文章编号:1671 - 6833(2009)03 - 0130 - 04

一种混合递增 NEM 的空间聚类算法

贾俊杰,张 勤

(长安大学 地质工程与测绘工程学院,陕西 西安 710064)

摘 要:由于 EM 算法不适合空间聚类对空间信息的要求,而邻域 EM 算法虽然结合了空间惩罚项,但是 NEM 在 E - step 步需要大量的迭代.为了既能满足空间信息的要求,又能避免过多的计算量,本文提出了 EM 与 NEM 二者相结合的混合递增 NEM 算法,算法首先在随机子样本中进行 EM 训练,直到似然判断条件下降,根据增量因子进行样本更新,然后样本转向 NEM 训练一次,如此进行循环递增的交叉训练,使得计算量降低,性能提高.实验结果显示,MNEM 只需要较少的运算便可达到收敛,聚类质量结果优于 NEM.

关键词:空间聚类;NEM 算法;高斯混合;空间惩罚项

中图分类号:TP 311 **文献标识码**:A

0 引言

空间聚类是指将数据对象集分组成为由类似的对象组成的簇,使得同一簇中的对象之间具有较小的相异度,而不同簇之间相异度较大.作为一种非监督学习方法,空间聚类不依赖于预先定义的类和带类标号的训练实例.实际上,几乎每一个样本点都与它的邻点相关,例如,附近相邻的房屋趋向于受到相似价格的影响.通常情况下,不同的区域具有不同的分布,这种分布被认为是空间异质性.具有相似特征的空间数据通常显示正自相关性和空间连续性.

高斯混合模型是一种常见的数学模型,它是指多个高斯分布的混合分布.笔者通过运用高斯混合模型,结合了 EM 算法[1]和邻域 EM (Neighborhood EM,NEM)算法[2]的思想,提出了一种混合递增 NEM 算法(Mixed Increasing NEM,MNEM),MNEM 算法利用样本逐步递增训练的思想,交替进行 EM 步骤和 NEM 步骤,直到递增后的样本与完全样本一致时,算法按照 NEM 的停止条件结束.实验证明,MNEM 算法在聚类质量和计算效率方面均优于 NEM 算法.

1 EM 算法

EM 算法是在观察数据为具有隐含变量(即不完全数据)时,对观测数据进行极大似然估计,通过多步迭代,使得似然值收敛于某个最优值的迭代算法.EM 算法被广泛应用于概率模型的参数估计.设 n 个空间点集合 $S=\{s_i\}_{i=1}^n$,有邻近关系 $N\subseteq S\times S$,点 s_i 和 s_j 相邻,当且仅当 $(s_i,s_j)\in N, i\neq j$,令 $N(s_i)=(s_j;(s_i,s_j)\in N)$ 表示 s_i 的邻点.假设 N 被邻接矩阵 W 给定,$W(i,j)=1$,当且仅当 $(s_i,s_j)\in N$,否则 $W(i,j)=0$.每个 s_i 都关联一个 d 维的属性特征向量 $x_i=x(s_i)\in\mathscr{X}^d$,每个对象 x_i 具有一个类标记 $y_i\in\{1,\cdots,m\}$.

假设观测样本 $X=\{x_i\}_{i=1}^n$,有包含 n 个成分的有限混合分布模型:

$$p(X|\Phi)=\sum_{i=1}^{m}\pi_i p_i(X|\theta_i) \tag{1}$$

$p_i(X|\theta_i)$ 是带有参数 θ_i 的第 i 个成分的概率密度函数;π_i 是第 i 个成分的先验概率,且有 $\sum_{i=1}^{m}\pi_i=1$;$\Phi=\{\pi_i,\mu_i,\sum_i\}_{i=1}^m$ 表示在高斯混合模型下所有参数集合;μ_i 和 $\sum_i$ 分别表示第 i 个成分的均值和协方差,有对数似然函数:

$$L(\Phi)=\sum_{j=1}^{n}\ln\left[\sum_{i=1}^{m}\pi_i p_i(x_j|\theta_i)\right] \tag{2}$$

通过 $\partial L/\partial\Phi=0$ 求解最大似然估计一般是困难的.因此 EM 尝试在不完全数据的情况下,为每个 x 增加一个缺失值 $y\in\{1,\cdots,m\}$,y 指明 x 来

收稿日期:2009 - 01 - 27;修订日期:2009 - 03 - 10

基金项目:国家自然科学基金资助重点项目(40534021)

作者简介:贾俊杰(1974 -),男,甘肃兰州人,长安大学博士研究生,从事空间数据挖掘与地理信息系统应用方面的研究.

自哪个成分，对最大值 L 进行多步迭代，即 $p(x|y=i)=p_i(x|\theta_i)$. 实际上，算法形成了一个估计序列 $\{\Phi^t\}$，Φ^0 是初始估计，*EM* 算法分两步：

E - step：估计函数 Q 是完全数据 $\{x,y\}$ 的对数似然的条件期望，$E_p[g]$ 表示关于 y 的分布 $p(y)$ 的期望，且 $p_{\Phi^{t-1}}(y)=p(y|X,\Phi^{t-1})$，有

$$Q(\Phi,\Phi^{t-1})=E_{P_{\Phi^{t-1}}}[\ln(p(\{x,y\}|\Phi))] \tag{3}$$

即计算 $p_{ij}^{t-1}=\dfrac{\pi_i^{t-1}p_i(x_j|\theta_i^{t-1})}{\sum_{i=1}^{m}\pi_i^{t-1}p_i(x_j|\theta_i^{t-1})}$，$p_{ij}^{t-1}$ 表示样本 x_j 在第 i 个成分中的概率.

M - step：计算参数 Φ_i^t.

$$\alpha_i^t=\frac{1}{n}\sum_{j=1}^{n}p_{ij}^{t-1} \tag{4}$$

$$\mu_i^t=\frac{\sum_{j=1}^{n}x_jp_{ij}^{t-1}}{\sum_{j=1}^{n}p_{ij}^{t-1}} \tag{5}$$

$$\sum\nolimits_i^t=\frac{\sum_{j=1}^{n}p_{ij}^{t-1}(x_j-\mu_i^t)(x_i-{}_i^t)^T}{\sum_{j=1}^{n}p_{ij}^{t-1}} \tag{6}$$

EM 算法通过迭代地进行 *E - step* 和 *M - step*，直到参数收敛为止. 即 $\|Q(\Phi^i|\Phi^{t-1})-Q(\Phi^{i-1}|\Phi^{t-1})\|$ 充分小为止. *EM* 算法在理论上能够收敛到参数空间的局部极值.

2　邻域 *EM*

NEM 算法为了结合空间信息，对于包括所有样本点分布 $p(y)$ 的函数 Q，加入一个空间惩罚项 $G(p)$. 如果相邻点有相似的 $p(y)$，则惩罚项将被最大化. $p(y_i)$ 是一个 m 维列向量 $[p_{i1},L,p_{im}]$. 事实上，由 $[p(y_1),L,p(y_n)]$ 形成的矩阵可以被认为是一个模糊分类矩阵[3].

$$\begin{aligned}G(P)&=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}W(i,j)p_{im}p_{jm}\\&=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}W(i,j)P(y_i)gP(y_j)\end{aligned} \tag{7}$$

因此得到一个 NEM 算法的新的估计函数 $F(p,\Phi)$，平滑因子 $\beta>0$ 是一个固定参数[4-5]. β 的缺省值是 1，以保证向量的尺度关系在转换后是无损的，NEM 建议 $\beta\in[0.5,1]$.

$$F(p,\Phi)=Q(p,\Phi)+\beta G(p) \tag{8}$$

类似于 Q，F 可以交替最大化估计它的两个参数，直到参数收敛为止，收敛条件与 EM 算法类似. NEM 算法通过 p 不变，可以计算 M - step. 在 E - step，Φ 是不变的. 如果在 p^* 处 F 被最大化，那么 $\partial F'/\partial p_{ik}=0$，$F'$ 是考虑 p 的约束的情况下 F 的拉格朗日函数，由公式(11)求解 p_{ik}^*. 实际上 NEM 算法只是对 EM 算法的 E - step 做了加入内积的操作[6].

$$p_{ik}^*=\frac{\pi_kp_k(x_i|\theta_k)\exp\left(\beta\sum_{j=1}^{n}W(i,j)p_{jk}^*\right)}{\sum_{l=1}^{K}\pi_lp_l(x_i|\theta_l)\exp\left(\beta\sum_{j=1}^{n}W(i,j)p_{jl}^*\right)} \tag{9}$$

3　混合递增 NEM

MNEM 算法步骤主要分为两步：1. 样本按几何方式递增的 EM 步骤；2. 用 NEM 步骤剔除空间不相邻的样本. 1 步和 2 步交替执行，直到递增后的样本与完全样本一致，按照 NEM 停止条件结束. 两步之间交叉运行的条件需要分析给定.

3.1　增量因子

由于初始样本的选取直接影响了算法的性能. 这里根据文献[7]给定增量因子 $d=\ln(N)/2$，初始子样本数量选取为 $M=N/d$. 当子样本在达到最佳拟合其真实分布的前提下，加入新的样本，再次进行拟合，若为最佳拟合，再次加入新的样本…，直到子样本的数量与完全样本一致时停止增加. 在这个递增的过程中，子样本逐渐逼近完全样本的真实分布. 若将子样本看作一个完全样本，则每次的样本增加量看作是选取的新样本. 因此根据初始样本选取的方法进行子样本增量的选取，令每次子样本新增样本数 $h=M/d$，注意，这里的 M 是样本增加前一次的 M 值，则更新后的 $M=M+M/d$，这种增量方式是逐步递增的过程.

通过 M 的递增可以看到子样本呈几何级数增加，整个样本训练过程所遍历的样本数远小于 NEM 算法. 但是下一个关键问题在于每次对 EM 步骤中的样本集进行样本增加时的判别条件是什么？根据 EM 算法的停止条件可以看到，每次迭代都要进行极大似然值的比较，当似然值不大于前一次的迭代值时，算法停止，即当前样本集针对其真实分布达到了最佳拟合，若要与完全样本集的真实分布达到最佳拟合，就必须增加新的样本. 鉴于这种思想，MNEM 算法也在每次 EM 迭代过程中对每个样本子集进行这种似然判断. 唯一与 EM 算法不同的是，MNEM 算法每次在 EM 迭代过程中，当似然值大于上一次的迭代值时，就不增加新的样本，而是再次训练原样本，直到不大于前一次的迭代值时才增加新的样本，并转到 NEM 进行训练.

3.2　递增判别条件

令 $(x_1,\cdots,x_N)$ 表示完整的数据集，$(x_{t_1},\cdots,x_{t_M})\subset(x_1,\cdots,x_N)$ 是随机选择的大小为 M 的子样

本,针对 EM 步骤则可以得到公式(3)的近似:

$$Q_M(\Phi,\Phi^{t-1}) = E_{P_{\Phi_{t-1}}}[\ln(p(\{x,y\}|\Phi))] \quad (10)$$

令 $B = Q_M(\Phi|\Phi^t) - Q_M(\Phi|\Phi^{t-1})$ 表示 t 次迭代与 $t-1$ 次迭代的似然差值,令 δ 为一充分小的数. 若 $B>\delta$,说明子样本集与其相应的子样本模型并不匹配,需要继续迭代;若 $B<=\delta$,说明子样本集与其相应的子样本模型匹配,而与完全样本集的高斯模型进行匹配还需要一段距离,因此需要增加额外的信息量来进行样本估计,也就是需要增加新的样本点来进行训练. 其实整个过程可以认为是子样本对整体样本的高斯模型进行逼进过程.

MNEM 算法在 EM 转向 NEM 后的条件下,只进行了一次内积操作,这是由于在 EM 过程中,虽然没有考虑样本点之间的空间邻近关系,但是具有类似属性特征的样本点已经按照高斯混合分布分别聚类,或者说在非空间条件下,每个样本点已基本稳定. 如果在此条件下考虑样本的空间邻近关系,就使得相邻点的分类确定性增强. 利用空间邻接矩阵 W 对属性相似但空间不邻近的样本点剔除,因此,在 *NEM* 条件下只需运行 *E-step* 一次就可以确定相邻点的分类.

3.3 MNEM 算法描述

MNEM 算法先是针对小样本,然后是大样本,直至完成样本的训练过程,这种 *EM* 与 *NEM* 之间交叉训练使算法涉及的运算量比 *NEM* 大幅减少.

如果点 s_i 在来自同一类的所有点中具有最大值 $p(y)$,即 $\exists k, \forall s_i \in \{s_j\} \cup N(s_j), p_{ik} = \max_l \{p_{il}\}$,则认为该点为聚类中心. 在子样本中,NEM 条件下,$p(y)$ 转化为一个硬分类,即除了最大值为 1 外,所有值的概率为 0. 这是因为,在空间聚类中,如果空间连续存在,那么,大多数的中心点将被同类的点包围. 随着子样本的增加,当前中心点的划分也将发生变化,这与 k-均值[8]算法的思想相似.

考虑当前中心点的划分,当递增后的子样本与完全样本一致的时候,那些中心点的 p 就稳定下来,并且不需要再估计. 设 n 是所有样本数据的大小,预先指定 m 个混合高斯成分和平滑因子 β,增量因子 d,初始子样本 $M = N/d$,初始参数 Φ^0,令 δ 为一充分小的数. 下面给出 MNEM 算法执行步骤描述:

①对长度 M 的子样本,由公式(3)计算 p_{ik}^{t-1}.

②计算参数 α_i^t, μ_i^t 和 $\sum_i^t$.

③计算似然判别式 B.

i. 如果 $B>\delta$,将参数 α_i^t, μ_i^t 和 $\sum_i^t$ 返回到①进行计算;

ii. 如果 $B<=\delta$,由公式(9)计算 P_{ik}^t;

④$M = M + M/d$.

⑤判断 M.

i. 如果 $M<n$,将 p_{ik}^t 代入②步执行.

ii. 如果 $M>=n$,令 $M=n$,执行 NEM 算法:

a. 计算参数 $\alpha_i^{t+1}, \mu_i^{t+1}$ 和 $\sum_i^{t+1}$;

b. 由公式(9)计算 P_{ik}^{t+1};

c. 若 $B>\delta$,执行 i 步,否则算法停止.

4 实验

实验采用某地区的 200 个区域(乡镇)的房价样本. 聚类就是基于每个居住点的房价均值. 对地区的经纬坐标进行归一化处理,应用 Matlab6.0 分析,相应的房价柱状图见图 1,若将整个样本集由 2 个成分构成的高斯混合分布来近似模拟,则得到房价在 3 000 元以下和 3 000~5 000 元之间两个聚类结果.

NEM 算法经过 20 次迭代,MNEM 算法经过 13 次迭代,得到运行时间 t、迭代次数 d 和极大似然值 L(见表 1),可以看到,MNEM 算法性能要高于 NEM 算法. 针对 NEM 和 MNEM 的两个聚类结果见图 2(a)和图 2(b),可以看到 MNEM 产生了一个在空间连续性上强于 NEM 的聚类结果,该聚类结果依赖于空间惩罚项.

表 1 房价数据的聚类比较

Tab.1 Clustering compare on house price data

参数	NEM	MNEM
t/s	0.267 5	0.137 2
d(num)	20	13
$-L(10^4)$	1.401 4	1.394 6

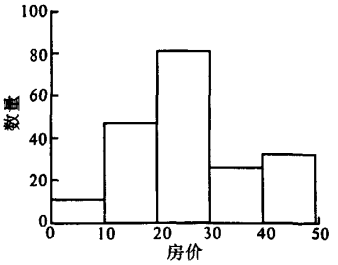

图 1 房价柱状图

Fig.1 The corresponding histogram of house price distribution in 200 towns

5 结论

由于 EM 算法不适合含有空间信息的空间聚类,而 NEM 算法虽然增加了空间惩罚项,但是需要在 E-step 进行多次迭代. 为了结合空间信息,

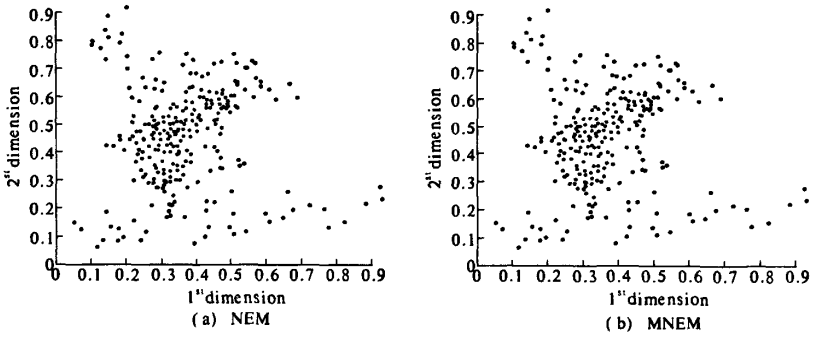

图 2　(a)和(b)分别是 NEM 和 MNEM 的聚类结果

Fig. 2　Two sample clustering results are shown in (a) and (b) for NEM and MNEM

又能避免额外的运算，笔者提出了结合 EM 和 NEM 的 MNEM 算法，MNEM 算法在每次 EM 步骤与 NEM 步骤交替过程中，首先利用 EM 算法对子样本进行非空间关系的聚类，然后通过 NEM 算法对结果进行空间邻近关系的筛选，基于这种循环交替的方式，最终逐渐得到完全样本集的极大似然估计. 实验证明 MNEM 聚类算法在计算效率和聚类质量方面都优于 NEM 算法. 类似于大多数 EM 算法，MNEM 的聚类效率也是要依赖于样本的初始化[9-10]，那么如何初始化 EM 参数将是下一步的研究工作.

参考文献：

[1] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm [M]. J. Roy. Statist. Soc. 1977, B(39): 1 - 38.

[2] AMBROISE C, GOVAERT G. Convergence of an EM - type algorithm for spatial clustering [J]. Pattern Recognition Lett, 1998, 19(10): 919 - 927.

[3] HATHAWAY R J. Another interpretation of the EM algorithm for mixture distributions [J]. Statist. Probab. Lett. 1986, 4, 53 - 56.

[4] KULLBACK S, LEIBLER R A. On information and sufficiency[J]. Ann. Math. Statist. 1951, 22, 79 - 86.

[5] ZAYANE O R, HAN J, ZHU H. Mining recurrent items in multimedia with progressive resolution refinement [C]. in: Proceedings of the 2000 International Conference on Data Engineering (ICDE'00), San Diego, CA, February 2000, 461 - 470.

[6] HU T M, SUNG S Y. A hybrid EM approach to spatial clustering[J]. Computational Statistics & Data Analysis, 2006, (50): 1188 - 1205.

[7] 张志兵. 空间数据挖掘关键技术研究[D]. 武汉：华中科技大学硕士学位论文, 2004.

[8] KAUFMAN K L, ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis [M]. New York, USA: John Wiley and Sons, 1990. 30 - 66.

[9] 岳　佳，王士同. 双重高斯混合模型的算法的聚类问题研究[J]. 计算机仿真, 2007, 24(11): 110 - 113.

[10] 谢勤岚. 基于 EM 算法的混合模型的参数估计[J]. 计算机与数字工程, 2006, 34(12): 42 - 44.

A Study of Mixed Increasing NEM Approach to Spatial Clustering

JIA Jun - Jie, ZHANG Qin

(Institute of Geology Engineering and Geomatics, Chang'an University, Xi'an 710064, China)

Abstract: EM algorithm is inappropriate for spatial clustering which requires consideration of spatial information. Although neighborhood EM algorithm incorporates a spatial penalty term, it needs more iterations in every E - step. To incorporate spatial information and avoid too much additional computation, this paper proposed mixed increasing NEM algorithm that combines EM and NEM. In MNEM, algorithm first train data based on random sub - sampling in EM till the likelihood - judgement condition begins to decrease, and update sub - sampling. Then training is turned to NEM and runs iteration of algorithm once. Because of this cross train of cycle, MNEM algorithm' computational complexity is decreased and capability is advanced. Experimental results show that less passes are needed in MNEM to converge and the final clustering quality is better than standard NEM.

Key words: spatial clustering; NEM algorithm; gaussian mixture; spatial penalty term