doi:10.3969/j.issn.1001-4616.2017.03.010

一种基于逆模拟退火和高斯混合模型的半监督聚类算法

王　垚,柴变芳,李文斌,吕　峰

(河北地质大学信息工程学院,河北 石家庄 050031)

[摘要]　基于节点标记的半监督高斯混合模型(Semi-supervised Gaussian Mixture Model,SGMM)可利用少量标记样本提高模型参数估计的准确率,但参数估计算法(SGMM Expectation Maximization,SGMM-EM)的准确率和收敛速度受高斯分布之间的重叠度和混和系数差异度影响. 为提高 SGMM 模型参数估计的准确率和收敛速度,将逆模拟退火框架与 SGMM 模型的 EM 算法相结合,提出一种基于逆模拟退火框架的半监督高斯混合模型聚类算法(Anti-annealing SGMM-EM,ASGMM-EM). 该算法逆温度参数从一个较小且大于 0 的值逐渐增加到大于 1 的上界,再逐渐降回 1. 在每个逆温度参数下执行半监督聚类算法 SGMM-EM 并迭代至收敛. 人工数据和真实数据上实验表明提出的算法 ASGMM-EM 优于仅用半监督技术或逆模拟退火技术的基于高斯混合模型的 EM 算法.

[关键词]　高斯混合模型,期望最大化算法,逆模拟退火,半监督聚类
[中图分类号]TP391　[文献标志码]A　[文章编号]1001-4616(2017)03-0067-07

A Semi-supervised Clustering Algorithm Based on Anti-annealing and Gaussian Mixture Model

Wang Yao,Chai Bianfang,Li Wenbin,Lü Feng

(School of Information Engineering,Hebei GEO University,Shijiazhuang 050031,China)

Abstract:Semi-supervised Gaussian mixture model(SGMM)based on labeling nodes can improve the accuracy of model parameter estimation. However, the accuracy and convergence of the Expectation Maximization (EM) algorithm are affected by the amount of overlap and mixing coefficients among the Gaussian distributions. In order to improve the accuracy and speed of the SGMM parameter estimation, the Anti-annealing is combined with the EM algorithm of SGMM. A clustering algorithm of the semi-supervised Gaussian mixture model based on anti-annealing (ASGMM-EM) is proposed. The inverse temperature parameter of the algorithm increases from a smaller value to an upper bound that more than 1 and then back to 1. The semi-supervised clustering EM algorithm is implemented at each inverse temperature parameter. Experiments on synthetic and real data show that the ASGMM-EM is better compared to the algorithms only using semi-supervised or anti-annealing technique.

Key words:Gaussian mixture model,expectation maximization algorithm,anti-annealing,semi-supervised clustering

聚类是一种广泛应用于计算机科学、生物信息学[1]、图像识别[2]、文档分类[3]等很多领域的数据分析技术. 聚类算法种类繁多,基于模型的聚类算法因其具有较强的理论基础和可解释性,在现实问题中得到了广泛应用. 高斯混合模型(Gaussian Mixture Model,GMM)是其中最典型、最常用的一种聚类模型. GMM 假定数据集合为多个高斯分布的混合,每一个高斯分布对应于一个类. 基于传统 GMM 的聚类算法可实现未标记样本集合的聚类,但是有些数据集合已知少量样本的标记信息,故无监督 GMM 无法利用这些先验信息,而半监督高斯混合模型(Semi-supervised GMM,SGMM)可利用不同形式的先验信息提高基于传统 GMM 的聚类性能.

收稿日期:2017-03-18.
基金项目:国家自然科学基金(61503260)、河北省研究生创新资助项目(CXZZSS2017131).
通讯联系人:李文斌,博士后,教授,研究方向:机器学习、复杂网络等. E-mail:25304189@qq.com

先验信息的形式主要包括样本约束对标记和样本标记. 其中,基于约束对标记的先验方式主要分两类:一类将先验信息融入模型的似然函数[4-6],另一类将先验信息以正则化项形式惩罚似然函数[7-8]. 虽然 SGMM 可利用除未标记样本之外的先验信息以提高聚类性能,但是 SGMM 仍存在些许缺陷. 例如,SGMM 没有考虑数据空间的基础流型结构,故其参数初始化的计算复杂度非常高. 对此,文献[9]提出了一种半监督 LCGMM 算法(Semi-LCGMM),该算法结合了局部一致高斯混合模型[10](Locally Consistent GMM,LCGMM)和 SGMM 两者的优点,显著提高了处理具有复杂底层结构的数据集的聚类性能. 文献[11]通过组合缩减的标记样本集和大量未标记样本集,可显著提高图像分割聚类精度.

基于节点标记的 SGMM 使用半监督期望最大化算法 SGMM-EM 估计模型参数,SGMM-EM 算法直接利用节点的类标记作为先验信息,并直接以概率值参与模型参数估计[11-12]. 由于 GMM 在类不均衡或类间重叠度大时,EM 算法参数估计不准确且收敛速度慢. 文献[13]提出一种基于逆模拟退火框架的高斯混合模型参数估计算法(Anti-annealing GMM-EM,AGMM-EM),但其主要针对无监督 GMM 的 EM 算法参数估计问题. 本文针对 SGMM 的 EM 参数估计算法存在的缺陷,提出一种基于逆模拟退火框架的半监督聚类算法. 该算法逆温度参数从一个较小的值开始增加到大于 1 的上界,再降回 1. 在每个逆温度参数下执行半监督聚类算法 SGMM-EM. 实验表明此改进可有效提高 SGMM-EM 算法在 SGMM 在类不均衡或类间重叠度大时的聚类性能.

1　半监督高斯混合模型

假设样本由高斯混合模型产生,其概率密度函数定义为:

$$p(x)=\sum_{i=1}^{k}\alpha_i\cdot\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i),\tag{1}$$

$$\text{s.t. }\alpha_i>0,\ \sum_{j=1}^{k}\alpha_i=1,\tag{2}$$

式中,α_i 为第 i 个高斯分布的混合系数,需满足条件(2),$\boldsymbol{\mu}_i$ 为第 i 个高斯分布的 n 维均值向量,$\boldsymbol{\Sigma}_i$ 为第 i 个高斯分布的 $n\times n$ 的协方差矩阵. $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ 为样本 x 由第 i 个高斯分布生成时的概率密度.

无监督 GMM 实现对无标记样本集合的聚类,假设每个样本由高斯混合模型产生,基于最大似然准则得到聚类的目标函数,采用 EM 算法估计模型参数. 基于节点标记的 SGMM 通过引入少量的已标记的数据作为先验信息,在先验信息的指引下可提升传统高斯混合模型的聚类性能.

给定有标记样本集 $D_l=\{(\boldsymbol{x}_1,\lambda_1),(\boldsymbol{x}_2,\lambda_2),\cdots,(\boldsymbol{x}_l,\lambda_l)\}$ 和未标记样本集 $D_u=\{\boldsymbol{x}_{l+1},\boldsymbol{x}_{l+2},\cdots,\boldsymbol{x}_{l+u}\}$. 其中,$\boldsymbol{x}_i$ 表示第 i 个样本,λ_i 表示其标记,D_l 为有标记样本集合,D_u 为无标记样本集合. $l\ll u$,$l+u=n$. SGMM 的基于极大似然准则的目标函数为:

$$LL(D_l\cup D_u)=\sum_{(\boldsymbol{x}_j,\lambda_j)\in D_l}\ln\Big(\sum_{i=1}^{k}\alpha_i\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\cdot p(\lambda_j\mid\boldsymbol{C}_i,\boldsymbol{x}_j)\Big)+\sum_{\boldsymbol{x}_j\in D_u}\ln\Big(\sum_{i=1}^{k}\alpha_i\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\Big),\tag{3}$$

采用 EM 算法求解模型参数,迭代更新公式如下:

E 步,根据当前模型参数计算未标记样本 $\boldsymbol{x}_j$ 属于各混合分布的后验概率:

$$\varphi_{ji}=\frac{\alpha_i p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\sum_{i=1}^{k}\alpha_i p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}.\tag{4}$$

M 步,利用属于当前类的所有样本及后验概率 φ_{ji} 更新模型参数:

$$\boldsymbol{\mu}_i=\frac{\Big(\sum_{\boldsymbol{x}_j\in D_u}\varphi_{ji}\boldsymbol{x}_j+\sum_{(\boldsymbol{x}_j,\lambda_j)\in D_l\wedge\lambda_j=i}\boldsymbol{x}_j\Big)}{\sum_{\boldsymbol{x}_j\in D_u}\varphi_{ji}+l_i},\tag{5}$$

$$\boldsymbol{\Sigma}_i=\frac{1}{\sum_{\boldsymbol{x}_j\in D_u}\varphi_{ji}+l_i}\Big(\sum_{\boldsymbol{x}_j\in D_u}\varphi_{ji}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^{\mathrm{T}}+\sum_{(\boldsymbol{x}_j,\lambda_j)\in D_l\wedge\lambda_j=i}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^{\mathrm{T}}\Big),\tag{6}$$

$$\alpha_i=\frac{1}{n}\Big(\sum_{\boldsymbol{x}_j\in D_u}\varphi_{ji}+l_i\Big),\tag{7}$$

式中,l_i 为属于第 i 个高斯混合分布的已标记样本个数.

2　基于逆模拟退火的半监督高斯混合模型

确定性退火 EM[14](Deterministic Annealing EM,DAEM)算法是一种改善算法易收敛于局部最优问题的重要技术. DAEM 算法虽然能够有效改善 EM 算法易陷入局部最优的缺陷,但是对于存在类不均衡或类间重叠度大的数据集合,DAEM 算法的参数估计比较差且收敛速度很慢. 为了解决此问题,文献[13]提出了 Anti-annealing EM 算法. 该算法的温度参数 β 首先由 $\beta_{\min}\approx 0$ 缓慢地增长到 $\beta_{\max}>1$,再从 $\beta_{\max}>1$ 缓慢地降低至 $\beta=1$. Anti-annealing EM 算法只针对传统 GMM 无监督聚类的类不均衡、类间重叠带来的性能降低问题,没有考虑半监督聚类问题. 为此,本文提出一种基于逆模拟退火算法的半监督高斯混合模型 EM(Anti-annealing Semi-supervised GMM EM,ASGMM-EM)算法.

ASGMM-EM 算法在执行 E 步时计算无标记样本 $\boldsymbol{x}_j$ 属于类 i 的后验概率为

$$\varphi_{ji}=\frac{(\alpha_i\cdot\mathcal{N}(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i))^{\beta}}{\sum_{l=1}^{k}(\alpha_l\cdot\mathcal{N}(\boldsymbol{x}_j\mid\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l))^{\beta}},\tag{8}$$

式中,β 为逆温度参数. M 步时估计模型参数 α_i、$\boldsymbol{\mu}_i$、$\boldsymbol{\Sigma}_i$,计算公式与(5)-(7)相同. ASGMM-EM 算法描述如图 1 所示.

该算法的参数初始化采用"K 均值"[15-16](K-means)算法思想,相似度采用欧氏距离度量. 其基本思想为:对于未标记样本 $\boldsymbol{x}_j\in D_u$,计算 $\boldsymbol{x}_j$ 到第 i 类已标记样本的平均欧式距离 d_{ji},定义样本 $\boldsymbol{x}_j$ 的初始类标记为

$$\varphi_j^{(0)}=\underset{i\in\{1,2,\cdots,k\}}{\operatorname{argmin}}\ d_{ji}.\tag{9}$$

通过式(9)可给所有未标记样本进行初始化类标记. 然后计算各类的初始均值向量 $\boldsymbol{\mu}_i^{(0)}$、初始协方差矩阵 $\boldsymbol{\Sigma}_i^{(0)}$ 和初始混合系数 $\alpha_i^{(0)}$. ASGMM-EM 算法采用文献[13]中的停止条件,即:

$$\frac{|L(\boldsymbol{\Theta}^k)-L(\boldsymbol{\Theta}^{k-1})|}{|L(\boldsymbol{\Theta}^k)|}<\tau,\tag{10}$$

式中,$\boldsymbol{\Theta}^k$ 为第 k 次迭代后的模型参数,$L(\boldsymbol{\Theta}^k)$ 为 $\boldsymbol{\Theta}^k$ 的最大似然估计值,τ 为阈值,根据需要提前设定. 当 $|L(\boldsymbol{\Theta}^k)-L(\boldsymbol{\Theta}^{k-1})|$ 与 $|L(\boldsymbol{\Theta}^k)|$ 的比值小于阈值 τ 时,算法终止.

ASGMM-EM 复杂度为 $O(mI_{\text{avg}}(uk+(l+u)k))$,其中 m 为逆温度参数个数,I_{avg} 为所有逆温度参数下 EM 算法的最大迭代次数,u 为未标记样本个数,l 为标记样本个数. 与半监督 EM 算法 SGMM-EM 复杂度 $O(I(uk+(l+u)k))$ 相比,复杂度增加了一个数量级 m,但平均每个逆温度参数下的迭代次数的平均值 I_{avg} 比 I 小. 因为随着逆温度参数的变化,ASGMM-EM 算法越接近收敛值,相应的逆温度参数下 SGMM-EM 收敛速度越快,迭代次数越少. 因此,ASGMM-EM 算法的复杂度与半监督高斯混合模型的参数估计算法 SGMM-EM 的复杂度相当.

输入:已标记样本集 $D_l=\{(\boldsymbol{x}_1,\lambda_1),(\boldsymbol{x}_2,\lambda_2),\cdots,(\boldsymbol{x}_l,\lambda_l)\}$;未标记样本集 $D_u=\{\boldsymbol{x}_{l+1},\boldsymbol{x}_{l+2},\cdots,\boldsymbol{x}_{l+u}\}$;高斯混合分布个数 k;
输出:类划分 $C=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k\}$($\boldsymbol{C}_i$ 表示第 i 个类的样本集合,$i=1,2,\cdots,k$);

```
统计有标记样本属于各类的个数 l_i 并将有标记样本分类;
初始化参数 Θ=[α_i,μ_i,Σ_i](i∈{1,2,…,k})和逆温度表 T={β_1,β_2,…,β_m};
for l=1 to m
    while
        E 步:根据当前模型参数 Θ 利用式(8)计算各未标记样本的后验概率 φ_ji;
        M 步:利用未标记样本当前的后验概率 φ_ji 和已标记样本个数 l_i 更新模型参数 Θ;
        if(停止条件)
            break;
        end if
    end while
end for
根据未标记样本最终的后验概率 φ_ji 对未标记样本进行类划分 C={C_1,C_2,…,C_k};
```

图 1　ASGMM-EM 算法

Fig. 1　ASGMM-EM algorithm

3　实验结果分析

本文实验运行的环境为 Intel(R)Core(TM)i7-3770 CPU @ 3.40 GHz,8.00 GB 内存,Windows 8 64 位操作系统. 实验平台采用 Matlab 7.0. 进行比较的 4 种算法分别为:算法 1 为无监督高斯混合模型 EM 算法,在此简记为 GMM-EM;算法 2 为结合逆模拟退火算法的无监督高斯混合模型 EM 算法,在此简记为 AGMM-EM;算法 3 为半监督高斯混合模型 EM 算法,在此简记为 SGMM-EM;算法 4 为结合逆模拟退火的半监督高斯混合模型 EM 算法,在此简记为 ASGMM-EM.

3.1　实验评价标准

本文实验评价标准采用对称 KL 散度[13](Symmetric Kullback-Leibler Divergence,SKLD)和归一化互信息[17](Normalized Mutual Information,NMI).

SKLD 是比较算法的模型估计参数与模型的真实参数接近程度的一种度量方式,此值越小说明算法估计的模型参数与真实参数越接近,即算法性能越好. SKLD 值利用如下公式计算.

$$\mathrm{SKLD}(\boldsymbol{\Theta}_e,\boldsymbol{\Theta}_t)=\sum_{j=1}^{k}D_s[(\boldsymbol{\mu}_{ej},\boldsymbol{\Sigma}_{ej}),(\boldsymbol{\mu}_{tj},\boldsymbol{\Sigma}_{tj})]=$$

$$\sum_{j=1}^{k}\left\{\frac{1}{2}\mathrm{Tr}[\boldsymbol{\Sigma}_{ej}^{-1}\boldsymbol{\Sigma}_{tj}+\boldsymbol{\Sigma}_{tj}^{-1}\boldsymbol{\Sigma}_{ej}]+\frac{1}{2}(\boldsymbol{\mu}_{ej}-\boldsymbol{\mu}_{tj})^{\mathrm{T}}(\boldsymbol{\Sigma}_{ej}^{-1}+\boldsymbol{\Sigma}_{tj}^{-1}](\boldsymbol{\mu}_{ej}-\boldsymbol{\mu}_{tj})\right\}. \tag{11}$$

NMI 评价指标主要反应算法聚类效果的准确度,NMI 值为 0 至 1 之间的实数,NMI 值越接近 1 表示算法聚类准确度越高. NMI 值计算公式为

$$\mathrm{NMI}=\frac{H(A)+H(B)}{H(A,B)}, \tag{12}$$

式中,

$$H(A)=-\sum_{a}P_A(a)\log P_A(a), \tag{13}$$

$$H(B)=-\sum_{b}P_B(b)\log P_B(b), \tag{14}$$

$$H(A,B)=-\sum_{a,b}P_{AB}(a,b)\log P_{AB}(a,b), \tag{15}$$

式中,$P_A(a)$、$P_B(b)$表示 A、B 的概率分布,$P_{AB}(a,b)$表示 A 和 B 的联合概率分布.

3.2　实验结果与分析

实验一共分两组进行. 第一组实验在人工数据上进行测试比较,第二组实验在 UCI 数据集上进行测试比较.

3.2.1　人工数据测试

本组实验在人工数据集对 4 种算法进行测试比较. 分别测试 4 种算法的 SKLD 值、NMI 值和运行时间随着迭代次数增加的变化趋势. 其中 SGMM-EM 和 ASGMM-EM 算法中有标记样本比例设置为 0.25%. 另外,测试比较 ASGMM-EM 和 SGMM-EM 算法的 SKLD 值和 NMI 值随着先验比例增加的变化趋势. 人工数据集参数设置如表 1 所示.

表 1　人工数据参数设置

Table 1　Parameter Setting of Synthetic Data

	Gaussian 1	Gaussian2
α	0.999	0.001
μ	(-6,0)	(6,0)
Σ	[6 0;0 6]	[6 0;0 6]
样本个数	20000	20

4 种算法均在各个迭代次数上均运行 10 次求其平均值,从而得到 4 种算法随迭代次数增加的平均 SKLD 值、平均 NMI 值和平均运行时间变化趋势如图 2、3、4 所示. ASGMM-EM 与 SGMM-EM 算法随基于节点标记的先验比例增加的平均 SKLD 值变化趋势和平均 NMI 变化趋势如图 5、6 所示.

由图 2 可看出 ASGMM-EM 算法要比其他 3 种算法的平均 SKLD 值要高并且收敛速度都要快很多. 其中,AGMM-EM 算法收敛速度较快,ASGMM-EM 算法收敛速度比 AGMM-EM 算法稍快,SGMM-EM 算法收敛速度最快,但是算法收敛呈现出强波动性,这是由于 EM 算法容易陷入局部收敛造成的. 而 ASGMM-EM 算法在逆模拟退火算法的控制下,可有效缓解这种波动性,能够相对快速稳定地收敛. 上述分析同样可由图 3 中得到,ASGMM-EM 算法相比于其他 3 种算法能够快速并且稳定地收敛. 由图 4 可看出 ASGMM-EM

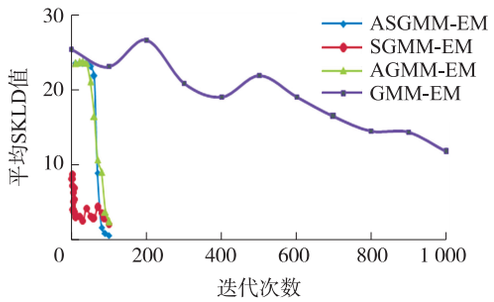

图 2　平均 SKLD 值变化趋势

Fig. 2　The variation of the average SKLD value

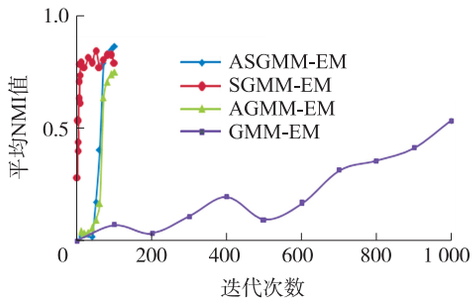

图 3　平均 NMI 值变化趋势

Fig. 3　The variation of the average NMI value

算法随着迭代次数的增加平均运行时间均稍高于其他 3 种算法,这是由于 ASGMM-EM 的时间复杂度均高于其他 3 种算法. 其中,比 SGMM-EM 增加了 1 个数量级 m(m 为逆温度参数的个数),但 ASGMM-EM 算法随着逆温度参数的变化,模型参数估计越趋近于真实值,迭代次数越来越少,故实际运行时间增加不到 1 个数量级 m. 虽然 ASGMM-EM 和 AGMM-EM 都在逆温度表的控制下,但是 ASGMM-EM 算法是半监督聚类算法,参数初始化和模型参数更新都需要利用已标记样本作为先验信息,故实际运行时间 ASGMM-EM 比 AGMM-EM 算法稍高.

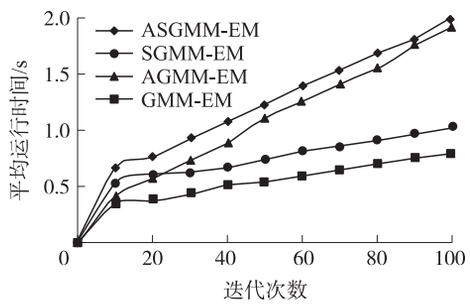

图 4　平均运行时间变化趋势

Fig. 4　The variation of the average run time

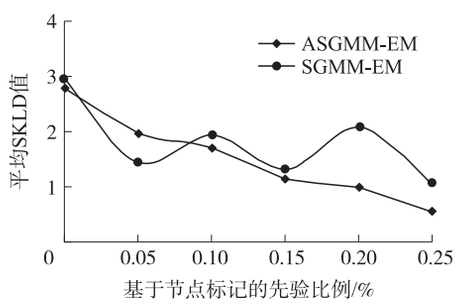

图 5　平均 SKLD 值变化趋势

Fig. 5　The variation of the average SKLD value

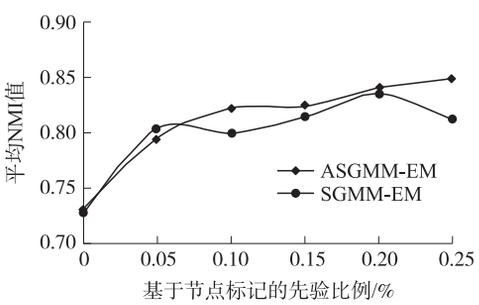

图 6　平均 NMI 值变化趋势

Fig. 6　The variation of the average NMI value

由图 5 可看出 ASGMM-EM 算法随着基于节点标记的先验比例增加平均 SKLD 值逐渐下降,说明随着先验比例的增加模型参数估计准确度逐渐提高. SGMM-EM 算法虽然随着先验比例的增加平均 SKLD 值整体呈下降趋势,但是呈现出强波动性. 这仍是由于传统 EM 算法易于陷于局部最优这一缺陷,而 ASGMM-EM 算法在逆模拟退火算法的控制下可有效缓解这一缺陷,使其更大概率地跳出局部最优从而找到更优的模型参数. 上述分析同样可由图 6 得到,ASGMM-EM 相比于 SGMM-EM 能够相对稳定地收敛到更优的模型参数使平均 NMI 值有所提高.

3.2.2　UCI 数据集测试

本组实验分别在 5 个 UCI 数据集上测试 4 种算法性能. UCI 数据集的基本信息如表 2 所示.

表 2　UCI 数据集

Table 2　UCI Dataset

数据集	n	d	k
Iris	150	4	3
ecoli	336	7	8
glass	214	9	6
seeds	210	7	3
wine	178	13	3

表 2 中,n 表示数据集样本个数,d 表示样本维数,k 表示数据集类别数. 由于无法得到 UCI 数据集的真实分布参数,所以本组实验不采用 SKLD 评价指标. 所以,分别测试 4 种算法在 5 种不同 UCI 数据集上随着迭代次数的增加平均 NMI 值的变化趋势. 其中,ASGMM-E 和 SGMM-EM 算法中有标记样本比例设置为 10%. 实验结果如图 7-11 所示.

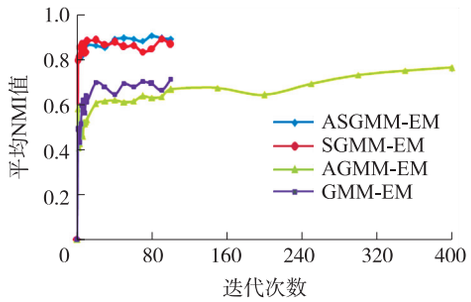

图 7　在 Iris 数据集上平均 NMI 值变化趋势

Fig. 7　The variation of the average NMI value on Iris

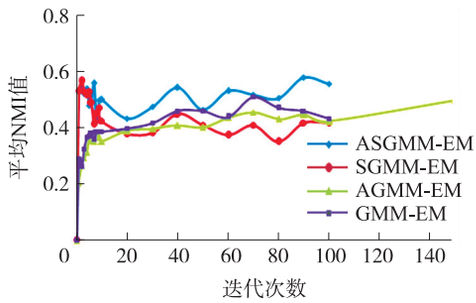

图 8　在 ecoli 数据集上平均 NMI 值变化趋势

Fig. 8　The variation of the average NMI value on ecoli

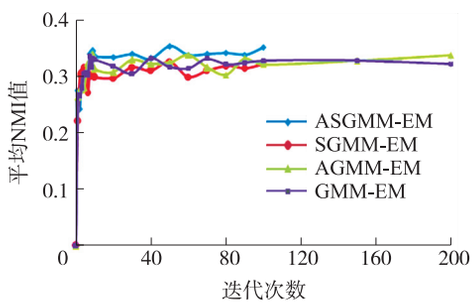

图 9　在 glass 数据集上平均 NMI 值变化趋势

Fig. 9　The variation of the average NMI value on glass

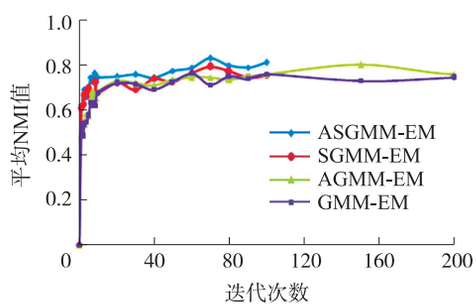

图 10　在 seeds 数据集上平均 NMI 值变化趋势

Fig. 10　The variation of the average NMI value on seeds

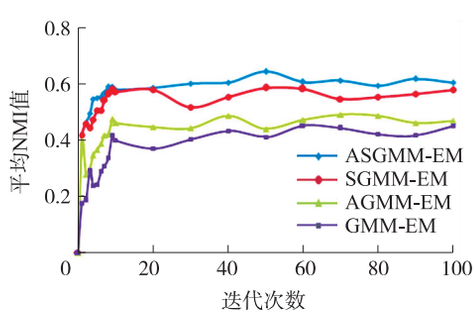

图 11　在 wine 数据集上平均 NMI 值变化趋势

Fig. 11　The variation of the average NMI value on wine

由图 7 可看出,ASGMM-EM 算法比 SGMM-EM 算法随着迭代次数增加的平均 NMI 值稍好,而 AGMM-EM 算法和 GMM-EM 算法在 Iris 数据集上性能表现较差,说明 ASGMM-EM 算法在处理这种类平衡的数据集时仍具有较好的聚类性能. 由图 8 和图 9 可看出,当数据集的分类数和样本维度都增加时,ASGMM-EM 算法的平均 NMI 值优于其他 3 种算法. 由图 10 和图 11 可看出,在数据集分类数不变时,随着样本维度增加,ASGMM-EM 算法的平均 NMI 值优于其他 3 种算法. 因此,ASGMM-EM 算法在处理这种类相对平衡的 UCI 数据集时,算法性能仍优于其他 3 种算法. 这是因为在算法初始迭代阶段,逆模拟退火算法温度表加大了搜索步长,并且随着迭代的进行使搜索步长逐渐减小,这样不仅可加快算法收敛速度,同时可以更大概率跳出局部最优解,从而在最优解的搜索上表现出相对稳定性. 此外,ASGMM-EM 算法引入了基于节点标记的先验信息,不仅使模型参数初始值更好,而且使模型参数在一定程度上朝着最优值方向更新.

4　结语

在 ASGMM-EM 中,约束信息基于节点标记,由用户先验提供. 针对不同的数据集需要适当调整温度表以适应不用的数据分布. 由于 ASGMM-EM 算法在估计模型参数时引入了基于节点标记的先验信息,因此,相比传统高斯混合模型 EM 算法收敛速度更快,准确率更高. 与此同时,又将逆模拟退火算法与 EM 算法相结合,使高斯混合模型能够适应一些类不均衡、类间重叠度高的数据分布情况,使模型更加稳定. 下一步工作是在 ASGMM-EM 算法的基础上利用 Hadoop 平台研究分布式算法,使高斯混合模型具有处理大规模数据集能力,同时进一步提高高斯混合模型的聚类速度及准确率.

[参考文献]

[1]　YEUNG K Y, YEUNG K Y, HAYNOR D R, et al. Validating clustering for gene expression data[J]. Bioinformatics, 2001,

17:309-318.
[2] YANG Y,XU D,NIE F,et al. Image clustering using local discriminant models and global integration[J]. IEEE Trans Image Process,2010,19:2 761-2 773.
[3] XU W, LIU X, GONG Y. Document clustering based on non-negative matrix factorization[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(ACM). Nuray,2003:267-273.
[4] CHAPELLE O,SCHÖCKOPF B,ZIEN A. Semi-Supervised Learning[C]//Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Springer-Verlag,2006:588-595.
[5] YANG M S,LAI C Y. A robust EM clustering algorithm for Gaussian mixture models[J]. Patter recognition,2012,45(10):3 950-3 961.
[6] PORTELA N M,CAVALCANTI G D C,REN T I. Semi-supervised clustering for MR brain image segmentation[J]. Expert systems with applications,2014,41(4):1 492-1 497.
[7] 於跃成,生佳根,邹晓华. 基于约束正则化的生成聚类分析[J]. 系统工程与电子设计,2014,36(4):777-783.
[8] HE X F,CAI D,SHAO Y L,et al. Laplacian regularized Gaussian mixture model for data clustering[J]. IEEE Trans on Knowledge and Data Engineering,2011,23(9):1 406-1 418.
[9] GAN H T,SANG N,HUANG R. Manifold regularized semi-supervised Gaussian mixture model[J]. Journal of the optical society of America. A,optics and image science,2015,32(4):566-575.
[10] LIU J,CAI D,HE X. Gaussian mixture model with local consistency[C]//The 24th AAAI Conference on Artificial Intelligence(AAAI). Atlanta,USA,2010:512-517.
[11] MARTINEZ U A,PLA F,SOTOCA J M. A semi-supervised Gaussian mixture model for image segmentation[C]//Proc of the 20th International Conference on Pattern Recognition. Istanbul Turkey,2010:2 941-2 944.
[12] 周志华. 机器学习[M]. 北京:清华大学出版社,2015:293-295.
[13] IFTEKHAR N,DANIEL G. Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients[C]//ICML,Edinburgh,Scotland,2012.
[14] UEDA N,NAKANO R. Deterministic annealing EM algorithm[J]. Neural networks,1998,11(2):271-282.
[15] STEINHAUS H. Sur la division des corpmaterielsen parties[J]. Bulletin of Acad Polon Sci,IV(C1. Ⅲ),1956:801-804.
[16] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]//Fifth Berkeley Symposium on Mathematics,Statistics and Probability. California,USA, 1967:281-297.
[17] ASUNCION A, NEWMAN D. UCI machine learning repository[EB/OL]. [2014-02-18]. http://www.ics.uci.edu/~mlearn/MLRepository.html. 2007.

[责任编辑:顾晓天]