficients \( \pi \) and the model order \( M \), so to realize the classification of complex data is feasible and efficient, and it is worth popularizing.

**ACKNOWLEDGMENTS**

This research is supported by the National Natural Science Foundation of China Grant 71601183.

Biography:
- Hongmei Zhang (1970 — ), professor. Research Interest: Computer technology and application.
- Wenqian Zhang (1993 — ), master degree candidate. Research Interest: Airspace control and management.
- Peng Xie (1983 — ), associate professor. Research Interest: Computer technology and application.

**References**

1. Chengfan Li, Jingyuan Yin. A Multispectral Remote Sensing Data Spectral Unmixing Algorithm Based on Variational Bayesian ICA[J]. Journal of the Indian Society of Remote Sensing, 2013, 412.
2. LI Ling-li, A Review on Classification Algorithms in Data Mining[J]. Journal of Chongqing Normal University (Natural Science 2011,04) pp.44-47.
3. Bishop C M. Pattern recognition and machine learning[M]. New York: Springer, 2006.
4. Sileye Ba, Xavier Alameda-Pineda, Alessio Xompero, Radu Horaud. An On-line Variational Bayesian Model for Multi-Person Tracking from Cluttered Scenes[J]. Computer Vision and Image Understanding, 2016.
5. Beal M J. Variational algorithms for approximate Bayesian inference[D]. University of London, 2003.
6. Cheng-Fan Li, Jing-Yuan Yin, Chun-Song Bai. Variational Bayesian independent component analysis for spectral unmixing in remote sensing image[J]. Arabian Journal of Geosciences, 2013, 64.
7. Chen Shen, Dingjie Xu, Wei Huang, Feng Shen. An Interacting Multiple Model Approach for State Estimation with Non-Gaussian Noise Using a Variational Bayesian Method[J]. Asian Journal of Control, 2015, 174.
8. XU Dingjie, SHEN Chen, SHEN Feng. Variational Bayesian Learning for Parameter Estimation of Mixture of Gaussians.[J]. Journal of Shanghai Jiaotong University, 2013, 07:1119-1125.