2020 年 2 月 第 2 期
Vol. 41 No. 2 2020

一种半监督机器学习的 EM 算法改进方法

夏筱筠[1],张笑东[1,2],王 帅[1,2],罗金鸣[3],崔露露[1,2],赵智阳[1,2]

[1](中国科学院 沈阳计算技术研究所,沈阳 110168)
[2](中国科学院大学,北京 100049)
[3](沈阳工程学院,沈阳 110136)
E-mail: sict_zxd@163.com

摘 要: EM(Expectation Maximization)算法是含有隐变量(latent variable)的概率参数模型最大似然估计、极大后验概率估计最有效的算法,但很容易进入局部最优现象,对此提出基于半监督机器学习机制的 EM 算法. 本文方法是在最大似然函数中加入惩罚最小二乘因子,同时引入非负约束作为先验信息,结合半监督机器学习方法,将 EM 算法改进转化为最小化求解问题,再采用最大似然方法求解 EM 模型,有效估计了混合矩阵和高斯混合模型参数,实现 EM 算法的改进. 仿真结果表明,该方法能够很好地解决了 EM 算法容易局部最优化问题.

关 键 词: 半监督机器学习; EM 算法; 改进分析; 局部最优
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2020)02-0230-06

Improved EM Algorithm of Semi-supervised Machine Learning

XIA Xiao-jun[1], ZHANG Xiao-dong[1,2], WANG Shuai[1,2], LUO Jin-ming[3], CUI Lu-lu[1,2], ZHAO Zhi-yang[1,2]
[1](Shenyang Institute of Computing Technology, Chinese Academy of Science, Shenyang 110168, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Shenyang Institute of Engineering, Shenyang 110136, China)

Abstract: EM algorithm is the most effective algorithm for maximum likelihood estimation and maximum posteriori probability estimation of probability parameter model with latent variable, but it is easy to enter the local optimum phenomenon. An EM algorithm based on semi-supervised machine learning is proposed. The least squares penalty term is added to the maximum likelihood function, non-negative constraints are introduced as a priori information, and combined with semi-suervised machine learning method, the improved EM algorithm is transformed into a minimization problem. The maximum likelihood algorithm is used to solve the EM model, effectively estimating the parameters of the mixed matrix and the Gaussian mixed model, and the improvement of the EM algorithm is realized. The simulation results show that this method can solve the problem of EM algorithm easily entering local optimum well.
Key words: semi-supervised machine learning; EM algorithm; improvement analysis; local optimum

1 引 言

期望最大化(Expectation Maximization, EM)算法是一种被广泛应用于极大似然估计的迭代计算方法,用于处理大规模数据不完整问题.

EM 算法实际上是一种基于梯度上升的方法,在应用于模型参数估计时能够保证迭代之后的似然函数是递增的,因此有一个很大的弊端就是通常只能得到一个局部最优解. 对于 EM 算法,国内外学者提出了很多改进方式,比较经典且应用比较广泛的有 PX-EM 算法、ECME 算法、MCEM 算法. PX-EM 算法利用协方差对 M 步的计算进行修正,通过获取数据的额外信息达到加快收敛速度的目标,在应用上更加简化,而且未破坏 EM 算法单调收敛性质. 但 PX-EM 算法那也有其缺点,在扩充参数模型上很难找到一个通用的标准,所以它在局部最优解问题上并没有得到很好的优化[1]; ECME 算法对 E 步进行了简化处理,由于它在极大化过程中针对的一直是实际对象,而不是实际对象的附近量,其拥有的收敛速度比 EM 和 PX-EM 算法都要大. 然而,ECME 算法在局部收敛的问题上依然没有给出有效的处理方法,依然延续了 EM 算法所固有的性质[2]; MCEM 算法用蒙特卡罗的方法对 EM 算法进行优化,对 EM 算法中的期望显示表示进行了改善,从而加速了收敛速度,但由于这种方法过于灵活,在模拟容量选择和收敛性准则的确认上很难,而且失去了 EM 算法的单调性,难以对其收敛特性进行估计,依然在局部最优问题上没有给出有效解决方案[3].

综上所述,EM 算法及其改进算法存在在局部最优化问题上还存在着很大的不足,国内外学者在改进方法上侧重于解决其收敛速度问题,针对 EM 算法进入局部最优问题还未

收稿日期: 2019-07-03 收修改稿日期: 2019-08-27 基金项目: 国家科技重大专项课题项目(2017ZX04011004)资助; 国家自然科学基金项目(61803271)资助. 作者简介: 夏筱筠,男,1967 年生,博士,研究员,博士生导师,研究方向为先进制造与自动化、智能化; 张笑东,男,1989 年生,博士研究生,研究方向为大数据与智能计算; 王 帅,男,1990 年生,博士研究生,研究方向为机器学习、机器人视觉; 罗金鸣,男,1983 年生,博士,讲师,研究方向为智能控制、切换控制; 崔露露,女,1994 年生,硕士研究生,研究方向为先进制造与自动化; 赵智阳,男,1994 年生,硕士研究生,研究方向为先进制造与自动化.

提出有效方法. 对此,本文将半监督机器学习方法用于 EM 算法改进,首先在最大似然函数中加入最小二乘惩罚项,引入非负约束作为先验信息,然后结合半监督机器学习法,将 EM 算法改进方法转化为处理最小化的问题,再采用最大似然算法求解 EM 问题,有效地求解了混合矩阵和高斯混合模型参数,实现 EM 算法的改进. 通过实验理论与实验结果分析表明改进后 EM 算法不会陷入局部最优,得到的结果更加可靠.

2　基于半监督机器学习的 EM 算法改进分析

2.1　改进 EM 算法先验信息的获取

以先验理论为基础依据概率惩罚定理以及二项分布概率函数[4],将最大似然函数表示为:

$$L(\beta) = [\alpha_{0i}^{y0i}, \alpha_{1i}^{y1i}, \alpha_{2i}^{y2i}, \cdots, \alpha_{ji}^{yji}] = \prod_{i=1}^{n}\{\prod_{j=0}^{j}\alpha_{ji}^{yji}\} \tag{1}$$

公式(1)中,$\alpha_{ji} = \alpha_j(x_i) = P(y = j \mid x_i)$,$y_{ji}$是 $j \times 1$ 的某一个一维向量,代表第 i 个观测变量,$x_i = (x_1, x_2, \cdots, x_p)$用于描述协变量.

通过计算,L 对数似然函数课表示成:

$$l(\beta) = \sum_{j=0}^{j}\sum_{i=1}^{n} y_{ji}\ln[\alpha_{ji}] \tag{2}$$

对公式(2)关于 $j \times p$ 进行一阶偏导求解,那么得分函数表示为:

$$S(\beta) = \sum_{i=1}^{n} x_{pi}(y_{ji} - \alpha_{ji}) \tag{3}$$

对 $S(\beta) = 0$ 进行求解,从而获取参数估计值 β.

将惩罚项添加到最大似然函数中会大大减小得分函数所带来的误差[5],本文添加的惩罚项为最小二乘惩罚项,公式为:

$$f_z^{k+1} = \frac{\operatorname{argmin}(p - Hf^k)^T(p - Hf^k)\chi Q(f)}{\beta} \tag{4}$$

公式(4)中,$p = [p_1, p_2, \cdots, p_i]^T$ 代表估计向量;$f = [f_1, f_2, \cdots, f_j]^T$ 代表惩罚向量;$Q(f)$ 为惩罚项;χ 代表对惩罚项的影响程度,χ 值可利用实验获取;H 为 $i \times j$ 系统概率矩阵;k 代表迭代次数.

利用最优化法对上述惩罚项进行求解,并且通过 OSL 算法获取迭代公式:

$$f_{zj}^{k+1} = f_j^k \frac{\sum H_{ij}p_i\beta}{\sum\{H_{ij}\sum H_{ij}f_j\} + \chi Q(f)} \tag{5}$$

需要指出,$Q(f)$ 主要包含二次与非二次,二次惩罚项结构简单,通常会导致数据丢失,而非二次惩罚项可避免该弊端[6,7],本文选用非二次惩罚项.

$$Q(f) = \chi\frac{f_{zj} - M}{\beta M} \tag{6}$$

式中,$M = Med(f_{zj}, j' \in N_{zj})$,$N_{zj}$代表 f_{zj} 的邻域.

将非二次惩罚因子加到最大似然函数中后,那么含有多个参数的估计误差为:

$$b(\beta) = \frac{b_1(\beta)}{n} + \frac{b_2(\beta)}{n} + \cdots \tag{7}$$

移除得分函数中的误差首项,对得分函数求解,即,$S(\beta_j)^* = 0$ 再次获取参数估计值 $\hat{\beta}_j$,以降低估计误差.

事实上,就是在似然函数添加惩罚最小二乘项 $|Q(\beta)|^{0.5}$,似然函数可描述成:

$$L(\beta)^* = L(\beta)|Q(\beta)|^{0.5} \tag{8}$$

对数似然函数可描述成:

$$l(\beta_j)^* = S(\beta)0.5\log|Q(\beta)| \tag{9}$$

得分函数的基本形式可描述成:

$$S(\beta_j)^* = S(\beta) - Ab(\beta)/n \tag{10}$$

其中,

$$b(\beta)/n = (X^TWX)^{-1}X^TW\varepsilon \tag{11}$$

$$A = X^TWX \tag{12}$$

利用得分函数与添加惩罚项之后的最大似然函数,得到先验信息如下:

$$S(\beta)^* = S(\beta) - X^TW\varepsilon \tag{13}$$

2.2　EM 算法改进的实现

在上述处理最大似然函数的基础上,本节首先对 EM 算法实现过程进行分析,然后引入非负约束先验信息,采用半监督机器学习的方式对 EM 算法实现改进.

EM 算法指的是观测数据是否含隐含变量条件下,也就是所用观测数据是非完整数据的情况下,使用极大似然函数对这些数据进行处理,经过多次循环,最终获得最优值的算法[9,10].

在传统 EM 算法中,通常会使用最大似然准则完成对参数估量. 假设存在 N 个数据 $X = [x_1, x_2, \cdots, x_N]$,这些数据是通过特定分布 $p(X \mid \theta)$独立采样获取[11],结合先验知识,则似然函数可描述成:

$$p(X|\theta) = \prod_{i=1}^{N} p(x_i \mid \theta) = L(\theta \mid X) \tag{14}$$

最大似然准则为找到符合以下条件的模型参数:

$$\theta^* = \arg\max L(\theta|X) \tag{15}$$

为了简化计算过程,使用函数 $\log(L(\theta|X))$ 进行最终解的优化,其中,$\log(L(\theta|X))$ 为对数似然. 如果 Z 代表一个数据集,其中有两组数据,一组是已观测数据,另一组是未观测数据,用 X 代表已观测,用 Y 代表未观测,则 $Z = (X, Y)$ 一般被称作完全数据集与缺失数据集,则有:

$$p(z|\theta) = P(x, y|\theta) = p(y|x, \theta)\,p(x|\theta) \tag{16}$$

假定似然函数可表示成:

$$L(\theta|Z) = p(X, Y|\theta) \tag{17}$$

完整数据的对数似然函数一般是在 E 步骤中进行的,这里 X 为已知量,那么针对未知变量 Y 的期望[12] 用 R 函数表示:

$$\begin{aligned} R(\theta, \theta'(t)) &\equiv E[\log p(X, Y|\theta) \mid X, \theta'(t)] \\ &= \int \log p(X, y \mid \theta) f(y \mid X, \theta'(t))\, dy \end{aligned} \tag{18}$$

在 M-step 中,通过下式对模型参数进行更新处理:

$$\theta'(t+1) = \arg\max_{\theta} R(\theta, \theta'(t)) \tag{19}$$

EM 算法通常不断对 E 步与 M 步进行循环处理,直至参数达到收敛. 它在理想状态下可获得局部最优化值[13].

EM 能够适应不同的问题环境,为了充分考虑不同应用场景,统一表示成为:

$$y = Cx \tag{20}$$

其中,矢量 $y \in R^{m\times 1}$ 代表测量数据,矢量 $x \in R^{n\times 1}$ 代表观测对象密度分布估计值,矩阵 $C \in R^{m\times n}$ 代表映射矩阵.

考虑到大部分物理过程均为衰减过程,x 与 y 均为非负

矢量,然而矢量 δU 与 $\delta\sigma$ 无法保证非负约束,为此,引入先验信息对 δU 进行计算:

$$\delta U = (-1)^{q}(U_{\text{mea}} - U_{\text{ref}}) \tag{21}$$

其中,U_{ref} 用于描述理论测量值,U_{mea} 用于描述实际测量值. q 值主要取决于先验信息(1.1 节得到),也就是如果 U_{ref} 和 U_{mea} 相比,U_{ref} 为较低的测量值,那么 $q=2$;否则,$q=1$.

在上述非负约束先验信息下,将半监督机器学算法思想应用于 EM 算法的优化之中.

通常来讲,如图 1 所示,在 EM 算法的每次循环中.

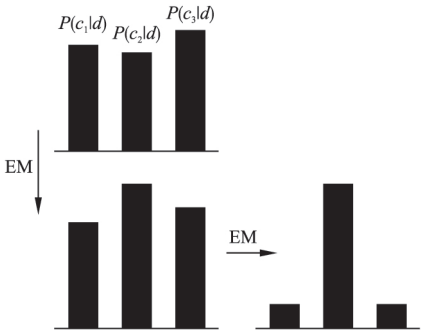

图 1 EM 算法迭代过程示意图

Fig. 1 Schematic diagram of iteration process of EM algorithm

设定没有标注样本 d_i 按照权重 $P(c_j|d_i;\theta')$ 分配,也就是在确定 c^* 值的情况下,下一轮计算 d_i 也仅可用于占其 $P(c^*|d_i;\theta')$ 部分参与和 c^* 的训练,并且需考虑其和其余类别相关 $P(c_j|d_i;\theta')$ 的可能性. EM 算法不判断为标定样本量,从而导致陷入局部最优,且收敛速度变慢. 在 EM 算法中,$P(c_j|d_i;\theta')$ 一直非零,所以本和其它种类相关的未标注的信息也在很大程度上会以 $P(c_j|d_i;\theta')$ 的部分对其余训练产生干扰,导致陷入局部最优.

半监督机器学习方法在每次训练时,可把确定结果引入到当前标注样本集,使用已经得到的结果训练,达到较好的训练效果,这是一种自训练的方式,大大避免了局部最优化问题.

本节把这种半监督机器学习方法用于 EM 改进中,在每次迭代,按照 E 步骤所得结果标注确定的样本数据,M 步骤训练新训练集下样本[14,15]. 使得每次迭代未标注样本集 U 逐步减小,加大迭代速度,并且防止干扰. 所以针对当前刚被添加至标注样本集的文本 $d_i \in L$,如图 2 所示,依据训练结果直接设置 $P(c_j|d_i;\theta') \in \{0,1\}$,其余信息不会对其它训练产生干扰.

本文给出半监督机器学习改进 EM 算法的方法,如下:

1) 假设 $t=0$.

2) 定义 $\theta^{(0)} = \arg\max_{\theta} P(\theta|L)$.

3) 当 $U! = \varnothing$ 时,执行 E 步骤,假设 $z^{(t+1)} = E[z|D;\theta^{(t)}]$,则有:

$$P(i^*, j^*) = \arg\max_{(i,j)}\{z_{ij}|d_i \in U\} \tag{22}$$

令 $L = L \cup \{(d_{i^*}, l_{j^*})\}$,$U = U \setminus \{d_{i^*}\}$,针对 $j = 1 \sim |C|$,令 $z_{i^* j} = 0$,否则 $z_{i^* j^*} = 1$

4) 那么 M 步有,$\theta^{(t+1)} = \arg\max_{\theta} P(\theta|D;z^{(t+1)})$.

5) 令 $t = t+1$,循环进行上述步骤,输出 $\theta^{(t)}$.

本节使用半监督机器学习的方式,在每轮 E 步中都会选择 z_j 中最大的 $z_{i^* j^*}$,同时把 d_{i^*} 从 U 中删除,把 (d_{i^*}, l_{j^*}) 添加至集合 L,并且将 $z_{i^* j}(j \neq j^*)$ 置 0,将 $z_{i^* j^*}$ 置 1. $M-step$ 在训练集上对 θ' 进行估计. 事实上,每轮迭代不是只选择一个没有标注的样本进行标注,可依据差异原则,标注若干可标注样本,实现标注效率的大幅度提升.

半监督机器学习方法在每轮迭代过程中会将完全确定的未标注样本添加标签,放入标注样本集合,为后续训练提供高质量标注样本,大大减少循环次数,从而避免了局部最优化问题的发生,提高训练性能.

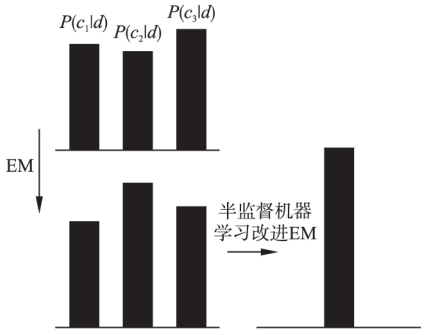

图 2 改进 EM 过程示意图

Fig. 2 Modified EM process diagram

在上述分析的基础上,将最大似然方法引入,利用其对 EM 模型进行求解计算,构造下述似然函数:

$$L(\delta\theta) = \sum_{i=1}^{m}[\delta U_i \log\lambda_i - \lambda_i] \tag{23}$$

上式中,$\lambda_i = (J\delta\theta)_i$ 代表期望值. 为了提高收敛速度,在似然函数中加入惩罚最小二乘,引入非负约束当成先验信息,转换成最小化问题:

$$\min W(\delta\theta) = \sum_{i=1}^{n}\{[J\delta\theta)_i + \omega^2] - (\delta\overline{U_i} + \omega^2) \times \log[J\delta\theta)_i + \omega^2]\} + \gamma \| L\delta\theta \|_2^2 \tag{24}$$

式中,$\delta\theta \geqslant 0$,$\overline{U_i} = \max\{U_i, 0\}$,$\gamma$ 用于描述非负正则化加权因子,参数 $\delta\theta_i$ 的非负约束和 ω^2 令最小化问题更加稳定. ω^2 代表噪声,L 代表正则化矩阵,$W(\delta\theta)$ 代表凸函数,J 代表系统矩阵,上式有全局最小解.

2.3 改进 EM 算法的应用

为更深入阐述本文优化方法,现使用本文优化后的方法对混合矩阵与高斯混合模型参数进行计算估量.

首先对混合矩阵进行剖析,$x = [x_1, x_2, \cdots, x_d]^T$ 是 d 维随机变量,它概率密度函数可表示成:

$$p(x|\theta) = \sum_{m=1}^{k}\partial_m p(x|\theta_m) \tag{25}$$

则该随机变量符合 k 个成分的有限混合分布,也就是该随机变量分布可通过混合矩阵进行描述. 其中,某一变量先验概率为 ∂_m,同时符合 $\partial_m \geqslant 0$,$\sum_{m=1}^{k}\partial_m = 1$.

用 θ_m 代表某个变量的密度模型中的参数,通过各 $p(x|\theta_m)$,获取差异混合矩阵. 混合矩阵全部参数集合可描述成 $\theta_m = \{(\partial_m, \theta_m)\}$.

混合高斯模型把各点出现的概率当成多个混合模型得到的结果,这是一种聚类的思维[16]. 混合高斯模型将每个模型

经过加权处理之后,可得其概率密度函数[17]:

$$p(X/\Theta) = \sum_{k=1}^{K} \pi_k p(x_i/\mu_k, \sum_k) \tag{26}$$

$$p(x_i/\mu_k, \sum_k) = \frac{1}{\sqrt{(2\pi)^d \det \sum_k}} \exp \left[-\frac{1}{2} (x-\mu_k)^T \sum_k^{-1} (x-\mu_k) \right] \tag{27}$$

其中,π_k 代表权重,代表模型被选中概率,μ_k、$\sum_k$ 分别代表均值和方差.

使用最大似然函数[18]的方式对混合高斯模型进行训练,可表示成为:

$$L(\Theta/X) = \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k p(x_i/\mu_k, \sum_k) \tag{28}$$

针对上式及混合矩阵参数的直接求解非常困难,本节采用改进 EM 算法迭代训练获取.

给出有限标记训练样本集,用 $G^l = \{(X_1, y_1), \cdots, (X_L, y_L)\}$ 进行描述,其中 X_i 代表某样本,该样本是高斯混合模型,y_i 代表对应类标. 在整个训练样本集 G^l 中,共存在 M 个类型,因此有 $y_i \in \{c_1, c_2, \cdots, c_M\}$.

$X = \langle x_1, x_2, \cdots, x_N \rangle$, $x_n \in R^d$. 针对各 S^l 中的 (X_i, y_i),假设其均互相独立同分布,并假设针对未标记样本 X,如果最大后验概率为 $P(c_j|X)$,那么 j 类中包含 X,可通过类条件概率密度函数 $p(X|c_j)$ 与相关法则对后验概率进行求解:

$$P(X|c_j) = \frac{p(X|c_j) P(c_j)}{p(X)} \tag{29}$$

式中,$P(c_j)$ 代表先验概率.

$$p(X) = \sum_{j=1}^{M} p(X|c_j) P(c_j) \tag{30}$$

针对于变量分类条件概率密度,使用混合高斯模型进行表示[19],如公式(31)所示.

$$p(x|c_j) = \sum_{k=1}^{K} \partial_{jk} p(x|\mu_{jk}, \sum_{jk}) \tag{31}$$

某一种混合变量的数量用 K 表示,混合权重用 ∂_{jk} 进行表示,使用 $p(x|\mu_{jk}, \sum_{jk})$ 代表多元混合高斯模型的概率密度.

完整高斯混合模型,也就是 $X = \langle x_1, x_2, \cdots, x_N \rangle$ 的类条件概率密度可写成公式(32)的结构形式.

$$p(X|c_j) = \prod_{n=1}^{N} p(x_n + c_j) = \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \partial_{jk} p(x_n | \mu_{jk}, \sum_{jk}) \right) \tag{32}$$

则可获取后验概率,$P(c_j|X)$,令:

$$j^* = \arg \max_j P(c_j + X) \tag{33}$$

利用上述过程实现样本标记.

$\theta_j = (\partial_{jk}, \mu_{jk}, \sum_{jk})$ 代表混合高斯模型,$\theta_m = \{(\partial_m, \theta_m)\}$ 代表混合矩阵全部参数集合,使用 EM 算法对它们的参数进行估量,在半监督机器学习过程中,如果学习样本集包含两种不同的集合组,即: $G = G^l \cup G^u$,可以为为标记和已标记两个组.

$$G = \{(x_1, y_1), \cdots, (x_l, y_l), (x_{l+1}, y_{l+1}), \cdots, (x_{l+1}, y_{l+u})\} \tag{34}$$

针对各未标记样本 x_i,对描述类别数量的 M 个隐含变量 z_{ij} 进行定义:

$$z_{ij} = \begin{cases} 1, if\ y_i = c_j \\ 0, else \end{cases} \tag{35}$$

θ_j 代表混合矩阵和混合高斯模型参数,使用改进之后的 EM 算法进行求解. 其中,初始迭代参数 θ^0 使用已经做过标注样本进行计算,再通过下述过程进行迭代:

针对 E 步骤: $\hat{z}^{(l+1)} = E[z|S; \theta^{(l)}]$,针对 M 步骤, $\hat{z}^{(l+1)} = E[z|S; \theta^{(l)}]$.

对于未分类的样本,在 E 步骤中使用最大后验概率进行标记,通过公式(36)计算:

$$E[z_{ij}|G; \theta^{(l)}] = \frac{P(x_i|c_j) P(c_j)}{P(x_i)} \tag{36}$$

在 M 步骤中,按照所作标记的新样本使用最大似然对参数进行重新求解,得到新的结果.

基于上述过程,设计改进 EM 算法估计参数流程为:

1) 假设 $t = 0$.

2) 进行初始化处理,针对 M 步骤,令 $\theta^{(0)} = \arg \max_\theta P(G^l|G))$,针对 E 步骤,令 $\hat{z}^{(l+1)} = E[z|G; \theta^{(l)}]$.

3) 针对所有 $i = L+1, \cdots, L+U$,假设 $j^* = \arg\max \hat{z}_{ij}^{(l+1)}$, $z_{ij}^{(l+1)} = \begin{cases} 1, if\ j = j^* \\ 0, else \end{cases}$.

4) 针对 M 步骤,令 $\theta^{l+1} = \arg \max_\theta P(G, z^{(l+1)}|\theta)$.

5) 令 $t = t+1$,对以上步骤循环至收敛.

6) 输出 θ^l,需要注意的是,每一类的混合高斯模型的参数都是通过当前类别中的样本集进行估计的.

3　实验与分析

3.1　实验理论

对于 EM 算的局部最优化问题,主要原因为其初始值的设置不够合理.

在本次实验中,首先对 EM 算法的初始化值进行训练学习,同时,将传统方法作为对照实验组,与所提方法进行对比,并给出对比结果.

衡量算法有效性的另一个参数为算法拟合值,所谓算法拟合效果即算法能否按照原设定进行运算和应用. 本次实验获取文献[4]算法、文献[6]算法以及所提算法的拟合曲线,具体获取方法如下:

将待测数据集 H 划分为两个子集 H_1 与 H_2,其划分规则如下: 将每个待测数据随机的划分到两个子集的其中任意一个中去. 分别在两个观测样本子集 H_1, H_2 上将算法目标函数极大化,得到拟合值,并对多个拟合值进行整合,得到拟合曲线. 利用拟合曲线可以看出算法的拟合效果,拟合的吻合程度越高,则说明实验方法的性能越高.

3.2　实验过程

针对 EM 方法在执行过程中经常会经过很多次学习来筛选出最优的初始值. 表 1 为传统方法和改进之后的方法性能对比.

分析表 1 可知,初始化相同的情况下,改进之后的方法对局部最优化问题起到了明显的抑制作用.

利用复合抽样法形成三阶数据,参数设置成,样本数量是 2000 个. 样本采样序列以及理论曲线分别用图 3 和图 4 进行描述.

在进行实验过程中,本文分别将文献[4]方法和文献[6]

方法作为对比进行参数估计,预设的研究模型的阶数是 3,但本文方法模型阶数是 4,这主要是由于复杂曲线可利用各种组成成分构成.

表 1 传统方法和改进后的方法性能对比

Table 1 Performance comparison between traditional method and improved method

实验次数	算法	迭代次数	迭代时间/s	对数似然函数值
1	传统算法	50	19	-3.52e6
	改进算法	78	26	-3.44e6
2	传统算法	50	19	-3.53e6
	改进算法	43	18	-3.48e6
3	传统算法	83	27	-3.52e6
	改进算法	36	17	-3.48e6

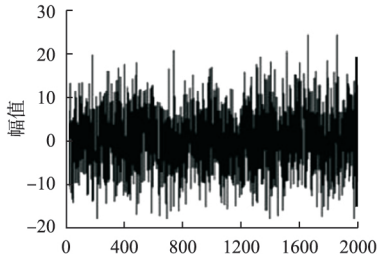

图 3 样本采样序列

Fig. 3 Sample sampling sequence

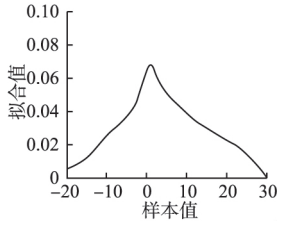

图 4 理论曲线

Fig. 4 Theoretical curve

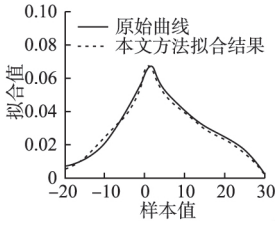

图 5 本文方法拟合结果

Fig. 5 Fitting results of this method

为了将本文方法和文献 [4] 方法、文献 [6] 方法进行比较,也把这两种方法的阶数设置成 6. 本实验利用不同方法拟合曲线和原始曲线相比,替代参数估计值与设定值的比较,从而便于分析.

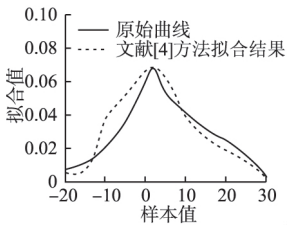

图 6 文献 [4] 方法拟合结果

Fig. 6 Fitting results of document [4]

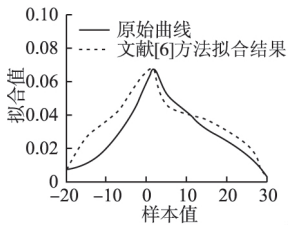

图 7 文献 [6] 方法拟合结果

Fig. 7 Fitting results of document [6]

采用本文方法、文献 [4] 方法、文献 [6] 方法进行拟合,得到的拟合曲线和原始曲线比较结果依次用图 5、图 6 和图 7 进行描述.

本文实验样本数量较为充足,将不同样本初始条件划分成 6 类,针对 6 种初始条件,获取本文方法拟合结果和文献 [4] 方法、文献 [6] 方法相比的性能可通过表 2 进行描述.

表 2 源信号处理前后功率谱峰值及对应频率

Table. 2 Comparison of fitting performance of three methods

初始条件	采用方法	迭代次数/次	迭代时间/s	似然函数值
初始条件 1	本文方法	42	15	-3.46E+06
	文献 [4] 方法	53	23	-3.59E+06
	文献 [6] 方法	82	35	-3.56E+06
初始条件 2	本文方法	49	19	-3.47E+06
	文献 [4] 方法	68	28	-3.61E+06
	文献 [6] 方法	75	33	-3.58E+06
初始条件 3	本文方法	38	12	-3.49E+06
	文献 [4] 方法	65	31	-3.55E+06
	文献 [6] 方法	89	25	-3.59E+06
初始条件 4	本文方法	32	20	-3.46E+06
	文献 [4] 方法	51	39	-3.54E+06
	文献 [6] 方法	55	51	-3.62E+06
初始条件 5	本文方法	45	18	-3.49E+06
	文献 [4] 方法	72	37	-3.59E+06
	文献 [6] 方法	88	42	-3.53E+06
初始条件 6	本文方法	41	13	-3.49E+06
	文献 [4] 方法	89	21	-3.62E+06
	文献 [6] 方法	60	36	-3.65E+06

3.3 实验结果分析

综合分析不同方法拟合曲线和原始曲线比较结果可以看出,本文提出的基于半监督机器学习的改进 EM 算法得到的拟合曲线和原始曲线的重合程度比文献 [4] 方法、文献 [6] 方法明显更优,表明本文方法拟合性能明显高于其它两种方法,这主要是因为本文改进 EM 算法对传统 EM 算法容易陷入局部最优的弊端进行了有效的优化,使得本文方法得到的最优解和实际值更加吻合,而文献 [4] 方法、文献 [6] 方法无法解决传统 EM 算法容易陷入局部最优的问题,得到的解并非最优解,导致拟合精度低.

分析不同方法拟合性能可以看出,在初始化相同情况下,本文方法对模型容易进入局部最优的问题起到了明显的抑制作用,能够获取更优的拟合结果. 不仅如此,本文方法迭代次数最少,迭代时间最短,整体性能优.

4 结 论

本文针对 EM 算法及其改进改进所存在的不足,使用半监督机器学习机制对其进行改进和优化.

使用二项分布概率函数以及惩罚概率定理以及,对最大似然函数进行描述,将惩罚因子加入到最大似然函数之中,大大降低了模型的最大似然估计的误差.

对 EM 算法实现过程进行分析,引入非负约束先验信息,针对 EM 算法很难在全局量空间中求解出最优解问题,结合半监督机器学习机制实现 EM 算法的优化改进,利用其中的一种自训练学习模式,在每次的训练的过程中,把确定标本放入到标记集合之中,经过自己所得到的结果进行训练,获得很好的训练结果,很好地避免了陷入局部最优. 在此基础上,EM 算法数学模型参数通过最大似然方法进行计算和优化,构造

似然函数,在最大似然函数增加惩罚最小二乘因子,引入非负约束当成先验信息,转换成最小化问题.

仿真实验结果表明,本文所提的基于半监督机器学习模式的改进 EM 算法得到的拟合曲线和原始曲线的重合程度比其它方法明显更优,说明所提方法不会陷入局部最优,拟合性能好. 除此之外,所提方法迭代次数最少,迭代时间最短,整体性能优.

References:

[1] David K W Ho, Bhaskar D Rao. Antithetic dithered 1-bit massive MIMO architecture: efficient channel estimation via parameter expansion and PML [J]. IEEE Transactions on Signal Processing, 2019, 67(9): 2291-2303.

[2] Espinozasánchez N A, Chimalramírez G K, Fuentespananá E M. Analyzing the communication between monocytes and primary breast cancer cells in an extracellular matrix extract(ECME)-based three-dimensional system [J]. Journal of Visualized Experiments, 2018, DOI: 10.3791/56589.

[3] Beesley L J, Taylor J M G. EM algorithms for fitting multistate cure models [J]. Biostatistics, 2018, 20(3): 1-17.

[4] Liu Zhe, Song Yu-qing, Xie Cong-hua, et al. Clustering gene expression data analysis using an improved EM algorithm based on multivariate elliptical contoured mixture models [J]. Optik, 2014, 125(21): 6388-6394.

[5] Lin Y, Schmidtlein C R, Li Q, et al. A krasnoselskii-mann algorithm with an improved EM preconditioner for PET image reconstruction [J]. IEEE Transactions on Medical Imaging, 2019, 38(9): 2114-2126.

[6] Sun B, Yan G, Ning L, et al. Multiple target counting and localization using variational bayesian EM algorithm in wireless sensor networks [J]. IEEE Transactions on Communications, 2017, 65(7): 2985-2998.

[7] Bellili F, Meftehi R, Affes S, et al. Maximum likelihood SNR estimation of linearly-modulated signals over time-varying flat-fading SIMO channels [J]. IEEE Transactions on Signal Processing, 2014, 63(2): 441-456.

[8] Peng Yuan-yuan, Xiao Chang-yan. Pulmonary fissure detection in CT images based on expectation maximization algorithm and surface fitting model [J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(10): 1870-1877.

[9] Liu J, Gasbarra D, Railavo J. Fast estimation of diffusion tensors under rician noise by the EM algorithm [J]. Journal of Neuroscience Methods, 2016, 257: 147-158, DOI: 10.1016/j.jneumeth.2015.09.029.

[10] Ranjan R, Huang B, Fatehi A. Robust Gaussian process modeling using EM algorithm [J]. Journal of Process Control, 2016, 42(5): 125-136.

[11] Tang Zheng, Song Yu-qing, Liu Zhe. Research on clustering based on improved particle swarm optimization and expectation maximization algorithm [J]. Journal of Chinese Computer Systems, 2015, 36(7): 1602-1606.

[12] Qi J, Wang J, Sun K. Efficient estimation of component interactions for cascading failure analysis by EM algorithm [J]. IEEE Transactions on Power Systems, 2018, 33(3): 3153-3161.

[13] Zhang Wen, Jiang Yi-pan, Yin Guang-da, et al. Handling missing values in software effort data based on naïve Bayes and EM algorithm [J]. Systems Engineering-Theory & Practice, 2017, 37(11): 2965-2974.

[14] Meng Bo, Zhu Ming. The application of EM algorithm to improve particle filter [J]. Journal of Image and Graphics, 2018, 14(9): 1745-1749.

[15] Zhao Yue, Li Hong. Chinese word segmentation cognitive model based on maximum likelihood optimization EM algorithm [J]. Bulletin of Science and Technology, 2016, 32(4): 178-181.

[16] Xu Xing-pei, Song Yu-qing, Lu Hu. Image segmentation based on deep learning features and community detection [J]. Journal of Chinese Computer Systems, 2018, 39(11): 2533-2537.

[17] Cao Wei-quan, Li Zhi-xiang, Wei Qiang, et al. Trajectory classification method based on probability density estimation of regional distribution [J]. Computer Engineering, 2018, 44(4): 262-267+286.

[18] Yi Qing-ming, Luo Chong, Shi Min. High dynamic GPS tracking loop based on fast maximum likelihood estimation [J]. Computer Engineering, 2016, 42(8): 300-304.

[19] Duan Suo-lin, Yan Xiang, Zhu Fang, et al. Object detection algorithm based on improved mixed Gaussian model and six-frame difference [J]. Computer Engineering, 2017, 43(7): 234-238.

附中文参考文献:

[8] 彭圆圆,肖昌炎. 基于 EM 算法和曲面拟合模型的肺裂检测算法 [J]. 计算机辅助设计与图形学学报, 2018, 30(10): 1870-1877.

[11] 汤　峥,宋余庆,刘　哲. 基于粒子群优化和 EM 算法的图像聚类研究 [J]. 小型微型计算机系统, 2015, 36(7): 1602-1606.

[13] 张　文,姜祎盼,殷广达,等. 基于朴素贝叶斯和 EM 算法的软件工作量缺失数据处理方法 [J]. 系统工程理论与实践, 2017, 37(11): 2965-2974.

[14] 孟　勃,朱　明. 采用 EM 算法对粒子滤波跟踪算法进行改进 [J]. 中国图象图形学报, 2018, 14(9): 1745-1749.

[15] 赵　越,李　红. 极大似然优化 EM 算法的汉语分词认知模型 [J]. 科技通报, 2016, 32(4): 178-181.

[16] 胥杏培,宋余庆,陆　虎. 一种结合深度学习特征和社团划分的图像分割方法 [J]. 小型微型计算机系统, 2018, 39(11): 2533-2537.

[17] 曹卫权, 李智翔, 魏　强,等. 基于区域分布概率密度估计的轨迹分类方法 [J]. 计算机工程, 2018, 44(4): 262-267+286.

[18] 易清明,罗　翀,石　敏. 基于快速最大似然估计的高动态 GPS 跟踪环 [J]. 计算机工程, 2016, 42(8): 300-304.

[19] 段锁林,严　翔,朱　方,等. 基于改进混合高斯模型与六帧差分的目标检测算法 [J]. 计算机工程, 2017, 43(7): 234-238.