◎数据库与信息处理◎

聚类模型参数自动选择的图库索引

郑爱华,汤 进,罗 斌
ZHENG Ai-hua,TANG Jin,LUO Bin

安徽大学 计算机科学与技术学院,合肥 230039
School of Computer Science and Technology,Anhui University,Hefei 230039,China
E-mail:ahzheng214@163.com

ZHENG Ai-hua,TANG Jin,LUO Bin.Graph indexing based on automatic clustering model selection.Computer Engineering and Applications,2007,43(22):158-160.

Abstract: A graph database indexing method,which is based on pattern clustering and automatic model selection,is proposed.The traditional Expectation Maximization (EM) algorithm provides an effective method for parameter estimation in mixture model clustering,but the number of model components need to be fixed before the processing progress,which will certainly reduce the accuracy of the high dimensional indexing.The proposed indexing method is based on the automatic mixture model selection algorithm,which uses the improved component-wise EM algorithm,the vector quantization method and probabilistic approximation mechanism.The experimental results show that the retrieval efficiency is increased while the true positive rate is kept in high level.
Key words: automatic model selection;Component-wise EM of Mixture algorithm;vector quantity;probabilistic approximation

摘 要:提出一种基于模式聚类和混合模型参数自动选择的图库索引方法。因为传统的EM(Expectation Maximization)算法为混合模型聚类问题中的参数估计提供了一个很好的解决方法,但需要事先指定聚类数,影响了高维数据索引的精度和效率。综合利用改进的CEM^2(Component-wise EM of Mixture)混合模型自动选择算法、矢量量化和概率近似的索引机制,在保证准确率同时有效提高了检索效率。

关键词:模型自动选择;改进CEM^{2}算法;矢量量化;概率近似索引
文章编号:1002-8331(2007)22-0158-03 **文献标识码**:A **中图分类号**:TP391.41

1 引言

混合模型聚类作为一种统计工具已被广泛应用于数据挖掘[1]、图像处理[2,3]、统计数据分析[4]等许多科学领域。聚类算法近些年来得到了较大的发展,如通过最小化均值方差误差函数来实现聚类的K-means方法[5]和通过迭代方法确定模型参数的EM算法[6,7]等。但这些算法的最大缺陷是不能自动确定聚类的类别数。除非用户实现指定的类别数与图库数据的真实类别数,否则将会大大影响检索的准确率,从而使聚类失去意义。因此,聚类数的确定成为聚类的关键问题。本文旨在不明显影响检索精度的前提下,合理的利用自动选择的聚类算法和索引机制有效地提高检索效率。针对国际上常用的图库[8]进行实验得到了较理想的实验结果。

2 混合模型自动选择的图库索引

本文提出的图库索引系统主要包括原始特征库生成、量化特征库生成和特征检索模块等。系统框图如图1所示。

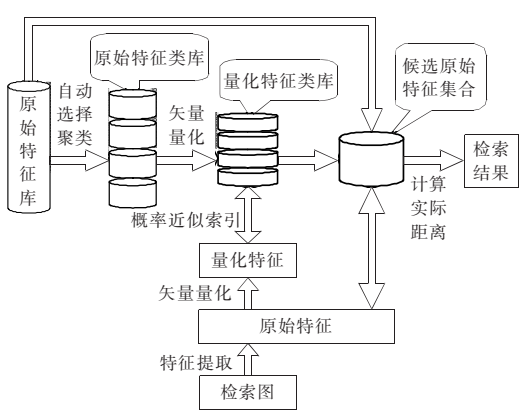

图1 基于模型自动选择的图库索引系统

2.1 图库的模型自动选择

如前面所述,聚类数的确定成为聚类的关键问题。本文方法在对CEM^{2}方法[9]加以改进后:不仅同时进行估计和模型选择,而且可以控制选择速度。只要给定其最大值k_{max}和最小值

基金项目:安徽省教育厅自然科学基金项目(No.KJ2007A045);安徽省高校青年教师科研资助计划项目(No.2006jql039)。
作者简介:郑爱华(1985-),女,硕士研究生,主要研究方向为图像处理和模式识别;汤进,博士生,讲师;罗斌,教授,博导。

$k_{\min}$，它可以在这个范围内自动选择一个最能反映图库数据真实分布的类别数 $bestk$。该方法适用于任何可能运用 EM 算法的参数混合模型。

设图库数据 $\boldsymbol{X}=[X_1,X_2,\cdots,X_N]^{\mathrm{T}}$ 由 N 个图组成,每个图用 d 维特征描述,并由 k_{nz} 个类别的混合模型产生,其中 k_{nz} 为混合概率非零的类别数。则其概率密度函数可以表示为:

$$f(x|\theta)=\sum_{i=1}^{k_{nz}}p_i f(x|\theta_i),\ \sum_{i=1}^{k_{nz}}p_i=1$$

其中,p_i 为混合概率,θ_i 为第 i 个类别的参数集。因此,整个图库数据的模型参数集可以表示成 $\theta=\{\theta_1,\cdots,\theta_{k_{nz}},p_1,\cdots,p_{k_{nz}}\}$。用 $Z=\{z^{(1)},\cdots,z^{(n)}\},z^{(i)}=[z_1^{(i)},\cdots,z_{k_{nz}}^{(i)}]$表示图库数据与类别的从属关系。

$$z_{ij}=\begin{cases}1 & x_i\text{是由第}j\text{个类别产生}\\ 0 & \text{其它}\end{cases}$$

则完整的 log 似然函数可以表示为:

$\log f(X,Z|\theta)=\sum_{i=1}^{n}\sum_{j=1}^{k_{nz}}z_j^{(i)}\log[p_j f(x^{(i)}|\theta_j)]$。模型自动选择实际上就是选择最佳 k_{nz} 下的参数估计。算法描述如下:

步骤 1 初始化 $k_{\max}$、$k_{\min}$、ε 和参数集 $\hat{\theta}(0)=\{\hat{\theta}_1,\cdots,\hat{\theta}_{k_{\max}},\hat{p}_1,\cdots,\hat{p}_{k_{\max}}\}$,并令 $k_{nz}=k_{\max}$。

步骤 2 改进的 CEM^2 算法进行模型自动选择。

(1)E-step:计算完全 log 似然函数的条件期望(conditional expectation)。

根据当前的 $\hat{\theta}(t)$ 的估计值,条件期望为:

$W=\{w^{(1)},\cdots,w^{(n)}\},w^{(i)}=[w_1^{(i)},\cdots,w_{k_{nz}}^{(i)}]$

$$w_j^{(i)}\equiv E[z_j^{(i)}|X,\hat{\theta}(t)]=\Pr[z_j^{(i)}=1|x^{(i)},\hat{\theta}(t)]=\frac{\hat{p}_j(t)f(x^i|\hat{\theta}_j(t))}{\sum_{m=1}^{k_{nz}}\hat{p}_m(t)f(x^i|\hat{\theta}_m(t))}\quad(1)$$

(2)M-step:更新参数估计。

$$\hat{p}_m(t+1)=\frac{\max\{0,(\sum_{i=1}^{n}w_m^{(i)})-\frac{N}{2}\}}{\sum_{j=1}^{k_{nz}}\max\{0,(\sum_{i=1}^{n}w_j^{(i)})-\frac{N}{2}\}}\quad(2)$$

$$\hat{\theta}_m(t+1)=\arg\max_{\theta}(\log f(X,W|\theta))$$

N 为决定类别的参数的个数。通过式(2)可以消除某些 $\sum_{i=1}^{n}w_m^{(i)}$ 较弱($\hat{p}_m(t+1)=0$)的类别。实际上,如果某些个别的类别存活系数(概率值)比其他类别都要小很多时,可以消除这些类别来提高模型选择的速度。因此,本文对式(2)进行改进:

$$\hat{p}_m(t+1)=\frac{\max\{\delta,(\sum_{i=1}^{n}w_m^{(i)})-\frac{N}{2}\}}{\sum_{j=1}^{k}\max\{\delta,(\sum_{i=1}^{n}w_j^{(i)})-\frac{N}{2}\}}\quad(3)$$

这样,就可以通过阈值 δ 来调整选择速度。由于消除这些类别后都将重新计算期望值,且所消除的类别本身存活系数就比较小，所以给整个模型自动选择带来的误差也将会是很小的。此时 $k_{nz}=k_{nz}-k_{died}$。k_{died} 为在式(3)中被消除的类别数。

收敛标准仍然采用 CEM^2 中的标准。通过对 MML(Minimum Message Length)标准[10]中很难计算的费希尔信息矩阵 $\boldsymbol{I}(\theta)$[11]近似处理得到。其代价函数为:

$$L(\theta,X)=\frac{N}{2}\sum_{i:p_i>0}\log(\frac{np_i}{12})+\frac{k_{nz}}{2}\log\frac{n}{12}+\frac{k_{nz}(N+1)}{2}-\log f(X|\theta)$$

若此时的代价与前一次的代价变化小于阈值 ε 则该步结束,否则重复该步。需要注意的是,与标准的 EM 算法不同的是,CEM^2 在 p_i 和 θ_i 更新之后都将重新计算 W 然后再更新 p_{i+1} 和 θ_{i+1},这样,如果某个类别消除($\hat{p}_m(t+1)=0$)后,立刻将它的概率块重新分配给其他的类别以提高它们的存活系数。

第 3 步:再次选择代价最小的参数估计。

当第 2 步中改进 CEM^2 的收敛后,可以得到 $L(\theta,X)$一个最小值。此时消除 p_m 最小的类别并返回第 2 步,直到 CEM^2 收敛。重复该过程直到 $k_{nz}\leqslant k_{\min}$。最后,选择使得 $L(\theta,X)$值最小的参数估计 $\hat{\theta}_{bestk}$。这样,就完成了模型自动选择。至此,便可以对图库数据 X 进行分类:(B_i 即为 x_i 所属的类别)

$$B_i=\arg_j(z_j^{(i)}=1),1\leqslant i\leqslant n,1\leqslant j\leqslant bestk$$

CEM^2 表面上似乎比标准的 EM 算法的计算量要大的多,因为要多次重复执行 E-step 来计算 W,但是,重复计算 W 时只有个别项需要重新计算，因此 CEM^2 的计算量仅比标准的 EM 略大[9]。

2.2　索引机制

在对数据进行准确分类后,索引机制的应用是必需的。“维数灾难”(Curse of Dimensionality)是推动检索机制发展的主要动力。在索引机制上,引用了量化误差较小的矢量量化方法[12]和概率近似索引[13]相结合的索引机制。这种将精确索引与近似索引相结合的方式,不仅大大降低了从磁盘上随机访问数据集里的原始数据的时间(I/O 时间),而且还有效的减少了数据之间的距离计算时间(CPU 时间)。

矢量量化与标量量化器相比,能更准确地描述数据的空间分布,从而减小的量化误差。它的过程大致分为以下几个步骤:(需要注意的是:以下各步骤都是在每个类别中独立进行的。)

步骤 1 用 KL 变换(Karhunen-Loeve Transform,KLT)[14]消除各维之间的相关性。

步骤 2 在总维数 b 不变的前提下,根据各维的能量(方差)σ_i^2 的不同分配不同的位数 b_i。这种非均匀的位数分配方法更符合数据的实际分布并能更加精确的估计距离的上下界。

步骤 3 根据每维的维数 b_i,用劳埃德算法[15]对每维进行空间划分。

步骤 4 可以通过原始数据在每维对应的区域生成对应的近似向量。

在索引机制的应用上,采用的概率近似的索引机制[13]实际上是利用概率的标准消除某些不太可能包含邻近点的区域所对应的类别的方式来简化检索过程,进而提高检索速度。虽然表面上看引入了一定的错误率，但由于图的特征本身的复杂性,即使不引入错误率,也无法保证检索结果的精确性。从后面的实验可以看出这种机制并未明显影响准确率。

3　实验结果与分析

在实验过程中,实验机器为 Lenovo 旭日 150 笔记本电脑,

PM 为 1.5 MHz,256 MB 内存,40 G 硬盘。实验图库是 huetB 图库中的 Dgraph 和 KNNgraph[8]为内容。对于实验数据,结果都是在相同环境下 20 次实验的平均值。在模型自动选择时,分别对不同的 k_{min} 和 k_{max} 进行实验,得到表 1 的对应的结果。

表 1 CEM^2 对图库不同特征自动选择结果

k_{min}	k_{max}	Dgraph 选择结果($bestk$)	KNNgraph 选择结果($bestk$)
1	10	8	5
10	20	16	15
20	30	28	26

为了更好地说明本文模型自动选择算法的有效性,给出 k=28 时的 Dgraph 数据投影到 2 维的聚类图(图 2)(为了更清楚的看到聚类结果,将以图中描述区域边界的纵横交错的椭圆线省略,而是选择各椭圆区域线的中心),并显示 k=15 的一个检索实例(图 3)。

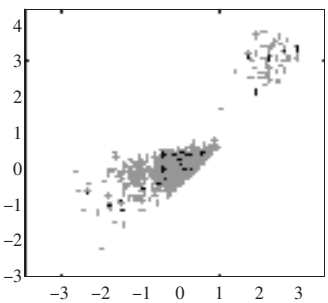

图 2 k=28 时 2 维的 Dgraph 聚类效果

图 3 k=15 时的一个检索实例

通过图 2、图 3,可以看出,该模型自动选择算法描述的数据分布与数据的实际分布是比较一致的。在整个检索过程中,不论是对量化数据的访问还是对原始数据的访问都是很少的。从表 2 可以很好的反应。

表 2 平均量化效率和平均访问原始数据的比率

检索方式	原始数据量化效率/%		访问原始数据的比率/%		
	2 bits/维	4 bits/维	k=8	k=12	k=16
本文	3.13	6.25	1.9	1.6	1.5
SS	100		100		

从时间复杂度来看,顺序检索的检索时间(Search Time,ST)包括提取检索图的特征时间 ST_F、图库中所有特征的距离计算时间 ST_{DIS} 和从磁盘读取数据的时间 $ST_{I/O}$:$ST_S=ST_F+ST_{DIS}+ST_{I/O}$。对本文采用的索引机制,由于对图库数据的聚类、量化都可以在检索前离线生成,所以,只增加了检索图的特征在索引文件的查询时间 ST_{index},而距离计算的时间可分为量化距离的计算时间 ST_{vdis} 和原始距离的计算时间 ST_{fdis}:$ST_I=ST_F+ST_{index}+ST_{vdis}+ST_{fdis}+ST_{I/O}$,实际上,由于实际参加距离计算的量化数据和原始数据比较少,使得它们相对于 ST_{DIS} 为整个检索过程节省下来的距离计算时间远远超过了花费在索引文件上的时间 ST_{index},从而使得检索时间明显减少。为了突出这个优点,分别计算了两种图库特征下顺序检索的 ST_{DIS} 和本文方法的 $ST_{index}+ST_{vdis}+ST_{fdis}$ 来说明。(量化程度均为 2 bits/维)。可以看出,索引机制下的检索时间比顺序检索有明显的减少(见图 4)。

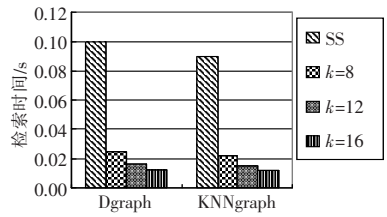

图 4 两种特征的不同检索方式的检索时间

在大幅度减少检索时间的同时,检索精度也是不容忽视的。在没有一定的检索精度下减少检索时间是没有意义的。表 3 是对检索精度的实验结果。

表 3 不同检索方式的两种特征下的平均准确率

检索方式	准确率/%	
	Dgraph	KNNgraph
k=8	45.23	69.66
k=12	45.96	75.12
k=16	54.43	82.93

从表 3 可以看出,在本文引用的方法上并没有明显影响准确率。

4 结论

本文结合混合模型自动选择与索引机制对图库进行索引,在不影响准确率的情况下,有效地提高了检索效率。该方法为解决未知图库具体分布高维数据的索引等问题提供了一个有效方案。同时,模型如何完全自动选择以及如何自动选择数据库的量化率(量化数据与原始数据的存储比率),使其能最大程度地体现图特征的信息仍然是值得研究的问题。

(收稿日期:2007 年 2 月)

参考文献:

[1] Fayyad U,Piatetsky-Shpiro G,Smyth P,et al.Advances in knowledge discovery and data mining[M].[S.l.]:MIT Press,1996.

[2] Lim Y W,Lee S U.On the color image segmentation algorithm based on the thresholding and the fuzzy C-means techniques[J].Pattern Recognition,1990,23(9):935-952.

[3] Uchiyama T,Arib M A.Color image segmentation using competitive learning[J].IEEE Trans Pattern Analysis and Machine Intelligence,1994,16(12).

[4] Banfield J,Raftery A.Model-based gaussian and non-gaussian clustering[C]//Biometrics,1993,49:803-821.

[5] MacQueen J B.Some methods for classification and analysis of multivariate observations[C]//Proc Fifth Berkeley Symp Math Statistics and Probability.Berkeley,Calif:Univ of California Press,1967:281-297.

[6] Dempster A P,Laird N M,Rubin D B.Maximum likelihood from incomplete data via the EM algorithm[J].J Roya Statistical Soc:Series B,1977,39(1):1-38.

(下转 169 页)