

子空间聚类的非参数模型及变分贝叶斯学习

卿湘运 王行愚

(华东理工大学信息科学与工程学院 上海 200237)

摘 要 子空间聚类的目标是在不同的特征子集上对给定的一组数据归类. 此非监督学习方法试图发现数据“在不同表达下的相似”模式, 并且引起了相关领域大量的关注和研究. 首先扩展 Hoff 提出的“均值与方差平移”模型为一个新的基于特征子集的非参数聚类模型, 其优点是能应用变分贝叶斯方法学习模型参数. 此模型结合 Dirichlet 过程混合模型和选择特征子集的非参数模型, 能自动选择聚类个数和进行子空间聚类. 然后给出基于马尔可夫链蒙特卡罗的参数后验推断算法. 出于计算速度上的考虑, 提出应用变分贝叶斯方法学习模型参数. 在仿真数据上的实验结果及在人脸聚类问题上的应用均表明了此模型能同时选择相关特征和在这些特征上具有相似模式的数据点. 在 UCI “多特征数据库”上应用无需抽样的变分贝叶斯方法, 其实验结果说明此方法能快速推断模型参数.

关键词 混合模型; Dirichlet 过程; 非参数贝叶斯; 马尔可夫链蒙特卡罗; 变分学习

中图法分类号 TP181

Nonparametric Model and Variational Bayesian Learning for Subspace Clustering

QING Xiang-Yun WANG Xing-Yu

(College of Information Science and Technology, East China University of Science and Technology, Shanghai 200237)

Abstract The goal of subspace clustering is to group a given set of data represented by different feature subsets. As an unsupervised learning method, subspace clustering tries to discover the patterns of "similarity examined under different presentations" and has received a great deal of interest and research in the related domains. Firstly the "mean and variance shift" model proposed by Hoff is extended to a new nonparametric model of subspace clustering based on subsets of features. The advantage of the model is that variational Bayesian method can be applied. The model based on the integration of a Dirichlet process mixture model and a nonparametric model of selecting subsets of features can automatically choose the number of clusters and perform subspace clustering. Then posterior inference of the model is done using Markov Chain Monte Carlo. Due to computational considerations the authors propose a variational Bayesian method to learn the parameters of the model. Experimental results using simulated data and the application to the problem of clustering face images illustrate the model can simultaneously selecting the relevant features and the data points that have similar pattern under these features. Experiments on the "multiple feature database" from the UCI repository show that variational Bayesian method without sampling can fleetly inference the parameters of the model.

Keywords mixture model; Dirichlet process; nonparametric Bayes; Markov chain Monte Carlo; variational learning

1 引言

根据特征子集进行聚类是机器学习领域一项重要的基础工作. 聚类要在没有任何类标签信息的情况下自动将具有类似模式的目标对象归类, 是一项非监督学习问题. 类似图像分割、文本图像归类、网页语义抽取等都可视作聚类问题. 在这些实际应用中, 高维的特征中可能仅少量特征有区分不同类别的作用, 由此引出同时进行聚类与特征选择的研究. 但是许多传统的特征选择算法是面向监督学习问题的, 即知晓目标对象的类别信息, 选取对未知类别信息目标分类准确度最高的若干特征. 这些算法一般分为两类, 即 wrapper 方法, 根据特定的分类器进行特征选择; 另一类为 filter 方法, 根据不依赖特定分类器的某目标函数进行特征选择. 但是在聚类这类非监督学习问题中, 由于没有类标签信息, 因此很难评价特征子集的相关性. 特别地, 由于类别数是未知的, 且类别数与特征子集的显著性也是交织相关的, 使得根据特征子集进行聚类更加具有挑战性.

Law 等将特征显著性引入到基于高斯混合的聚类框架中, 从而将特征显著性作为一个概率问题^[1]. 根据最小消息长度标准(MML), 利用期望值最大化(EM)算法估计特征显著的概率大小, 不相关的特征显著性概率趋向零. 本方法同时也能估计聚类个数.

Constantinopoulos 等根据上面算法中描述特征显著性的概率模型, 将模型选择与特征选择相结合, 扩充为一个贝叶斯框架^[2]. 应用变分推断方法能同时得到聚类个数、特征显著性程度与各混合分量参数. 试验结果表明此方法在高维稀疏数据集上较前算法更加鲁棒.

Roth 等在聚类任务中应用 wrapper 策略即直接根据某些划分算法的判别能力进行特征选择, 并提出了一个优化算法保证局部最优收敛^[3]. 在几个两类问题上的试验结果显示此方法有类似监督学习算法的分类精度.

以上 3 个算法都假定所有类别的特征显著程度是一致的, 即对每个聚类都选择相同的特征. 然而, 每个聚类可能作用于不同的特征子集, 这些子集可能重叠, 不一定相等, 这种聚类方法称之为子空间聚类. 1998 年 Agrawal 等给出了首个子空间聚类算法 CLIQUE^[4], 此后陆续有不少算法问世, 这些算法综述可参考文献[5-6]等.

Friedman 等根据目标加权特征的差异, 迭代产生数据对之间的不相似距离, 再应用基于距离的聚类方法进行子空间聚类^[7]. 这方法(COSA)本质上是一个启发式的聚类算法, 不能决定聚类个数, 而且一个潜在的假设是每个聚类的特征子集不重叠.

Hoff 等提出了基于“均值与方差平移”模型的子空间聚类算法^[8](以下简称 Hoff 模型), 并根据多变量 Dirichlet 过程混合模型选取聚类个数, 根据一个或更多特征属性均值与方差的不同对目标自动归类. 对比 COSA 算法, 此算法能发现方差没有变化、仅均值平移的一类目标, 而 COSA 对此类目标不敏感.

本文的工作建立在 Hoff 模型的基础上, 主要贡献如下:

(1) 对选择均值或方差平移的特征子集类似进行聚类的 Dirichlet 过程引入非参数模型. 而 Hoff 的方法则是应用先验优势对数(prior log-odd)法选取平移的特征子集. 因此本方法对模型选择即决定聚类分量个数和特征子集的选择建立了一个统一的非参数贝叶斯模型. 传统的参数模型对于数据的生成、分量个数及特征的选择作了较强的先验假设, 而非参数模型的假设条件则较弱, 让数据“自己说话”. 许多非参数模型能看作有限参数模型趋向无穷时的模型, 特别地 Dirichlet 过程是一类重要的非参数模型.

(2) 在利用马尔可夫链蒙特卡罗(Markov Chain Monte Carlo, MCMC)推断模型参数的基础上, 给出了利用变分贝叶斯(Variational Bayes, VB)来推断模型参数的算法. 利用 MCMC 方法特别是 Gibbs 抽样方法, 虽然只要有足够的抽样, 可能得到模型参数的无偏估计, 保证能收敛到逼近的后验分布, 但是 MCMC 收敛速度慢, 而且判断一条马尔可夫链何时收敛也是一个难以解决的问题. 而变分贝叶斯学习方法则利用易处理的一簇分布来逼近隐变量的后验分布, 最大化变分参数的对数似然目标函数的下界来求得模型参数, 从而加快算法参数推断速度.

2 子空间聚类的非参数模型

假定要进行聚类的数据集 $X = \{x_i | i = 1, 2, \dots, N\}$, 每个数据 x_i 是一个 D 维的特征矢量 $x_i = \{x_{ij} | j = 1, 2, \dots, D\}$. 假定这些数据是由一个混合模型生成的, 共 K 个分量, 其混合权值为 π_k . 再进一步假定

此模型每个混合分量密度可因子分解的,即为各个特征分量概率密度的乘积. 根据 Hoeff 等的“均值与方差平移”模型, x_i 的概率密度也如文献[2]所示,是一个两层混合模型

$$p(x_i) = \sum_{k=1}^K \pi_k \prod_{j=1}^D \varphi(x_{ij}),$$

$$\varphi(x_{ij}) = \omega_{kj} N(x_{ij} | \mu_j + \delta_{kj} \times \omega_{kj}, \omega_{kj}^2 \times \sigma_j^2) + (1 - \omega_{kj}) N(x_{ij} | \mu_j, \sigma_j^2).$$

为便于选择分量个数和每个分量的特征子集,引入隐变量 $z_{ik} \in \{0, 1\}$, 且 $\sum_{k=1}^K z_{ik} = 1$ 和 $r_{kj} \in \{0, 1\}$, 数据 X 假定独立地从高斯分布抽取得到

$$p(X | z, r, \mu, \sigma^2, \delta, \omega)$$

$$= \prod_{i=1}^N \prod_{k=1}^K \left[\prod_{j=1}^D N(x_{ij} | \mu_j + \delta_{kj} \times \omega_{kj}, \omega_{kj}^2 \times \sigma_j^2)^{r_{kj}} \times N(x_{ij} | \mu_j, \sigma_j^2)^{1-r_{kj}} \right]^{z_{ik}} \quad (1)$$

r_{kj} 与 Law 等的特征显著性概率类似,但是由于此算法是基于子空间的聚类,因此 r_{kj} 表示某类相对一基准类别特征是否需要平移 δ_{kj} , 或方差是否伸缩 ω_{kj}^2 , 反映的是类别之间一个相对的特征显著概念,可以认为是对每个聚类进行特征选择,而不是对所有类别进行相同的特征选择.

根据式(1)在最大似然框架下可以得到模型各参数值. 最大似然法存在过适应问题,另一个更主要的问题是不能进行模型选择. 由于我们试图决定聚类个数,本文应用贝叶斯方法进行模型选择,因此需对层次模型参数赋予合适的共轭先验分布假设,一方面克服最大似然法存在的过适应问题及模型选择问题,另一方面为了计算方便,特别对于变分贝叶斯推断方法采用共轭先验分布能求得参数后验期望值的闭形式解. 以下给出模型各参数的先验分布

$$\mu_j \sim N(m_j, c) \quad (2)$$

$$\sigma_j^2 \sim IG(v_1, v_{2j}) \quad (3)$$

其中 $IG(v_1, v_2)$ 表示形状参数为 v_1 、尺度参数为 v_2 的倒伽母分布.

聚类过程,即决定每个数据的类别 z_i 的过程则描述为 Dirichlet 过程. 由于混合权值 π_k 必须为正,且和为 1,故取为 Dirichlet 先验分布

$$p(\pi_1, \pi_2, \dots, \pi_K | \alpha) \sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{j=1}^K \pi_j^{\alpha/K-1} \quad (4)$$

其中 $\Gamma(\cdot)$ 表示伽母函数. 先验参数 α 取形状参数为 a_α 、尺度参数为 b_α 的伽母分布

$$\alpha \sim G(a_\alpha, b_\alpha) \quad (5)$$

在给定混合权值 π_k 情况下,得到的每个目标 x_i 归属某个类别 z_i 的联合分布为多项式分布

$$p(z_1, z_2, \dots, z_N | \pi_1, \pi_2, \dots, \pi_K) = \prod_{k=1}^K \pi_k^{n_k} \quad (6)$$

$$n_k = \sum_{i=1}^N I(z_i = k)$$

其中 $I(\cdot)$ 为指示函数,如果括号里的条件满足则为 1,否则为 0. 利用标准 Dirichlet 积分,则能得到在先验参数 α 下目标对象类别 z_i 的概率分布

$$p(z_1, z_2, \dots, z_N | \alpha) = \int p(z_1, \dots, z_N | \pi_1, \dots, \pi_K) p(\pi_1, \dots, \pi_K | \alpha) d\pi_1 \dots d\pi_K$$

$$= \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_{k=1}^K \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)} \quad (7)$$

如果考虑聚类个数 k 未知,假定为无限类,即 $K \rightarrow \infty$,此模型成为无限混合模型. 但是实际上由于观测的数据个数 N 有限,因此至少包含一个目标对象的类别数 K_+ 为有限. 因此上式在 $K \rightarrow \infty$ 的情况下为^[9-10]

$$p(z_1, z_2, \dots, z_n | \alpha) = \alpha^{K_+} \left\{ \prod_{k=1}^{K_+} (n_k - 1)! \right\} \Gamma(\alpha) / \Gamma(\alpha + N) \quad (8)$$

各个分量参数分布的先验假设为

$$\delta_{kj} \sim N(0, \tau_j^2 = \eta \times \sigma_j^2) \quad (9)$$

$$\omega_{kj}^2 \sim IG(a_w, b_w) \quad (10)$$

$$\eta \sim IG(a_\eta, b_\eta) \quad a_w \geq b_w + 1 \quad (11)$$

对于二值变量 r_{kj} ,我们假定第 k 类第 j 个特征均值平移或方差伸缩的概率为 g_j ,且相互独立. 生成二值变量的概率分布一般用贝努利(Bernoulli)分布表示,故有

$$r_{kj} \sim \text{Bernoulli}(g_j) = g_j^{r_{kj}} (1 - g_j)^{1-r_{kj}} \quad (12)$$

贝努利分布的共轭先验分布为 Beta 分布,因此

$$g_j \sim \text{Beta}(\beta/D, 1) = (\beta/D) \times g_j^{(\beta/D)-1} \quad (13)$$

由于能将贝努利分布的均值 g_j 积分出来

$$\int \prod_{k=1}^T p(r_{kj} | g_j) p(g_j) dg_j = \frac{\beta \Gamma(t_j + \beta/D) \Gamma(T - t_j + 1)}{D \Gamma(T + 1 + \beta/D)},$$

产生 r_{kj} 的先验仅依赖于选择特征 j 的类别个数 t_j , 因此上述产生 r_{kj} 的模型也是一个非参数统计模型. 在 Griffiths 等的工作中类似 Dirichlet 过程推广了当 $j \rightarrow \infty$ 即特征个数为无限时选取特征的分^[9].

再对 Beta 过程先验参数 β 设定合适的先验分布

$$\beta \sim G(a_\beta, b_\beta) \quad (14)$$

其形状超参数 a_β 和尺度超参数 b_β 控制二值矩阵 \mathbf{r} 元素 0 和 1 总个数的比例。

3 后验推断

由以上概率模型和先验分布, 可得到各个变量的后验分布。值得一提的是, 应用 Dirichlet 过程进行 Gibbs 抽样的过程也被称作中餐馆过程 (Chinese Restaurant Process, CRP)。类似地, 当假定有无限特征时各目标对象进行特征选择的过程被形象地称作印度自助餐过程 (Indian Buffet Process, IBP)^[9]。本文的统计模型假定无限分量混合, 但特征个数有限。其抽样算法主要步骤如下。

1. 对每个数据, 抽样 $z_i, i=1, 2, \dots, N$;
2. 在当前的类别 $k = \{1, 2, \dots, T\}$, 抽样 $r_{kj}, j=1, 2, \dots, D$;
3. 抽样 w_{kj}^2, δ_{kj} ;
4. 抽样 $\mu_j, \sigma_j^2, \eta, \alpha$ 和 β 等。

迭代执行直至收敛。我们将中间隐变量 π_k 和 g_j 积分出来, 因而不要求后验分布予以抽样更新。文献 [8] 已给出了模型部分参数的后验推断结果, 为内容的完整性和易于理解起见, 本文仍将其结果附上。限于篇幅, 这里只直接给出各参数的后验分布。

3.1 抽样 z_i

假定 T 为当前类别数, 由于

$$p(z_i = k | z_{-i}, x, \theta) \propto p(x_i | z_i, \theta) p(z_i | z_{-i}),$$

且有

$$p(z_i = k, k \leq T | z_{-i}) = n_{-i,k} / (N - 1 + \alpha),$$

$$p(z_i = k, k > T | z_{-i}) = \alpha / (N - 1 + \alpha),$$

$$p(x_i | z_i, \theta) = \prod_{j=1}^D N(x_{ij} | \mu_j, \sigma_j^2, w_{z(i),j}^2, \delta_{z(i),j})^{r_{z(i),j}} \cdot N(x_{ij} | \mu_j, \sigma_j^2)^{1-r_{z(i),j}},$$

$n_{-i,k}$ 为不考虑数据 i 时类 k 包含的数据个数, 因此在每次抽样过程中有可能增加或减少一个聚类, 或保持不变, 完全视当前数据似然及很弱的先验分布假设而定。

3.2 抽样 r_{kj}

$$p(x_{ij} : z_i = k | r_{kj} = 1, \theta)$$

$$= \iint \left\{ \prod_{i: z(i)=k} (2\pi)^{-1/2} \times (w_{kj}^2 \sigma_{kj}^2)^{-1/2} \times \exp[-(x_{ij} - \mu_j - \delta_{kj} \times w_{kj})^2 / (2w_{kj}^2 \sigma_{kj}^2)] \right\} \times \\ ((b_w)^{a_w} / \Gamma(a_w)) \times (w_{kj}^2)^{-a_w-1} \exp(-b_w / w_{kj}^2) \times \\ (2\pi)^{-1/2} \times (\eta \sigma_j^2)^{-1/2} \times \exp[-\delta_{kj}^2 / (2\eta \sigma_j^2)] d\delta_{kj} dw_{kj}^2$$

$$= (2\pi)^{-n_k/2} \times (\eta n_k + 1)^{-1/2} \times (\sigma_j^2)^{-n_k/2} \times \\ (\Gamma(a_w + n_k/2) / \Gamma(a_w)) \times (b_w)^{a_w} \times \\ \left\{ b_w + \frac{n_k}{2\sigma_j^2} \left[\overline{\xi_{kj}^2} - \frac{\eta}{\eta + 1/n_k} (\bar{\xi}_{kj})^2 \right] \right\}^{-(a_w + n_k/2)}.$$

在将分量参数积分出来后, 此概率密度实际上具有多变量 t -分布的概率密度形式。

$$p(x_{ij} : z_i = k | r_{kj} = 0, \theta)$$

$$= \prod_{i: z(i)=k} (2\pi)^{-1/2} \times (\sigma_j^2)^{-1/2} \times \exp[-(x_{ij} - \mu_j)^2 / (2\sigma_j^2)] \\ = (2\pi)^{-n_k/2} \times (\sigma_j^2)^{-n_k/2} \times \exp[-n_k \times \overline{\xi_{kj}^2} / (2\sigma_j^2)],$$

其中, $\xi_{kj} = \{x_{ij} - \mu_j : z_i = k\}$, $\bar{\xi}_{kj}$ 为向量 ξ_{kj} 的均值, $\overline{\xi_{kj}^2}$ 为 ξ_{kj} 平方的均值。由于

$$p(r_{kj} = 1 | r_{-k,j}) = \int_0^1 p(r_{kj} | g_j) p(g_j | r_{-k,j}) dg_j = \\ (t_{-k,j} + \beta/D) / (T + \beta/D),$$

$$p(r_{kj} = 0 | r_{-k,j}) = (T - t_{-k,j}) / (T + \beta/D),$$

其中 $t_{-k,j}$ 为不考虑类 k 情况下选择第 j 个特征的类别个数, 特别地考虑到每个特征假定是独立生成的, 则抽样 r_{kj} 的过程为

$$p(r_{kj} = 1 | x, r_{-(kj)}, \theta) \propto$$

$$p(x_{ij} : z_i = k | r_{kj} = 1, \theta) p(r_{kj} = 1 | r_{-(kj)}) =$$

$$p(x_{ij} : z_i = k | r_{kj} = 1, \theta) p(r_{kj} = 1 | r_{-k,j}),$$

$$p(r_{kj} = 0 | x, r_{-(kj)}, \theta) \propto$$

$$p(x_{ij} : z_i = k | r_{kj} = 0, \theta) p(r_{kj} = 0 | r_{-(kj)}) =$$

$$p(x_{ij} : z_i = k | r_{kj} = 0, \theta) p(r_{kj} = 0 | r_{-k,j}).$$

此计算过程与 Hoff 模型在计算上的主要区别是: 抽样 r_{kj} 由观测数据、超参数及选择 j 特征的聚类个数决定, 因此可以较好地控制聚类之间均值平移或方差伸缩的特征个数, 同时使得一些没有鉴别信息的特征的 r_{kj} 设置为 0, 通过此模型可以得到两类之间具有判别信息的特征差异。

3.3 抽样 w_{kj}^2, δ_{kj}

当 $r_{kj}=1$ 时,

$$\Sigma_j = \sigma_j^2 I_{n_k \times n_k} + \tau_j^2 1_{n_k \times n_k},$$

$$w_{kj}^2 \sim IG(a_w + n_k/2, b_w + (\xi_{kj}^T \Sigma_j^{-1} \xi_{kj})/2),$$

$$\hat{\tau}_{kj}^2 = (n_k / \sigma_j^2 + 1 / \tau_j^2)^{-1},$$

$$\hat{\delta}_{kj} = \sum_{i: z(i)=k} (x_{ij} - \mu_j) / (w_{kj} \times (n_k + 1/\eta)),$$

$$\delta_{kj} \sim N(\hat{\delta}_{kj}, \hat{\tau}_{kj}^2),$$

当 $r_{kj}=0$ 时

$$w_{kj}^2 \sim IG(a_w, b_w),$$

$$\delta_{kj} \sim N(0, \tau_{kj}^2).$$

3.4 抽样 $\mu_j, \sigma_j^2, \eta, \alpha$ 和 β 等

$$\hat{c}_j = \{1/c + \sum_{i=1}^N 1/[(w_{z(i),j}^2)^{r_{z(i),j}} \times \sigma_j^2]\}^{-1},$$

$$\begin{aligned}\epsilon_{ij} &= x_{ij} - r_{z(i),j} \times w_{z(i),j} \times \delta_{z(i),j}, \\ \hat{\mu}_j &= \hat{c}_j \times \left\{ m_j / c + \sum_{i=1}^N \epsilon_{ij} / [(w_{z(i),j}^2)^{r_{z(i),j}} \times \sigma_j^2] \right\}, \\ \mu_j &\sim N(\hat{\mu}_j, \hat{c}_j), \\ \sigma_j^2 &\sim IG(v_1 + (N + T)/2, v_{2j} + \sum_{i=1}^N (\epsilon_{ij} - \mu_j)^2 / \\ &\quad (2 \times (w_{z(i),j}^2)^{r_{z(i),j}}) + \sum_{k=1}^T \delta_{kj}^2 / (2 \times \eta)),\end{aligned}$$

$$\eta \sim IG(a_\eta + D \times T/2, b_\eta + (\sum_{k=1}^T \sum_{j=1}^D \delta_{kj}^2 / \sigma_j^2) / 2),$$

$$\begin{aligned}p(\alpha | \cdots) &\propto p(z_1, \cdots, z_n | \alpha) p(\alpha) \propto \\ &\alpha^{T+a_\alpha-1} e^{-b_\alpha \alpha} \Gamma(\alpha) / \Gamma(\alpha + N), \\ p(\beta | \cdots) &\propto p(r_{11}, \cdots, r_{TD} | \beta) p(\beta) \propto \\ &\left(\prod_{j=1}^D \int \left(\prod_{k=1}^T p(r_{kj} | g_j) p(g_j) dg_j \right) p(\beta) \propto \right. \\ &\left. \left(\prod_{j=1}^D \frac{(\beta/D) \Gamma(t_j + (\beta/D)) \Gamma(T - t_j + 1)}{\Gamma(T + 1 + (\beta/D))} \right) \times \right. \\ &\left. \beta^{a_\beta-1} e^{-b_\beta \beta}, \right.\end{aligned}$$

其中 $t_j = \sum_{k=1}^T r_{kj}$.

虽然 α 和 β 的后验分布没有标准的分布形式，但可采用自适应拒绝抽样方法或重要抽样方法予以更新。

4 变分贝叶斯推断方法

设在贝叶斯层次模型中所有超参数集为 ϑ ，所有中间隐变量集为 θ ，变分推断方法就是试图找到一个具有因子分解形式的 $Q(\theta)$ 逼近后验分布 $p(\theta | X, \vartheta)^{[10]}$ 。因此最小化 $Q(\theta)$ 和 $p(\theta | X, \vartheta)$ 的 KL 散度目标函数：

$$\begin{aligned}D(Q(\theta) \parallel p(\theta | X, \vartheta)) &= E_Q[\ln Q] - \\ &E_Q[\ln p(\theta | X, \vartheta)] + \ln p(X | \vartheta),\end{aligned}$$

其中 $E_p\{a\}$ 表示随机变量 a 关于概率分布或密度 p 的期望值。应用 Jensen 不等式，最小化 KL 散度目标函数也即最大化对数边缘似然的下界 $L[Q, \vartheta]$ ：
 $\ln p(X | \vartheta) \geq L[Q, \vartheta] = E_Q[\ln p(\theta | X, \vartheta)] - E_Q[\ln Q]$
可得到后验 $p(\theta | X, \vartheta)$ 的逼近 $Q(\theta)$ 。

由于 Dirichlet 过程是无限混合模型，中间参数 π_k 可能有无穷个，故无法直接应用变分推断方法来求解。因此一般用“断棒”(stick-breaking)表示形式截断无限混合为有限混合^[11]。引入随机变量

$$V_k \sim Beta(1, \alpha),$$

则混合权值 π_k 可表示为

$$\pi_k(V) = V_k \prod_{j=1}^{k-1} (1 - V_j),$$

同时，为减少中间变量的耦合，调整 δ_{kj} 的分布形式为

$$\delta_{kj} \sim N(0, \tau_j^2 = \eta \times w_{kj}^2 \times \sigma_j^2),$$

$$\begin{aligned}p(X | z, r, \mu, \sigma^2, \delta, w) &= \\ &\prod_{i=1}^N \prod_{k=1}^K \left[\prod_{j=1}^D N(x_{ij} | \mu_j + \delta_{kj}, w_{kj}^2 \times \sigma_j^2)^{r_{kj}} \times \right. \\ &\left. N(x_{ij} | \mu_j, \sigma_j^2)^{1-r_{kj}} \right]^{z_{ik}}.\end{aligned}$$

至此，一个统一的非参数贝叶斯层次模型如图 1 所示。

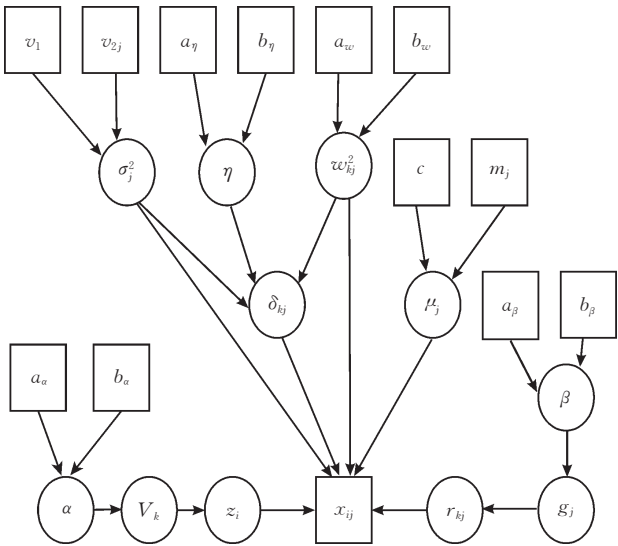


图 1 子空间聚类非参数贝叶斯层次模型

设截断的混合分量个数为 T ，一般取较大值，甚至可为所有数据个数。设各分量中间参数为 θ_k ，中间参数的先验超参数为 ϑ_0 ，则有

$$\begin{aligned}Q(V, z, \theta^* | \gamma, \theta, \phi) &= \\ &\prod_{k=1}^{T-1} Q(V_k | \gamma_k) \prod_{k=1}^T Q(\theta_k^* | \theta_k) \prod_{i=1}^N Q(z_i | \phi_i),\end{aligned}$$

其中 θ_k^* 为各分量参数 θ_k 的后验共轭， $Q(V_k | \gamma_k)$ 为 Beta 分布 $Beta(\gamma_{k,1}, \gamma_{k,2})$ ， $Q(z_i | \phi_i)$ 为 N -维的多项式分布。因此

$$\begin{aligned}L[Q, \vartheta] &= \sum_{k=1}^T E_Q[\ln p(V_k | \alpha^*)] + \\ &\sum_{k=1}^T E_Q[\ln p(\theta_k^* | \vartheta_0)] + \\ &\sum_{i=1}^N E_Q[\ln p(z_i | V)] + \\ &\sum_{i=1}^N E_Q[\ln p(x_i | \theta^*, z_i)] + \\ &E_Q[\ln p(\alpha^* | a_\alpha, b_\alpha)] - E_Q[\ln Q(V, z, \theta^*)],\end{aligned}$$

其中各分量中间变量可分解为

$$Q(\theta^*) = Q(\sigma^2)Q(\eta)Q(w^2)Q(\beta)Q(g)Q(\delta)Q(\mu)Q(r),$$

变分推断的任务就是要求得各因子的分布形式. 以下

下标记 $\langle y \rangle$ 为 $E_Q(y)$, $\Psi(y) = \frac{\partial}{\partial y} \Gamma(y)$, 设定

$$Q(z_i = k) = \phi_{ik},$$

$$Q(z_i > k) = \sum_{t=k+1}^T \phi_{it},$$

对于 $Q(V_k)$ 有

$$\gamma_{k,1} = 1 + \sum_{i=1}^N \phi_{ik},$$

$$\gamma_{k,2} = \langle \alpha \rangle + \sum_{i=1}^N \sum_{\varsigma=k+1}^T \phi_{i\varsigma},$$

则 $Q(V_k) \propto \text{Beta}(\gamma_{k,1}, \gamma_{k,2})$

$$\langle \ln V_k \rangle = \Psi(\gamma_{k,1}) - \Psi(\gamma_{k,1} + \gamma_{k,2}),$$

$$\langle \ln(1 - V_k) \rangle = \Psi(\gamma_{k,2}) - \Psi(\gamma_{k,1} + \gamma_{k,2}),$$

因此对于 Φ_{ik} 有

$$(\phi_{i1}, \phi_{i2}, \dots, \phi_{iT}) \propto$$

$$\text{Dirchlet}(\exp(S_1), \exp(S_2), \dots, \exp(S_T))$$

$$\phi_{ik} \propto \exp(S_k) \quad (15)$$

$$S_k = \langle \ln V_k \rangle + \sum_{\varsigma=1}^{k-1} \langle \ln(1 - V_{\varsigma}) \rangle +$$

$$\sum_{j=1}^D \varphi_{kj} [-\ln(2\pi) + \langle \ln(1/w_{kj}^2) \rangle +$$

$$\langle \ln(1/\sigma_j^2) \rangle] / 2 -$$

$$\sum_{j=1}^D \varphi_{kj} [(x_{ij}^2 + \langle \delta_{kj}^2 \rangle + \langle \mu_j^2 \rangle -$$

$$2x_{ij} \langle \delta_{kj} \rangle - 2x_{ij} \langle \mu_j \rangle +$$

$$2x_{ij} \langle \delta_{kj} \rangle \langle \mu_j \rangle) / (\langle w_{kj}^2 \rangle \langle \sigma_j^2 \rangle)] / 2 +$$

$$\sum_{j=1}^D (1 - \varphi_{kj}) [-\ln(2\pi) + \langle \ln(1/\sigma_j^2) \rangle] / 2 -$$

$$\sum_{j=1}^D (1 - \varphi_{kj}) [(x_{ij}^2 + \langle \mu_j^2 \rangle - 2x_{ij} \langle \mu_j \rangle) / \langle \sigma_j^2 \rangle] / 2,$$

$$\sum_{k=1}^T \langle \ln p(V_K | \alpha^*) \rangle$$

$$= \sum_{k=1}^T [\ln \langle \alpha \rangle + (\langle \alpha \rangle - 1) \langle \ln(1 - V_k) \rangle]$$

$$= T \ln \langle \alpha \rangle + (\langle \alpha \rangle - 1) \cdot$$

$$\sum_{k=1}^T [\Psi(\gamma_{k,2}) - \Psi(\gamma_{k,1} + \gamma_{k,2})];$$

对于 $Q(\alpha)$ 有

$$\bar{a}_\alpha = a_\alpha + T - 1,$$

$$\bar{b}_\alpha = b_\alpha - \sum_{k=1}^{T-1} \langle \ln(1 - V_k) \rangle,$$

则 $Q(\alpha) \propto G(\bar{a}_\alpha, \bar{b}_\alpha)$,

$$\langle \alpha \rangle = \bar{a}_\alpha / \bar{b}_\alpha \quad (16)$$

$$\langle \ln p(\alpha^* | a_\alpha, b_\alpha) \rangle = \Psi(\bar{a}_\alpha) - \ln(\bar{b}_\alpha).$$

而求各分量变分参数则可参考基于高斯混合的聚类 EM 算法, 也可看作前面给出的 MCMC 抽样中类别和特征选择软赋值 (soft assignment) 时的后验推断. 特别地本模型各参数选用的先验分布与后验分布构成共轭对, 因而可容易地写出各变分参数值.

(1) $Q(r_{kj})$

$$FS_{kj0} = \langle \ln(1 - g_j) \rangle + \sum_{i=1}^N \phi_{ij} \langle \ln(1/\sigma_j^2) \rangle / 2 -$$

$$\sum_{i=1}^N \phi_{ik} [(x_{ij}^2 + \langle \mu_j^2 \rangle - 2x_{ij} \langle \mu_j \rangle) / \langle \sigma_j^2 \rangle] / 2,$$

$$FS_{kj1} = \langle \ln g_j \rangle + \sum_{i=1}^N \phi_{ij} [\langle \ln(1/\sigma_j^2) \rangle + \langle \ln(1/w_{kj}^2) \rangle] / 2 -$$

$$\sum_{i=1}^N \phi_{ik} [(x_{ij}^2 + \langle \delta_{kj}^2 \rangle + \langle \mu_j^2 \rangle - 2x_{ij} \langle \delta_{kj} \rangle - 2x_{ij} \langle \mu_j \rangle +$$

$$2x_{ij} \langle \delta_{kj} \rangle \langle \mu_j \rangle) / (\langle w_{kj}^2 \rangle \langle \sigma_j^2 \rangle)] / 2,$$

$$Q(r_{kj}) \propto \text{Beta}(\exp(FS_{kj0}), \exp(FS_{kj1})),$$

$$Q(r_{kj} = 1) = \varphi_{kj},$$

$$Q(r_{kj} = 0) = 1 - \varphi_{kj},$$

$$\varphi_{kj} = \exp(FS_{kj1}) / (\exp(FS_{kj0}) + \exp(FS_{kj1})) \quad (17)$$

(2) $Q(g_j)$

$$\bar{a}_{g_j} = (\langle \beta \rangle / D) + \sum_{k=1}^T \varphi_{kj},$$

$$\bar{b}_{g_j} = 1 + \sum_{k=1}^T (1 - \varphi_{kj}),$$

则 $Q(g_j) \propto \text{Beta}(\bar{a}_{g_j}, \bar{b}_{g_j})$,

$$\langle \ln(g_j) \rangle = \Psi(\bar{a}_{g_j}) - \Psi(T + (\langle \beta \rangle / D) + 1) \quad (18)$$

$$\langle \ln(1 - g_j) \rangle = \Psi(\bar{b}_{g_j}) - \Psi(T + (\langle \beta \rangle / D) + 1) \quad (19)$$

(3) $Q(\beta)$

$$\bar{a}_\beta = a_\beta + D - 1,$$

$$\bar{b}_\beta = b_\beta - \sum_{j=1}^D \langle \ln(g_j) \rangle / D,$$

则 $Q(\beta) \propto G(\bar{a}_\beta, \bar{b}_\beta)$,

$$\langle \beta \rangle = \bar{a}_\beta / \bar{b}_\beta \quad (20)$$

$$\langle \ln p(\beta | a_\beta, b_\beta) \rangle = \Psi(\bar{a}_\beta) - \ln(\bar{b}_\beta).$$

(4) $Q(\eta)$

$$\bar{a}_\eta = a_\eta + T \times D / 2,$$

$$\bar{b}_\eta = b_\eta + \frac{1}{2} \sum_{k=1}^T \sum_{j=1}^D \langle \delta_{kj}^2 \rangle / (\langle w_{kj}^2 \rangle \langle \sigma_j^2 \rangle),$$

则 $Q(\eta) \propto IG(\bar{a}_\eta, \bar{b}_\eta)$,

$$\langle \eta \rangle = \bar{b}_\eta / \bar{a}_\eta \quad (21)$$

$$\langle w_{kj}^2 \rangle = \bar{b}_{wkj} / \bar{a}_{wkj} \quad (28)$$

$$(5) Q(\mu_j)$$

$$\langle \ln(1/w_{kj}^2) \rangle = \Psi(\bar{a}_{wkj}) - \ln(\bar{b}_{wkj}).$$

$$\bar{c}_j = \left(\frac{1}{c} + \sum_{i=1}^N \sum_{k=1}^T \phi_{ik} \varphi_{kj} / (\langle w_{kj}^2 \rangle \langle \sigma_j^2 \rangle) + \right.$$

$$\left. \sum_{i=1}^N \sum_{k=1}^T \phi_{ik} (1 - \varphi_{kj}) / \langle 1/\sigma_j^2 \rangle \right)^{-1},$$

$$\bar{m}_j = \bar{c}_j \times \left(\frac{m_j}{c} + \sum_{i=1}^N \sum_{k=1}^T \phi_{ik} \varphi_{kj} \times (x_{ij} - \delta_{kj}) / \right.$$

$$\left. (\langle w_{kj}^2 \rangle \langle \sigma_j^2 \rangle) + \sum_{i=1}^N \sum_{k=1}^T \phi_{ik} (1 - \varphi_{kj}) \times x_{ij} / \langle \sigma_j^2 \rangle \right),$$

$$\text{则 } Q(\mu_j) \propto N(\bar{m}_j, \bar{c}_j)$$

$$\langle \mu_j \rangle = \bar{m}_j \quad (22)$$

$$\langle \mu_j^2 \rangle = \bar{c}_j + \bar{m}_j \times \bar{m}_j \quad (23)$$

$$(6) Q(\sigma_j^2)$$

$$\bar{v}_{1j} = v_1 + \left(\sum_{i=1}^N \sum_{k=1}^T \phi_{ik} + T \right) / 2,$$

$$\bar{v}_{2j} = v_{2j} + \left[\sum_{i=1}^N \sum_{k=1}^T \phi_{ik} \varphi_{kj} (x_{ij}^2 + \langle \delta_{kj}^2 \rangle + \langle \mu_j^2 \rangle - \right.$$

$$\left. 2x_{ij} \langle \delta_{kj} \rangle - 2x_{ij} \langle \mu_j \rangle + 2 \langle \delta_{kj} \rangle \langle \mu_j \rangle \right] / 2 +$$

$$\left[\sum_{i=1}^N \sum_{k=1}^T \phi_{ik} (1 - \varphi_{kj}) (x_{ij}^2 + \langle \mu_j^2 \rangle - 2x_{ij} \langle \mu_j \rangle) \right] / 2 +$$

$$\left[\sum_{k=1}^T \langle \delta_{kj}^2 \rangle / (\langle \eta \rangle \langle w_{kj}^2 \rangle) \right] / 2,$$

$$\text{则 } Q(\sigma_j^2) \propto IG(\bar{v}_{1j}, \bar{v}_{2j}),$$

$$\langle 1/\sigma_j^2 \rangle = \bar{v}_{1j} / \bar{v}_{2j} \quad (24)$$

$$\langle \sigma_j^2 \rangle = \bar{v}_{2j} / \bar{v}_{1j} \quad (25)$$

$$\langle \ln(1/\sigma_j^2) \rangle = \Psi(\bar{v}_{1j}) - \ln(\bar{v}_{2j}).$$

$$(7) Q(\delta_{kj}^2)$$

$$\bar{\tau}_{kj}^2 = \sum_{i=1}^N \phi_{ik} \varphi_{kj} / (\langle w_{kj}^2 \rangle \langle \sigma_j^2 \rangle) + 1 / (\langle \eta \rangle \langle w_{kj}^2 \rangle \langle \sigma_j^2 \rangle),$$

$$\bar{\delta}_{kj} = \sum_{i=1}^N \phi_{ik} \varphi_{kj} (x_{ij} - \langle \mu_j \rangle) / (\langle w_{kj}^2 \rangle \langle \sigma_j^2 \rangle),$$

$$\text{则 } Q(\delta_{kj}) \propto N(\bar{\delta}_{kj}, \bar{\tau}_{kj}^2),$$

$$\langle \delta_{kj} \rangle = \bar{\delta}_{kj} / \bar{\tau}_{kj}^2 \quad (26)$$

$$\langle \delta_{kj}^2 \rangle = \langle \delta_{kj} \rangle \langle \delta_{kj} \rangle + 1 / \bar{\tau}_{kj}^2 \quad (27)$$

$$(8) Q(w_{kj}^2)$$

$$\bar{a}_{wkj} = a_w + \left(\sum_{i=1}^N \phi_{ik} \varphi_{kj} + 1 \right) / 2,$$

$$\bar{b}_{wkj} = b_w + \left[\sum_{i=1}^N \phi_{ik} \varphi_{kj} (x_{ij}^2 + \langle \mu_j^2 \rangle + \langle \delta_{kj}^2 \rangle - \right.$$

$$\left. 2 \langle \mu_j \rangle x_{ij} - 2 \langle \delta_{kj} \rangle x_{ij} + 2 \langle \mu_j \rangle \langle \delta_{kj} \rangle x_{ij} \right] / 2 +$$

$$\left[\langle \delta_{kj}^2 \rangle / (\langle \eta \rangle \langle \sigma_j^2 \rangle) \right] / 2,$$

$$\text{则 } Q(w_{kj}^2) \propto IG(\bar{a}_{wkj}, \bar{b}_{wkj}),$$

迭代执行如上步骤,直至下界函数 $L[\mathbf{Q}, \mathcal{D}]$ 不再增加. 由于“断棒”法对各分量混合权值先验是按分量大小有序排列的,因此在每一次迭代后根据逼近的各分量大小重新排列各分量并调整相应的分量参数次序.

5 实验结果

以下给出变分法的实验结果. MCMC 抽样方法在文献[8]中显示了此模型的一些优点,但是要经历几万次的迭代抽样,在当前的模式识别与机器学习应用领域不是理想的方法. 然而一些生物数据的特征数远远大于样本个数且对聚类准确度要求较高时,对其进行 MCMC 抽样则能够发挥它的优点.

如果两个或更多的分量参数相似, Gibbs 抽样易陷入局部模式,收敛速度非常慢,不能正确聚类. 特别地本模型还要抽样 r_{kj} , 其值与分量参数高度耦合,如不设定合适的迭代初始值,即使应用 split-merge 方法^[12-13],也很难保证 MCMC 抽样收敛. 文献[14]也注意到 Dirichlet 混合模型存在的这个严重问题,应用 A^* 搜索方法求 Dirichlet 混合的聚类结果,或将其结果作为 MCMC 抽样的迭代初始值. 本文在进行 MCMC 抽样试验时的迭代初始值利用 k 均值聚类结果,其类别数 T 一般比真实的类别数大,采用欧氏距离作为 k 均值聚类的距离度量,随机设定类中心点,迭代 100 次. 设最后得到的各类中心为 $\hat{\delta}_k$, 观测数据均值为 μ_x , 则 μ 的先验超参数 m 取 μ_x , 各类平移初始值 $\delta_k = \hat{\delta}_k - \mu_x$, 参数 w^2 和 σ^2 按先验分布抽取, $\eta=1, r_{kj}$ 随机取 0 或 1. 而且由于 Gibbs 抽样每次只移动一个观测数据到一个新的混合分量中,因此在一遍 Gibbs 抽样后再利用 split-merge 算法移动一组观测数据到一个新的混合分量中. 在整个迭代结束后可将只包含很少几个观测数据的分量去除.

变分法不需抽样,但其性能可能更依赖于初始值的选择. 以下试验各参数初始值选择如下: $\Phi_{ik} \approx 1/T$; φ_{kj} 取标准分布随即数; 设 μ_x 为数据集的均值, S_x 为数据集各特征的方差, μ 的先验超参数 m 取 μ_x , 先验超参数 c 取 10^6 ; $v_1 = 1/2$, $v_{2j} = S_{xj}/2$; $\langle \sigma_j^2 \rangle = S_{xj}$; $a_\eta = 1/2$, $b_\eta = 1/2$, $\langle \eta \rangle = 1$; $a_w = 3$, $b_w = 2$, $\langle w^2 \rangle = 0.1$; $\langle \delta_{kj} \rangle \sim N(0, \langle \eta \rangle \langle w_{kj}^2 \rangle \langle \sigma_j^2 \rangle)$. 而 α 和 β 及它们的超参数则根据用户对可能聚类个数及特征

选择的先验知识而设定. β 一般取两类之间可能有特征均值平移或方差伸缩的特征数. 在本文试验中截断的聚类个数 T 一般大于目标聚类个数. 在迭代过程中对于混合权值后验均值很小的分量也予以去除.

本算法的主要功能是在不同的特征子集上进行聚类, 不是特征选择. 评价子空间聚类的标准在文献[6]中已有一定的研究, 本文仍采取校正的 Rand 指数作为性能评价标准^[15]. 本文模型是对 Hoff 模型的扩充, 对特征平移或方差伸缩的特征选择引入了非参数统计模型, 但在进行 MCMC 抽样时两者聚类性能没有本质差异, 由于是经抽样得到聚类, 也很难比较两者的聚类性能. 本文改进的模型对比 Hoff 模型的一个主要优点是便于利用变分法推断模型参数, 而变分法则无需抽样, 在计算速度上得到很大的提高. 因此以下实验给出的 MCMC 结果均为根据本文模型前述参数后验分布抽样迭代的结果.

5.1 仿真数据

仿真数据类似文献[7-8]中的仿真数据, 100 个目标对象, 1000 个特征. 其中 85 个对象为一类, 每个特征服从均值为 0、方差为 1 的标准分布; 其余 15 个目标为另一类, 其中有 150 个特征服从均值为 1.5、方差为 0.2 的标准分布, 其他特征也服从均值为 0、方差为 1 的标准分布. 两类之间的部分特征既有均值平移也有方差伸缩. 应用变分贝叶斯学习方法的执行结果如图 2 所示, 迭代执行 50 次就能正确聚类(校正的 Rand 指数值为 1), 约 100 次左右目标函数下界不再增加, 对 φ_{kj} 二值化得到 r_{kj} , 含 15 个目标的类别 k 对应应有均值平移和方差伸缩的 15 个特征 r_{kj} 值为 1, 其余的 r_{kj} 值都为 0. 应用 MCMC 方法也只需要 100 次左右的迭代就能达到均衡分布. 同时说明对于这样一类具有很强聚类模式的数据,

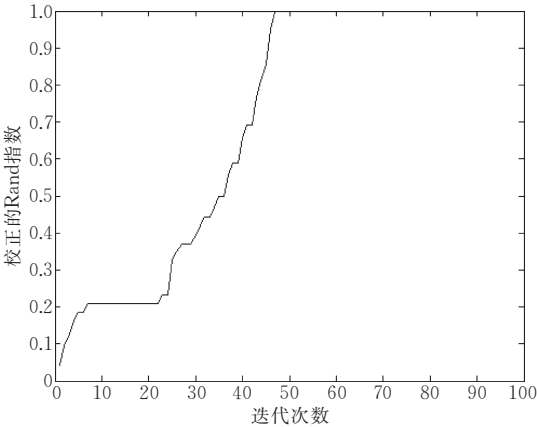


图 2 仿真数据校正的 Rand 指数

应用变分贝叶斯方法和 MCMC 方法都能很快地得到正确的聚类结果.

5.2 真实数据

首先的实验是一个关于人脸聚类的问题. 数据来源于 Stirling 人脸数据库^①, 选择所有的 68 个前视(类 1)和 105 个侧视(类 2)灰度人脸图像. 对这些人脸进行特征脸抽取的过程如文献[3]所示, 选取人脸在 10 个主特征分量上的投影系数为特征值. 根据主分量分析的性质, 经主分量分析得到的投影系数服从高斯分布, 故本文的模型很适合此类问题. 类 1 和类 2 数据各特征的分布情况分别如图 3, 4 所示, 均用箱线图表示. 图 5 表示观测数据总体及各类均值. 从上述三张图可大致直观地得到两类在各特征之间的差异: 在特征 1 和 4 上存在显著的均值平移, 而在特征 2、3、5、6、9 上两类的方差差异较均值差异更加明显, 在特征 7、8 和 10 上两类则没有均值显著平移或方差大小显著不同的现象.

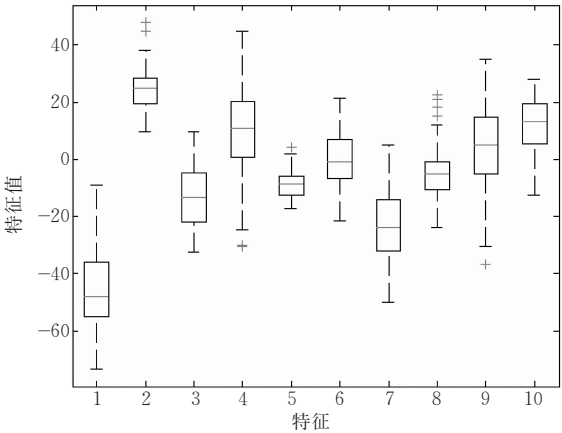


图 3 类 1 各特征分布

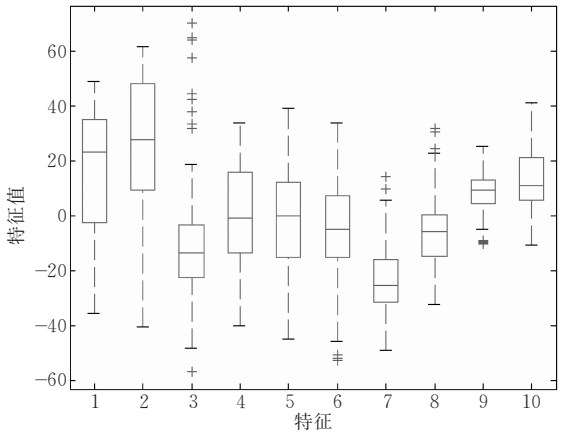


图 4 类 2 各特征分布

① <http://pics.psych.stir.ac.uk/cgibin/PICS/New/pics.cgi>

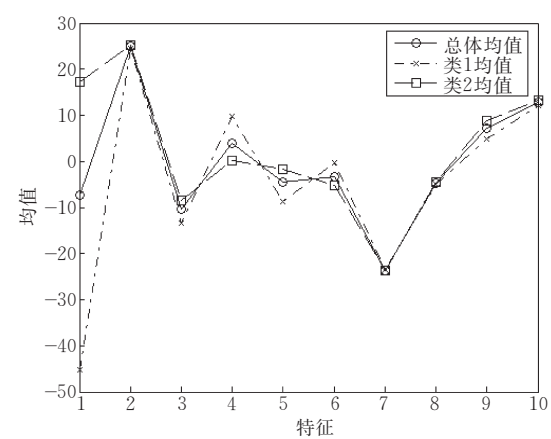


图 5 数据总体及各类均值

应用本文提出的变分贝叶斯推断算法根据不同的初始值运行 20 次,平均错误区分前、侧视人脸个数为 4~5 个,低于文献[3]的 7 个.应用 MCMC 抽样方法的准确度更高,只误分其中的 3 个,聚类结果的基准均值 μ_j 和各类均值 $\mu_{kj} = \mu_j + r_{kj} \times w_{kj} \times \delta_{kj}$, $k=1,2$; $j=1,2,\dots,10$.

如图 6 所示:在特征 7、8 和 10 上两类均值点重合 ($r_{1j} = r_{2j} = 0, j=7,8,10$),即两类没有均值平移或方差伸缩,说明这些特征对所有类别都没有判别信息.类 2 特征 4 的均值与基准均值重合,但类 1 有平移 ($r_{24} = 0, r_{14} = 1$);类 1 特征 9 的均值与基准均值重合,类 2 有平移 ($r_{19} = 0, r_{29} = 1$),因此类 1 和类 2 在不同的特征子集上有判别信息存在,也正是子空间聚类思想的体现.对于其他特征两类相对基准均值都有平移项.因此实验结果反映了两类在不同特征子集上的聚类特性.

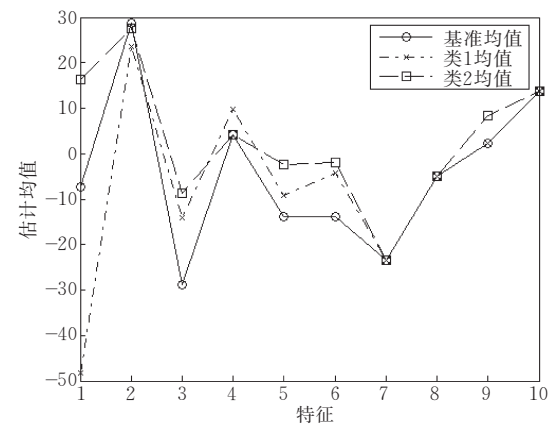


图 6 子空间聚类估计的基准均值及各类均值

接下来的试验是对手写数字(“0”~“4”)进行聚类.数据来自 UCI 的多特征数据库^①,每类 200 个样本,共 1000 个样本.对每个大小为 16×16 的灰度图像提取不同的识别特征,如文献[2]所述,本文选取

Zernike 矩(Zernike,47 个特征)、傅立叶系数(Fourier,76 个特征)和轮廓相关系数(Profile,216 个特征)等三个特征集,在每个特征集上应用本文算法和 k -均值聚类算法, k -均值聚类初始值设定也如前所述,只是将初始聚类个数设为理想的目标聚类个数,迭代 100 次或收敛标准已满足.两方法分别根据不同的随机初始值运行 20 次,计算平均的校正的 Rand 指数,结果如表 1 所示.由于进行 MCMC 抽样计算时间很长,其主要原因是对于高维大规模数据问题一方面需要同时运行几条马尔可夫链,在迭代过程交换部分参数,可能还需采用退火技术,否则易陷入局部模式;另一方面对 r_{kj} 抽样需利用 Metropolis-Hastings 接受概率方法,同时更新 r 矩阵一行中的几列值^[16],也是一个很耗时的过程.故没有给出应用 MCMC 方法的结果.

从表 1 可看出,当特征维数较低时,本文算法性能略低于 k -均值聚类算法,但是当特征维数较高时,如在 Profile 特征集上进行聚类时,本算法性能优于 k -均值聚类算法.一方面说明本算法对高维特征数据集有较好的聚类效果,能发现各个类别在不同特征子集上均值与方差的差异;另一方面说明由于本文模型假定各特征独立,而低维特征集之间可能存在较大的相关性,因此导致聚类性能下降.

表 1 应用本文方法与 k -均值算法手写数字聚类校正的 Rand 指数

| 特征集 | Rand 指数 | | |
|-----------|---------|---------|---------|
| | Zernike | Fourier | Profile |
| k -均值算法 | 0.61 | 0.57 | 0.78 |
| 本文的 VB 算法 | 0.55 | 0.52 | 0.87 |

另外,建立统计模型进行聚类相对 k -均值聚类及层级聚类启发式算法的优势还在于其后验推断能对聚类结果、各特征均值平移及方差伸缩等的不确定性进行度量,自动进行模型选择,特别的统计模型也能对含有缺失数据的目标进行聚类.

6 结语和讨论

在本文中,我们在子空间聚类 Hoff 模型的基础上对特征均值和方差平移引入非参数统计模型,并给出子空间聚类模型的变分贝叶斯参数推断方法,在几个模拟和真实数据集上验证了本算法能根据数据集本身的聚类特性和各特征均值与方差的不同,

① <http://www.ics.uci.edu/mllearn/MLRepository.html>

进行子空间聚类,运用变分贝叶斯方法逼近各参数的后验分布,大大减少了算法运行时间.

但是应注意到变分法存在依赖初始值且可能陷入局部较小的缺点. 作者在几个非常高维的生物信息数据集上运用变分贝叶斯方法不能取得聚类结果说明了此问题仍值得深入探究. 一方面需研究如何避免陷入局部极小,使得下界逼近更紧. 确定性退火技术是较可行的技术^[17];另一方面 Dirichlet 过程和选择特征均值平移或方差伸缩的非参数过程变分贝叶斯方法还有待深入研究,特别是生物信息数据特征在空间上具有马尔可夫特性,研究此类模型的变分贝叶斯推断方法将是本文接下来的工作.

针对 Hoff 模型各特征独立这一假设,已有一些假定特征非独立生成而同时进行聚类 and 特征选择的工作^[18-19],但是只能应用 MCMC 抽样方法进行后验推断,实验对象只有几十个数据. 对大规模数据集其计算量是难以接受的. 如应用变分贝叶斯方法,并采取自动相关决定 (Automatic Relevance Determination, ARD) 框架选取特征,但是协方差更新的计算量仍是相当大的.

参 考 文 献

- [1] Law M H, Figueiredo M A T, Jain A K. Simultaneous feature selection and clustering using a mixture model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(9): 845-889
- [2] Constantinopoulos C, Titsias M K, Likas A. Bayesian feature and model selection for Gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, 28(6): 1013-1018
- [3] Roth V, Lange T. Feature selection in clustering problems// *Proceedings of the Advances in Neural Information Processing Systems*. Vancouver, Canada, 2003, 16: 1237-1244
- [4] Agrawal R, Gehrke J, Gunopulous D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications// *Proceedings of the ACM SIGMOD International Conference Management of Data*. Seattle WA, 1998: 94-105
- [5] Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data, a review. *ACM SIGKDD Explorations News Letter*, 2004, 6(1): 90-105
- [6] Patrikainen A, Meila M. Comparing subspace clusterings. *IEEE Transactions on Knowledge and Data Engineering*, 2006, 18(7): 902-916
- [7] Friedman J H, Meulman J J. Clustering objects on subsets of attributes. *Journal Royal Statistical Society Series B Statistical Methodology*, 2004, 66(4): 815-849
- [8] Hoff P D. Model-based subspace clustering. *Bayesian Analysis*, 2006, 1(2): 321-344
- [9] Griffiths T L, Ghahramani Z. Infinite latent feature models and the Indian buffet process// *Proceedings of the Advances in Neural Information Processing Systems*. Vancouver, Canada, 2005, 18: 475-482
- [10] Beal M J. Variational algorithms for approximation Bayesian inference [Ph. D. dissertation]. Gatsby Computational Neuroscience Unit, University of College, London, UK, 2003
- [11] Blei D M, Jordan M I. Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis*, 2005, 1(1): 121-144
- [12] Jain S, Neal R. A split-merge markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 2004, 13(1): 158-182
- [13] Jain S, Neal R. Splitting and merging components of a non-conjugate Dirichlet process mixture model. *Bayesian Analysis*, 2005, 1(5): 1-38
- [14] Hal Daume III. Fast search for Dirichlet process mixture models// *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*. San Juan, Puerto Rico, 2007
- [15] Yang Chun-Mei, Wan Bai-Kun, Gao Xiao-Feng. Selections of data preprocessing methods and similarity metrics for gene cluster analysis. *Progress in Natural Science*, 2006, 16(6): 607-613
- [16] Meeds E, Ghahramani Z, Neal R, Sam R. Modeling dyadic data with binary latent factors// *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*. San Juan, Puerto Rico, 2007
- [17] Ghahramani Z, Hinton G E. Variational learning for switching state-space models. *Neural Computation*, 2000, 12(44): 831-864
- [18] Tadesse M G, Sha N, Vannucci M. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 2005, 100(470): 602-617
- [19] Kim S, Tadesse M G, Vannucci M. Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 2006, 93(4): 877-893



QING Xiang-Yun, born in 1977, Ph. D. candidate. His main research interests include statistical machine learning and pattern recognition.

WANG Xing-Yu, born in 1944, professor, Ph. D. supervisor. His research interests include intelligence control theory and pattern recognition etc.

Background

Various types of tasks in some specification domains, such as image segmentation, text and image classification, web semantic information extraction, etc., can be viewed as a clustering problem to solve. The goal of cluster analysis is to group a data set into clusters such that those data points in each cluster are more similar to each other than to those of other clusters. As one of the most fundamental unsupervised learning problems, it has been studied widely in the literature. However, data represented by a number of features may have discriminative information only on the subset of features. In particular, individual clusters may represent grouping on different (possibly overlapping) feature subsets, and it is interesting to discover such patterns that highlight different facets of the similarity between the data points. Therefore, subspace clustering was proposed in order to solve the problem of simultaneously choosing the subset of features and selecting the data points given those features.

The first subspace algorithm, CLIQUE was proposed by Agrawal R et al. in 1998 and was soon followed by many related methods. Friedman and Meulman (2004) developed a clustering algorithm on subset of attributes, whose clustering criteria and computational approaches were largely driven by heuristics. Hoff (2006) presented a model-based subspace clustering methods based on finding groups which differ from each other in terms of their means and/or variances at one or

more attributes. However, the model of "clustering shifts in mean and variance" learned by Markov Chain Monte Carlo and the computational cost may be prohibitive. So the authors extend their model to a new unified nonparametric model such that variational Bayesian method can be applied to accelerate estimation of parameters. Variational Bayesian approximations have been widely used in Bayesian learning to offset the high computational cost of exact Bayesian calculations. Nonparametric model only need little prior information and model selection is decided by data itself. The authors' model can simultaneously optimize over the number of components, the subsets of features to each of components and the parameters of the model under the MCMC and variational frameworks.

The research is partially supported by National Natural Science Foundation of China under grant No. 60674089 and the Doctoral Program of the Ministry of Education under grant No. 20040251010. One mission of these two projects is to develop a algorithm that can automatically partition signals into different clusters and discover latent patterns from human's EEG. The research work of this paper, as a part of fundamental theory work, will be applied to these projects. The study of the team aims at meeting with the new international research tendency and integrating the studies on control theory, machine learning and brain signal.