利用高斯混合模型实现概率密度函数逼近

袁礼海[1],李　钊[2],宋建社[1]
(1.第二炮兵工程学院 导弹工程研究所,陕西 西安 710025;
2.驻石家庄地区军事代表室,河北 石家庄 050081)

摘　要:针对图像的概率分布密度函数的不确定,利用有限高斯混合模型逼近图像的概率分布密度函数。理论上证明了有限高斯混合模型可以以任意精度正逼近实数上的非负黎曼可积函数,特别可以逼近任意的概率分布密度函数。实例表明有限高斯混合模型逼近已知分布密度函数或未知分布密度函数时,具有逼近精度高等优点,为函数逼近提供了理论和技术支持。

关键词:高斯混合模型;函数逼近;概率密度函数;高斯分布
中图分类号:TP301.6　　**文献标识码**:A　　**文章编号**:1003-3114(2007)02-20-3

Probability Density Function Approximation Using Gaussian Mixture Model

YUAN Li-hai[1], LI Zhao[2], SONG Jian-she[1]
(1. Institute of Missile Engineering of the Second Artillery Engineering Institute, Xi' an Shanxi 710025, China;
2. The Military Representative Office in Shijiazhuang, Shijiazhuang Hebei 050081, China)

Abstract: With the analysis of probability density functions, it can be approximated using Gaussian mixture model. Nonnegative Riemann integrable function in R space, especially probability density function, can be approximated with arbitrary precision by finite sum of Gaussian density function with different parameters. This can be proved to be effective. The computational examples show that this approach can obtain high accuracy such that it can provide new theoretical and technical supports for function approximation.

Key words: Gaussian mixture model; function approximation; probability density function; Gaussian distribution

0　引言

在图像处理中,经常需要知道图像的概率分布密度函数,然而图像的概率密度函数经常是很难准确求解的,通常的办法是通过某个已知的分布密度函数进行逼近,例如:高斯(Gaussian)分布、对数正态分布、伽玛分布、贝塔分布、指数分布、韦布尔分布、瑞利分布等,然而这类参数化分布密度要求是单峰形式,即只有一个极大值,而实际问题中,可能包含多峰的密度形式,在特征空间中往往表现为多种密度分布的混合,很难把这种复杂的分布通过单一的参数化密度函数表示出来[1]。

R.Wilson 在文献[2]中讨论了多分辨率高斯混合模型的函数逼近能力。在此基础上本文证明了有限高斯混合模型可以以任意精度正逼近实数上的非负黎曼可积函数,特别可以逼近任意的概率分布密度函数,并利用实例说明了有限高斯混合密度函数具有较强的逼近能力。

基金项目:国家自然科学基金(60272022)
收稿日期:2006-10-10
作者简介:袁礼海(1977-),男,工程师。主要研究方向:SAR 图像处理、信号分析等。

1　有限混合密度函数

在图像处理领域,有限混合分布理论的方法就是将全部像素值拟合成一个加权混合的概率密度函数,使每个权重正是该对象的像素在整个像素集里所占的比例。高斯模型涉及均值(μ)和方差(σ^2)的选择。

定义 1　有限混合密度模型

假设数据 $x(x \in R^p)$ 来自多个分布的混合体。那么,其概率密度就可表示为:

$$f(x;\theta) = \sum_{i=1}^{g} \alpha_i f_i(x;\theta_i), \qquad (1)$$

式中,g 为密度分支的个数;$\alpha_1, \cdots, \alpha_g$ 是各混合密度分支的权重,$\sum_{i=1}^{g} \alpha_i = 1, (\alpha_i \geq 0, i = 1, \cdots, g)$;$f_i$ 是第 i 个类别的密度;θ_i 是相应类别的未知参数,整个混合密度的参数为 $\theta = (\alpha_1, \cdots, \alpha_g, \theta_1, \cdots, \theta_g)$。通常采用基于期望最大化(简称 EM)算法的最大似然

模型去估计概率函数参数。理论上，增加密度分量的个数能提高逼近程度，但这样会增加个别参数求解的难度[3]。

2 有限混合密度函数的逼近能力

有限高斯混合模型可以以任意精度正逼近 R 上的非负黎曼可积函数，特别可以逼近任意的概率分布密度函数。在文献[4]的基础上，重新给出了相应的证明。

定义 2 “脉冲性”

考虑带参变量 θ 的概率分布密度函数 $p(x;\theta)$，其中，参变量 $\theta=\theta(\varepsilon)$，$\varepsilon$ 为正实数，若 $\varepsilon\to 0$ 时，有 $\int_{-\varepsilon}^{\varepsilon}p(x;\theta)\mathrm{d}x=1$，则称 $p(x;\theta)$ 具有“脉冲性”。

引理 1 满足条件 $\lim\limits_{\varepsilon\to 0}\dfrac{\varepsilon^2}{\sigma^2}=\infty\ (\varepsilon>0)$ 的高斯分布函数具有“脉冲性”。

证明：令

$$A=\int_{-a}^{a}\mathrm{e}^{-x^2}\mathrm{d}x, \tag{2}$$

于是

$$A^2=\int_{-a}^{a}\int_{-a}^{a}\mathrm{e}^{-(x^2+y^2)}\mathrm{d}x\mathrm{d}y。 \tag{3}$$

作坐标变换 $x=r\cos\theta$，$y=r\sin\theta$，则：

$$A^2\geq\int_0^{2\pi}\int_0^{a}r\mathrm{e}^{-r^2}\mathrm{d}r\mathrm{d}\theta=(1-\mathrm{e}^{-a^2})\pi, \tag{4}$$

有：

$$\int_{-\varepsilon}^{\varepsilon}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{x^2}{2\sigma^2}}\mathrm{d}x=\int_{-\frac{\varepsilon}{\sqrt{2}\sigma}}^{\frac{\varepsilon}{\sqrt{2}\sigma}}\frac{1}{\sqrt{\pi}}\mathrm{e}^{-x^2}\mathrm{d}x\geq\sqrt{1-\mathrm{e}^{-\frac{\varepsilon^2}{2\sigma^2}}}。 \tag{5}$$

由上式，若选择恰当的 σ，如 $\sigma=\varepsilon^2$，满足 $\lim\limits_{\varepsilon\to 0}\dfrac{\varepsilon^2}{\sigma^2}=\infty$，则：

$$1\geq\lim_{\varepsilon\to 0}\int_{-\varepsilon}^{\varepsilon}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{x^2}{2\sigma^2}}\mathrm{d}x\geq\lim_{\varepsilon\to 0}\sqrt{1-\mathrm{e}^{-\frac{\varepsilon^2}{2\sigma^2}}}=1,$$

即：

$$\lim_{\varepsilon\to\infty}\int_{-\varepsilon}^{\varepsilon}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{x^2}{2\sigma^2}}\mathrm{d}x=1。 \tag{6}$$

根据定义，可知引理结论成立。

定理 1 有限高斯混合模型严格正逼近任意非负可积函数。

设 $p_\theta(x)$ 是含参数为 θ 的高斯分布密度函数。对任意非负可积函数 $f(x)$ 有如下关系：对任意 $\delta>0$，存在正整数 N，使得 $\int_R\left|f(x)-\sum\limits_{i=1}^{N}\pi_i p_{\theta_i}(x)\right|\mathrm{d}x<\delta$，$\pi_i>0$，$\sum\limits_{i=1}^{N}\pi_i=1$，则称 $p_\theta(x)$ 有限混合模型严格正逼近函数 $f(x)$。

证明：下面分 3 步来加以证明。

第 1 步，$p_\theta(x)$ 的有限混合模型严格正逼近函数 $p_\theta * \chi_X(x)$，$\chi_X(x)$ 为量化示性函数。

$$\chi_X(x)=\begin{cases}\dfrac{1}{2\varepsilon}, & \text{若 } x\in X,\ X=[-\varepsilon,\varepsilon],\ \varepsilon \text{ 为正实数。}\\ 0, & \text{其他}\end{cases} \tag{7}$$

由卷积定义：

$$p_\theta * \chi_X(x)=\frac{1}{2\varepsilon}\int_X p_\theta(x-y)\mathrm{d}y=\lim_{N\to\infty}\sum_{i=1}^{N}\frac{\sigma\gamma_i}{2\varepsilon}p_\theta(x-y_i)=\lim_{N\to\infty}\sum_{i=1}^{N}\pi_i p_\theta(x-y_i), \tag{8}$$

这里，$\pi_i=\dfrac{\sigma\gamma_i}{2\varepsilon}>0$，$\sum\limits_{i=1}^{N}\pi_i=1$。

第 2 步，$p_\theta * \chi_X(x)$ 任意逼近 $\chi_X(x)$。

令 $\eta=\int_R\left|\chi_X(x)-p_\theta * \chi_X(x)\right|\mathrm{d}x$，

$$\eta=\int_X\left|\chi_X(x)-p_\theta(x) * \chi_X(x)\right|\mathrm{d}x+\int_{R-X}\left|\chi_X(x)-p_\theta(x) * \chi_X(x)\right|\mathrm{d}x=\int_{-\varepsilon}^{\varepsilon}\left|\frac{1}{2\varepsilon}-\frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon}p_\theta(x-y)\mathrm{d}y\right|\mathrm{d}x+\int_{R-X}\left|\frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon}p_\theta(x-y)\mathrm{d}y\right|\mathrm{d}x=\frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon}\left[1-\int_{-\varepsilon}^{\varepsilon}p_\theta(x-y)\mathrm{d}y\right]\mathrm{d}x+\frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon}\int_{R-X}p_\theta(x-y)\mathrm{d}x\mathrm{d}y,$$

则由

$$\int_{R-X}p_\theta(x-y)\mathrm{d}x=\int_R P_\theta(x-y)\mathrm{d}x-\int_X p_\theta(x-y)\mathrm{d}x=1-\int_{-\varepsilon}^{\varepsilon}p_\theta(x-y)\mathrm{d}x, \tag{9}$$

有：

$$\eta=2\times\frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon}\left[1-\int_{-\varepsilon}^{\varepsilon}p_\theta(x-y)\mathrm{d}y\right]\mathrm{d}x=2\left[1-\frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon}\int_{-\varepsilon}^{\varepsilon}p_\theta(x-y)\mathrm{d}y\mathrm{d}x\right]=2\left[1-\frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon}\int_{x-\varepsilon}^{x+\varepsilon}p_\theta(t)\mathrm{d}t\mathrm{d}x\right]=2\left[1-\int_{-\varepsilon}^{\varepsilon}p_\theta(t)\mathrm{d}t\right]\to 0(\sigma=\varepsilon^2;\varepsilon\to 0), \tag{10}$$

从而有：

$$\int_R\left|\chi_X(x)-p_\theta * \chi_X(x)\right|\mathrm{d}x\to 0。 \tag{11}$$

第 3 步，$\chi_X(x)$ 的有限混合模型严格正逼近任意非负 R 可积函数 $f(x)$。

引入 δ 函数，有 $\lim\limits_{\varepsilon\to 0}\chi(x)=\delta(x)$，

$$f(x;\theta)=\int_R f(y;\theta)\delta(x-y)\mathrm{d}y=\int_R f(y;\theta)\lim_{\varepsilon\to 0}\chi(x-y)\mathrm{d}y=\lim_{\varepsilon\to 0}\int_R f(y;\theta)\chi(x-y)\mathrm{d}y。 \tag{12}$$

3 应用举例

3.1 Rayleigh 分布密度函数逼近

现以 3 个高斯概率密度函数逼近 Rayleigh 分布的概率密度函数来阐述有限高斯混合密度函数的逼近能力。如图 1 所示，Rayleigh 分布的概率密度函数是不对称的，因而采用单个的高斯分布来逼近是很难完成的，而 3 个高斯密度函数能够很好地逼近原始分布。Rayleigh 概率密度函数为：

$$y_R = f(x;b) = \frac{x}{b^2}\exp(\frac{-x^2}{2b^2}), \quad (13)$$

这里，$b = 0.585$。而高斯分布的概率密度函数为：

$$y_G = f(x;\theta) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}), \quad (14)$$

式中，参数 $\theta = \{\mu,\sigma\}$，相应有 $\theta_1 = \{0.48,0.23\}$，$\theta_2 = \{0.83,0.35\}$，$\theta_3 = \{0.82,0.6\}$。逼近的公式表示为：

$$y_R \approx 0.3y_{G1} + 0.5y_{G2} + 0.2y_{G3}。 \quad (15)$$

图 1 中实线代表 Rayleigh 分布，短虚线代表有限混合分布，长虚线代表 Gaussian 分布，短虚线和实线的拟合程度表明通过 3 个高斯分布函数的加权和能够很好地对 Rayleigh 分布函数进行逼近。

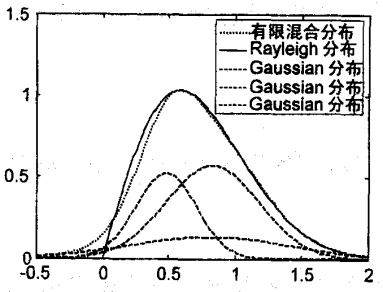

图 1 有限混合分布逼近 Rayleigh 分布示意图

3.2 图像概率分布密度函数的逼近

有限混合密度函数可以很好地逼近任意的连续分布密度函数，甚至可以估计未知概率密度函数。通过选择适当的分布作为密度分量来描述真实分布的局部区域，建立混合模型来近似复杂分布，这样就可以很好地从已知的观测数据来估计真实分布的局部特性，这是用单参数分布描述无法做到的。

合成孔径雷达(SAR)图像研究中，国内外一些专家常常预先假设各类别在特征空间的密度分布服从高斯密度分布[1]。实验中选用 Ku 波段机载 SAR 图像，图像中主要存在 3 类目标，即池塘、农田和树林。采用 3 个高斯混合密度函数去逼近 SAR 图像的密度函数，利用 EM 算法进行高斯混合密度函数的参数估计，最终的模型参数如表 1 所示，其中包括 3 种类别的权重、均值和方差。已知上述参数以后，可以很方便绘制出 3 种类别的高斯混合分布密度，再通过原始图像绘制出灰度值分布密度，2 条分布密度曲线如图 2 所示。带“+”的线表示有限高斯混合分布密度，另一条为灰度值分布密度，即灰度直方图。2 条曲线基本吻合表明有限高斯混合分布密度模型能够较好地逼近 SAR 图像的分布密度函数。

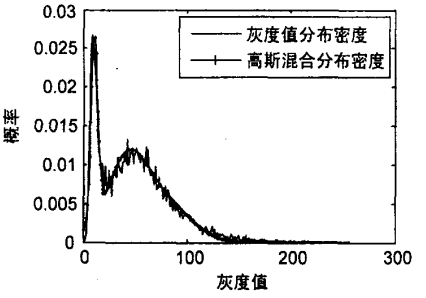

图 2 有限高斯混合分布密度逼近 SAR 图像的灰度值分布密

表 1 有限高斯混合分布密度模型参数

参数	α_1	α_2	α_3
参数值	0.28	0.19	0.53
参数	μ_1	μ_2	μ_3
参数值	80.03	9.50	41.27
参数	σ_1^2	σ_2^2	σ_3^2
参数值	600.8	10.37	400.21

4 结束语

本文研究了有限高斯混合分布密度逼近概率分布密度理论，并证明了其正确性。利用有限高斯混合分布密度逼近 Rayleigh 概率密度函数；在未知概率分布密度时，以 SAR 图像为例，使有限高斯混合分布密度能准确地逼近任意 SAR 图像分布密度，这为图像概率分布密度函数的逼近提供了一条有效的途径。

参 考 文 献

[1] O'SULLIVAN J A, DEVORE M D, Kedia V, et al. SAR ATR Performance Using a Conditionally Gaussian Model [J]. IEEE Transactions on Aerospace and Electronic Systems, 2001, 37(1):91 - 105.

[2] WILSON R. Multiresolution Gaussian Mixture Models: Theory and Application[R]. Research Report RR404, Department of Computer Science, University of Warwick, UK, 1999:1 - 10.

[3] BILMES J. A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models [R]. Department of Electrical Engineering and Computer Science, U. C. Berkeley, TR - 97 - 021, 1998:1 - 13.

[4] 连惠城. 多尺度随机模型与 SAR 图像无监督分割[D]. 西北工业大学硕士学位论文, 2004:11 - 18.

利用高斯混合模型实现概率密度函数逼近

作者： 袁礼海， 李钊， 宋建社， YUAN Li-hai， LI Zhao， SONG Jian-she
作者单位： 袁礼海,宋建社,YUAN Li-hai,SONG Jian-she(第二炮兵工程学院导弹工程研究所,陕西,西安,710025)， 李钊,LI Zhao(驻石家庄地区军事代表室,河北,石家庄,050081)
刊名： 无线电通信技术
英文刊名： RADIO COMMUNICATIONS TECHNOLOGY
年，卷(期)： 2007,33(2)
被引用次数： 12次

参考文献(4条)

1. O'SULLIVAN J A;DEVORE M D;Kedia V SAR ATR Performance Using a Conditionally Gaussian Model 2001(01)

2. WILSON R Multiresolution Gaussian Mixture Models:Theory and Application[Research Report RR404,Department of Computer Science,University of Warwick] 1999

3. BILMES J A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models[Department of Electrical Engineering and Computer Science,U.C.Berkeley,TR -97-021] 1998

4. 连惠城 多尺度随机模型与SAR图像无监督分割[学位论文] 2004

引证文献(7条)

1. 黄洪全,方方,龚迪琛,丁卫撑 基于HMM双重模型的伽马能谱漂移模拟[期刊论文]-物探与化探 2010(04)

2. 周里,杨承志 面向节能优化的加热钢坯特征参数模型研[期刊论文]-自动化技术与应用 2011(04)

3. 庄亚强,张晨新,张小宽,周超 基于高斯混合密度模型的隐身目标RCS统计分析[期刊论文]-空军工程大学学报（自然科学版） 2014(02)

4. 徐晓旻,肖仰华 KBAC：一种基于K-means的自适应聚类[期刊论文]-小型微型计算机系统 2012(10)

5. 徐晓旻,肖仰华 KBAC：一种基于K-means的自适应聚类[期刊论文]-小型微型计算机系统 2012(10)

6. 黄玲 面向机器解译的遥感图像质量评价关键技术研究[学位论文]硕士 2013

7. 江佩斯 多谐波源随机谐波电流叠加问题的研究[学位论文]硕士 2008