第 19卷第 1期
2005年 3月

山西师范大学学报（自然科学版）
Journal of Shanxi Teacher's University
Natural Science Edition

Vol. 19 No. 1
Mar. 2005

文章编号：1009-4490（2005）01-0046-04

基于高斯混合模型的 EM学习算法

王 源[1,2]，陈亚军[3]

（1. 西华师范大学计算机学院微机应用研究所，四川 南充 637002；
2. 淮南师范学院信息技术系，安徽 淮南 232001；
3. 西华师范大学物理与电子信息学院，四川 南充 637002）

摘 要：本文研究了一类基于无监督聚类学习的算法—— EM算法的算法实现. EM算法通常用于存在隐含变量时的聚类学习，由于引入了隐含变量，导致算法难以保证收敛和达到极优值. 本文通过将该算法应用于高斯混合模型的学习，引入重叠度分析的方法改进 EM算法的约束条件，从而能够确保 EM算法的正确学习.

关键词：高斯混合模型；EM算法；无监督聚类；机器学习

中图分类号：TP181 文献标识码：A

0 引言

在模式分类中，基于模型的无监督学习是一种自动学习的方式，不需要对学习样本做类别标记，利用已知的数学模型通过逐步逼近的方法，使给定数据集与数学模型之间达成最佳拟合. 在许多实际的机器学习问题框架中，相关实例特征中只有一部分可以被观察到，我们常常只能根据所观察到的样例去推断未知的数据. 也就是说，在许多现实世界中的问题存在着隐含变量（hidden variables），有时又称为潜在变量（latent variables）[1]，是指在学习过程中未完全观察到的数据. 事实上，某些变量有时能观察到，有时不能，通常的办法是：使用已经观察到的该变量的实例去在一定范围内估计未观察到的实例中的变量的值. 隐含变量的出现能够大幅度减少参数的数目，但在大幅度减少数据数量的同时需要设置学习参数，从而使学习的问题变得复杂.

聚类分析也称为数据分割，具有多种目标，但都涉及把一个对象集合分组或分割为子集或“簇”，使得每个簇内部的对象之间的相关性比其他簇中对象之间的相关性更紧密. 无监督聚类是在多种对象集合中辩识的问题，之所以叫无监督，是因为分类标志未事先给定，基于无监督聚类的机器学习称为无监督学习.

基于模型的聚类方法就是试图对给定数据与某个数学模型达成最佳拟合，这类方法经常是基于数据都是有一个内在的混合概率分布假设来进行的. 基于模型聚类方法主要有两种：统计方法和神经网络方法，本文采用的 EM算法属于统计学习方法，是从不完全数据中计算极大似然估计的重复统计技术，比照传统神经网络的学习方法，它具有低开销，不用设置学习步长、易收敛、收敛速度快和易于实现的特点，是当前机器学习领域的主流技术之一.

1 高斯混合模型

假定我们有一系列观察值由混合分布 P产生，该分布由 k个独立同方差的高斯分布构成，即有 k个成

收稿日期：2004-07-08
基金项目：四川省教育厅重点项目基金资助（2004A102）.
作者简介：王源（1971— ），男，安徽淮南人，淮南师范学院讲师，硕士，主要从事机器学习方面的研究.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分. 首先选取一个成分然后基于该成分产生一个样本从而得到数据点. 设定有 N个点组成了指定的数据集 D D= { x i} $_{i=1}^{N}$. 将数据集 D在 d维空间中的对应的点作为一定分布的样本值, 则此分布可由 k个高斯密度函数的加权平均所表示的概率密度函数描述如下:

$$p(x;\Theta)=\sum_{j=1}^{k}\alpha_j G(x;m_j,\sum_j),\quad \alpha\geqslant 0\text{且}\sum_{j=1}^{k}\alpha_j=1 \tag{1}$$

其中

$$G(x;m_j,\sum_j)=\frac{\exp[-\frac{1}{2}(x-m_j)^T\sum_j^{-1}(x-m_j)]}{(2\pi)^{d/2}|\sum_j|^{1/2}}$$

也即, p(x;Θ) 是多个高斯密度函数的有限组合, 称为高斯混合密度或高斯混合模型 (Gaussian MixtureModel), 简记为 GMM, 模型中 x表示随机向量, d是向量 x的维数, Θ是参数向量集合. α j是混合模型中基模型高斯密度函数的权重、m j为均值向量、$\sum_j=(\delta_{mn}^{(j)}\delta_{mn}^{(j)})_{d\times d}$为协方差阵 (正定矩阵).

2　EM算法[2]

通过上述模型定义, 我们可以在具有隐含变量的变量和实际数据之间建立上述概率模型, 这就是学习的目标.

采用 EM算法的基本思想是对于上述不完整数据集 D, 假设这些数据独立同分布于我们已知的某一模型, 如 GMM, 而我们知道该模型的参数, 因此可以根据该模型推出属于每个成分的各数据点的概率. 然后, 修改每个成分的值 (这里每个成分适合于整个数据集, 且每个点由属于该成分的概率是否有利而得到), 重复该过程直到收敛到结束条件. 本质上, 我们通过推断含有隐含变量的概率分布得到 "完整" 的数据, 每个数据点中都有这些隐含变量的成分, 且基于当前的模型.

对于高斯混合分布, 我们任意初始化该混合模型的参数, 学习步骤如下:

(1)初始化: 对各类别密度分布待估计的参数 Θ的初值设置, 包括各类别的比例、均值向量 μ和协方差矩阵∑.

(2)E—步: (期望步)计算隐含变量 (设为 Z ij)数据的期望值. 用随机变量 C指示数据成分, 则概率 p ij = P(C= i| x j)表示数据 x j由成分 i产生的概率, 也即由第 i个高斯分布产生的概率. 由贝叶斯公式, 有 p ij = p(x j| C= i) P(c= i), 其中 P(x j| C= i)即 x j在第 i个高斯分布中的概率. 而 P(C= i)是第 i个高斯分布的权重参数. 应用到上述定义的 GMM中, 表示如下:

$$p(j|x_i)=\frac{\alpha_j G(x_i;m_j,\sum_j)}{p(x_i;\Theta)},\quad j=1,2\cdots k \tag{2}$$

(3)M—步: (极大化步)在该步中, 主要通过求解对数似然方程, 计算出期望值到达极大值点时新的均值 m j、协方差矩阵∑ j及权重 α j, 用于下次叠代.

$$m_j\leftarrow\frac{\sum_{i=1}^{N}p(j|x_i)x_i}{\sum_{i=1}^{N}p(j|x_i)}=\frac{1}{\alpha_j N}\sum_{i=1}^{N}p(j|x_i)x_i \tag{3}$$

$$\sum_j\leftarrow\frac{\sum_{i=1}^{N}p(j|x_i)x_i x_i^T}{\sum_{i=1}^{N}p(j|x_i)}=\frac{1}{\alpha_j N}\sum_{i=1}^{N}p(j|x_i)x_i x_i^T \tag{4}$$

$$\alpha_j^{NEW}=\frac{1}{N}\sum_{i=1}^{N}\frac{\alpha_j^{OLD}G(x_i;m_j,\sum_j)}{\sum_{j=1}^{k}\alpha_j^{OLD}G(x_i;m_j,\sum_j)}=\frac{1}{N}\sum_{i=1}^{n}p(j|x_i) \tag{5}$$

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(4)满足结束条件则停止,否则转第(2)步.

由上可知,整个 EM算法分为两步:E步,又叫期望步,能通过计算隐含变量 Z_{ij}的期望值 p_{ij}得到.这里的 Z_{ij}值当 x_j由第 i个成分产生时为1,否则为0. M步,又叫极大化步,基于最大化隐含指示变量数据(已计算出期望值)的对数似然估计值寻找参数的新值.

其中初始 E步,虽然各混合密度函数的参数可以选取一个随机初值.但如果有一定知识支持,可以选取一个有效的初始值,以便于缩小 EM算法的搜索空间,基于此种方法的学习有时又称为半监督学习方法,在此不赘述. EM算法是一个与最大似然估计相一致的算法,能够收敛,但却无法保证收敛的正确性,也即收敛到与样本所服从分布的真参数相一致的解,尤其是当组成混合模型的支密度函数只局限于单种样本失去泛化能力时,如在某一稀疏分布区域仅存在一个或极少样本时,算法可能失败. Wu[3]证明了 EM算法在某种正则条件下能使原似然函数或对数似然函数收敛到它的极大值或局部极大值,但要映射到高维的正定矩阵进行运算,这通常难以实现.本文是基于各分支密度函数重叠度分析的方法.

3 GMM模型中混合支密度的重叠度讨论

我们首先定义高斯混合分布中各支分布的重叠度如下.

定义 3.1 对于高斯混合分布中第 i个高斯分布和第 j个高斯分布的重叠度为:

$$e_{ij}(\Theta)=\lim_{n\to\infty}\frac{1}{N}\sum_{t=1}^{n}|r_{ij}(x_t)|=\int|r_{ij}(x)|p(x,\Theta)\,dx,\quad i,j=1,2,\cdots,k \tag{6}$$

其中 $r_{ij}(x)=[\delta_j-h_i(x)]h_j(x)$,且 $h_i(x)=\dfrac{\alpha_i G(x,m_i,\sum_i)}{\sum_{k=1}^{k}\alpha_k G(x,m_k,\sum_k)}$ $i=1,2,\cdots,k$因为 $|r_{ij}(x)|\leqslant 1$,所以 $e_{ij}(\Theta)\leqslant 1$.

易得到 $e_i(\Theta)=\sum_{j\neq i}e_{ij}(\Theta)>0$

表示第 i个高斯分布和混合密度中其他高斯分布的重叠度.我们有以下的定义:

定义 3.2(最大重叠度)

$$e(\Theta)=\max_{ij}e_{ij}(\Theta)\leqslant 1 \tag{7}$$

条件 3.1 $\varepsilon D_{max}(\Theta)\leqslant D_{min}(\Theta)\leqslant\|m_i-m_j\|\leqslant D_{max}(\Theta)$

其中 $D_{max}(\Theta)=\max_{i\neq j}\|m_i-m_j\|$, $D_{min}(\Theta)=\min_{i\neq j}\|m_i-m_j\|$, $\varepsilon>0$.即混合密度的重叠度减小到0时,任意两个均值 m_i, m_j不能任意靠近.

条件 3.2 任意给定的正数 $\varepsilon>0$, $\xi>0$, $\lambda(\Theta)$是协方差矩阵族$\sum_1,\cdots,\sum_k$的最大特征数,则有 $\varepsilon\lambda(\Theta)\leqslant\lambda_{ij}\leqslant\xi\lambda(\Theta)$,其中 $i=1,2\cdots k$ $j=1,2,\cdots,d$.

条件 3.3 对于 GMM中任一分布的高斯密度函数的权重 α_i总 $\exists\varepsilon>0$,使得 $\alpha_i>\varepsilon$.

由以上条件可以得到(详见文献[2])以下定理:

定理 3.1 设 $\{x_t\}_1^N$为独立同分布的随机样本,来自参数为 Θ的 K个高斯混合分布, Θ^N是该样本的最大似然一致解,即 $\lim_{N\to\infty}\Theta^N=\Theta$,若参数 Θ满足条件 3.1、3.2、3.3时,最大重叠度 $e(\Theta)\to 0$,当 N充分大时,存在 Θ^N的闭邻域 $N(\Theta^N)$,对任意初始值 $\Theta^{(0)}\in N(\Theta^N)$, EM算法必唯一收敛到 Θ^N.

4 改进后的 EM算法

基于上述分析,我们可以对标准 EM算法做如下改进,以确保其正确收敛.

a)初始化:对各类别密度分布待估计的参数 Θ的初值设置,包括各类别的比例、均值向量 μ和协方差矩阵 Σ;

b)重叠度判定:如果满足定理 3.1的条件,则转入 c),否则转入预处理程序;

c) E步,计算(2)式;

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

d)M步:计算(3)、(4)、(5)三式;

e)收敛性判断:如果满足结束条件则停止,否则转 c).

这里的预处理程序可以对每一类密度分布进行再分解,能进一步细化重叠度,也可以引入稳健统计的方法消除孤立点的干扰.

5　EM算法的通用形式

事实上 EM算法是通过计算某一实例中隐含变量的期望值,然后再计算参数,用期望值代替观测值.令 x为所有实例中的各观测值, Z表示各实例中的隐含变量,为概率模型的参数.则 EM算法有如下的通用(简化)形式:

$$\Theta^{(i+1)} = \arg\max_{\Theta} \sum_{z} P(Z = z \mid x, \Theta^{(i)}) L(x, Z = z \mid \Theta^{(i)}) \tag{8}$$

未知样本的类别归属:得到各类别的最大似然参数后,逐点从实例中读入未知样本 x,根据贝叶斯判别公式决定 x的归属.通过上述分析,我们知道 EM算法是通过补充无标号样本数据来改善训练样本选取的不完备性,然后通过 EM迭代计算纠正各个类别分量的最大似然函数参数集的估计,使总体密度函数分布更接近于实际分布.

6　结束语

本文对 EM算法应用于无监督聚类学习进行了理论研究,并对 EM算法约束条件给予改进,使之能够在相关邻域内正确收敛.虽然 EM算法只能得到局部最优解,我们也可以采用重叠度分析的方法改善参数初值的设置,从而使求得的解更好地逼近全局最优解,增强本文方法的通用性.

参考文献:

[1] T M Mitchell. Machine Learning[M]. USA: McGraw Hill, 1997.

[2] 付淑群, 等. EM算法正确收敛性的探讨[J]. 汕头大学学报, 2002, 17(4): 1~12.

[3] Wu C F G. On the convergence properties of the EM algorithm[J]. Annals of Statistics, 1983, 11: 95~103.

[4] 茆诗松, 王静龙, 濮晓龙. 高等数理统计[M]. 北京: 高等教育出版社, 1998.

[5] Dempster A P, Laird N, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm (with discussion)[J]. J Royal Stat Soc B, 1977, 39: 1~18.

[6] George F Luger. Artificial Intelligence: Structure and Strategies for Complex Problem Solving[M]. UK: Pearson Education, 2002.

[7] Trevor Hastie et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction[M]. USA: Springer-Verlag, 2001.

A Study of EM Learning Algorithm Based on Gaussian Mixture Model

WANG Yuan[1,2], CHEN Ya-jun[3]

(1. Institute of Microcomputer Application, The School of Computer, West China Normal University, Nanchong, Sichuan 637002, China;
2. Information Technology Department, Huainan Teachers College, Huainan, Anhui 232001, China;
3. The School of Physics & Electronics, West China Normal University, Nanchong, Sichuan 637002, China)

Abstract: In this paper, We conducts a theoretical analysis into the method of Machine leaning with EM algorithm which is an unsupervised-clusting one. The EM algorithm used to estimate some clusting-learning parameters including hidden variables. Which lead to difficulties of converging correctly and obtaining to the local maximum points. We use EM algorithm to learn some parameters of Gaussian Mixture Model and demonstrate that the analysis of the mixture density's overlap measure can enforce the restrict conditions of EM algorithm, as a result, this analysis can assure the efficiency of learning.

Key words: Gaussian-Mixture-Model; EM-algorithm; unsupervised-clusting; Machine-Learning

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net