第35卷 第15期 计 算 机 工 程 2009年8月
Vol.35 No.15 Computer Engineering August 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)15—0187—03 文献标识码：A 中图分类号：TP391

利用高斯域的半监督回归和主动学习

崔 鹏[1,2]，张汝波[1]
(1. 哈尔滨工程大学计算机学院，哈尔滨 150001；2. 哈尔滨理工大学计算机学院，哈尔滨 150080)

摘 要：介绍一种定义近邻图上的高斯域(GF)及用于降维和分类的GF的相关知识，提出一种用于半监督回归的高斯域，能自动设置模型参数和近邻数，利用监督和无监督数据进行熵值查询选择从而进行主动学习。实验将其与半监督学习法进行比较并验证了GF的有效性。
关键词：高斯域；半监督回归；主动学习；熵；Cholesky分解

Semi-Supervised Regression and Active Learning with GF

CUI Peng[1,2], ZHANG Ru-bo[1]
(1. Department of Computer, Harbin Engineering University, Harbin 150001;
2. Department of Computer, Harbin University of Science & Technology, Harbin 150080)

【Abstract】 A Gaussian Fields(GF) on nearest neighbor graph is defined by using a non-parametric technique. On the basis of it, a MAP criterion which can automatically set model parameter and numbers of nearest-neighbor k is proposed and entropy maximization query selection method for active learning by using supervised and unsupervised information is specified. Experimental results demonstrate effectiveness of GF compared with semi-active learning method.
【Key words】 Gaussian fields(GF); semi-supervised regression; active learning; entropy; Cholesky decomposition

学习高维数据是一个具有挑战性的问题，已变得越来越重要。本文通过能量函数得到一种高斯域(Gaussian Fields,GF)模型，能量化预测中的不确定性对半监督回归以及主动学习有很大的实际意义。

1 背景及相关工作

1.1 定义在近邻图上的GF

给定高维输入集 $X=\{x_1, x_2, \cdots, x_n\}$，定义GF的第1步是找到 x_i 的 k 个最近的邻接点。本文的邻接点间测度都是在欧式距离基础上定义的。

用 y_i 表示 x_i 的一些标量输出，$\boldsymbol{y}=(y_1, y_2, \cdots, y_n)^{\mathrm{T}}$ 表示包含所有输出的 n 维矢量。已知最近邻接点，能量函数 $E(\boldsymbol{y})$ 为所有相邻最近的输入点对 x_i 和 x_j，它们输出的平方差乘以带权因子 a_{ij} 的和。若 $a_{ij}\neq 0$ 表示最近邻接，那么：

$$E(\boldsymbol{y})=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}(y_i-y_j)^2=\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{y}=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{y} \tag{1}$$

其中，矩阵 $\boldsymbol{D}$ 是邻接矩阵 $\boldsymbol{A}$ 行数和的对角阵；矩阵 $\boldsymbol{L}$ 为矩阵 $\boldsymbol{A}$ 的图Laplacian算子，并且 $\boldsymbol{L}$ 是正定的[1]。可通过增加一个正则化因子确保正定性：

$$E(\boldsymbol{y})=\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{L}+\alpha\boldsymbol{I})\boldsymbol{y}=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{y} \tag{2}$$

由此可见，$\boldsymbol{M}$ 和 $\boldsymbol{L}$ 有相同的特征向量，其特征值随 α 增加。若 $\boldsymbol{L}$ 为逆协方差阵，正则化因子以 α^{-1} 变化代替了GF密度方向1上的无限变化。因此，给定点集 X，在能量方程式(2)基础上，可以定义输出 $\boldsymbol{y}$ 的GF密度为

$$p(\boldsymbol{y})=N(\boldsymbol{y};0,\beta^{-1}\boldsymbol{C})\propto\exp-\frac{\beta}{2}E(\boldsymbol{y}) \tag{3}$$

其中，β 是控制密度平滑性的比例参数；$\boldsymbol{C}$ 表示 $\boldsymbol{M}$ 的逆。虽然逆特征矩阵是稀疏的，但是它常能推导出一个全特征矩阵 $\beta^{-1}\boldsymbol{C}$。

所有输出的联合密度可推出输出集的条件概率和边缘概率，由于半监督学习中有些数据 x_i 的输出是已知的，因此会固定一些值。不失一般性，假设监督点的所有指标都小于无监督的点，则 $\boldsymbol{y}=[\boldsymbol{y}_s^{\mathrm{T}}\ \boldsymbol{y}_u^{\mathrm{T}}]^{\mathrm{T}}$，$\boldsymbol{y}_u$ 表示无监督输出的子向量，$\boldsymbol{y}_s$ 表示监督输出的子向量。$\boldsymbol{M}_{ss}$, $\boldsymbol{M}_{us}$ 和 $\boldsymbol{M}_{uu}$ 表示相应 $\boldsymbol{M}$ 的块[2]。

$$\boldsymbol{M}=\begin{pmatrix}\boldsymbol{M}_{ss} & \boldsymbol{M}_{su}\\ \boldsymbol{M}_{ss} & \boldsymbol{M}_{uu}\end{pmatrix} \tag{4}$$

式(2)能够扩展为

$$E(\boldsymbol{y})=\boldsymbol{y}_u^{\mathrm{T}}\boldsymbol{M}_{uu}\boldsymbol{y}_u+\boldsymbol{y}_s^{\mathrm{T}}\boldsymbol{M}_{ss}\boldsymbol{y}_s+2\boldsymbol{y}_u^{\mathrm{T}}\boldsymbol{M}_{us}\boldsymbol{y}_s \tag{5}$$

显然，条件概率密度为

$$p(\boldsymbol{y}_u\mid\boldsymbol{y}_s)\propto\exp-\frac{\beta}{2}(\boldsymbol{y}_u^{\mathrm{T}}\boldsymbol{M}_{uu}\boldsymbol{y}_u+2\boldsymbol{y}_u^{\mathrm{T}}\boldsymbol{M}_{us}\boldsymbol{y}_s) \tag{6}$$

通过计算此条件概率密度的均值 $\boldsymbol{y}_u$* 预测无监督输出：

$$\boldsymbol{y}_u^{*}=-\boldsymbol{M}_{uu}^{-1}\boldsymbol{M}_{us}\boldsymbol{y}_s \tag{7}$$

为解式(9)，需用(稀疏)Cholesky分解 $\boldsymbol{M}_{uu}=\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}$，首先解 $\boldsymbol{R}^{\mathrm{T}}\boldsymbol{z}=-\boldsymbol{M}_{us}\boldsymbol{y}_s$，然后解 $\boldsymbol{R}\boldsymbol{y}_u^{*}=\boldsymbol{z}$。几个需要评估的量用Cholesky因子以类似的方法求得。

1.2 用于降维和分类的GF

对于2类-半监督分类，$\boldsymbol{y}_s$ 中的输出值为0或1，表示类隶属关系。分类是通过最小化定义在式(1)中的能量函数 $E(\boldsymbol{y})$ 进行的，$\boldsymbol{y}_u$ 固定为 $\boldsymbol{y}_s$[3]。极小值 $\boldsymbol{y}_u^{*}$ 由线性方程特征化：

$$\boldsymbol{L}_{uu}\boldsymbol{y}_u^{*}=-\boldsymbol{L}_{us}\boldsymbol{y}_s \tag{8}$$

若在 $\boldsymbol{y}_u$* 中的未标记点小于1/2，则它们会被分配到类0；否则会被分到类1。若忽视 α 调整，则这种方法算得的条件概率 $p(\boldsymbol{y}_u|\boldsymbol{y}_s)$ 均值和条件均值的初值都为1/2。

2 基于GF的半监督回归和主动学习

利用GF进行主动半监督回归主要包括2个问题：(1)能

作者简介：崔 鹏(1971－)，男，副教授、博士研究生，主研方向：机器学习，数据挖掘；张汝波，教授、博士生导师
收稿日期：2009-01-11 **E-mail：**cuipeng83@163.com

自动设置模型参数 β 和 k 的最大后验(MAP)标准；(2)用于主动学习的最大熵值查询选择过程。

2.1 MAP 标准

GF 密度由 k——连接到每点的近邻数及可变比例参数 β 进行参数化。由于很难手工设置这些参数，因此希望能自动找到最优值，利用 MAP 标准为 k 和 β 选定合适值。

若给定 $\boldsymbol{y}_s$ 中的监督输出，可得到基于参数 k,β 的后验分布： $p(k,\beta\,|\,\boldsymbol{y}_s)\propto p(\boldsymbol{y}_s\,|\,k,\beta)p(k,\beta)$。

由于 $p(\boldsymbol{y}\,|\,\beta,k)$ 是高斯的，因此可以相对容易地获得监督点的边缘似然性。用 n_s 表示监督点数，给定 k 和 β，$\boldsymbol{y}_s$ 表示边缘对数似然性:

$$\lg p(\boldsymbol{y}_s\,|\,\beta,k)=-\frac{1}{2}[\lg|\boldsymbol{C}_{ss}|-n_s\lg\beta+\beta\boldsymbol{y}_s^{\mathrm{T}}\boldsymbol{C}_{ss}^{-1}\boldsymbol{y}_s] \tag{9}$$

其中，$\boldsymbol{C}_{SS}$ 依赖于 k，通过求导，得到关于 β 的边缘对数最大值，最大值时 β 表示为 β^*:

$$\beta^*=\arg\max_{\beta}\lg p(\boldsymbol{y}_s\,|\,\beta,k)=n_s/(\boldsymbol{y}_{ss}^{\mathrm{T}}\boldsymbol{C}_{ss}^{-1}\boldsymbol{y}_s) \tag{10}$$

用其替换式(9)中的 β，则:

$$l^*=\lg p(\boldsymbol{y}_s\,|\,\beta^*,k)=-\frac{1}{2}[\lg|\boldsymbol{C}_{ss}|+n_s+n_s\lg(\boldsymbol{y}_s^{\mathrm{T}}\boldsymbol{C}_{ss}^{-1}\boldsymbol{y}_s/n_s)] \tag{11}$$

对邻居数 k*最佳值的选取需在 k={1, 2, …, $k_{\max}$}范围内，选取式(11)最大时的 k 值。关于正则化算子 α 的边缘似然性最大化不能在封闭形式下进行，但可以采用基于梯度下降法。在实验中，令 α 取值为 10^{-12}。

2.2 主动查询选择

实际上可用的监督数据通常是有限的，由于其获取代价很高，因此启发人们使用基于无监督和监督数据集确定对无监督数据和监督信号更有用的主动学习方法。

为了确定哪些固定大小的查询集是最有用的信息，需要考虑给定已标记查询集条件下无标记点条件密度的熵，以得到能导致最低条件熵值的查询集；同时希望在给定标记点情况下对无标记点预测的不确定性尽可能小。其基本思想是：如果假定域均值接近相应变量的真值，那么分布 $p(\boldsymbol{y}_u|\boldsymbol{y}_s)$应尽可能紧密地围绕在其均值周围[4]。

熵的链规则 $H(\boldsymbol{y})=H(\boldsymbol{y}_s)+H(\boldsymbol{y}_u|\,\boldsymbol{y}_s)$。由于完全域的熵 $H(\boldsymbol{y})$ 是固定的，因此为了计算 $H(\boldsymbol{y}_s)=\frac{1}{2}\lg|\boldsymbol{C}_{ss}|-\frac{1}{2}n_s\lg\beta+const$，采用前述 $\boldsymbol{M}$ 的 Cholesky 因数以找到 $\boldsymbol{C}_{ss}$。为找到带有最大熵的单变量，需计算 $\boldsymbol{C}$ 的对角元素的所有边缘变化。而从所有变量的一个随机子集中选取的有局部最大方差的变量，很可能也是全局最大方差的变量。给定一些初始查询集，随机地选择不在当前查询集中的一些变量 $\boldsymbol{y}_i$，直到满足停止的条件。如果用 $\boldsymbol{y}_i$ 替换一个当前查询集中的变量产生查询集的一个更高联合熵，就产生了一次交换。可用一个随机子集或贪婪过程产生初始查询集，逐渐将新元素添加到查询集中。

在查询集大小未知的情况下，当用户要对另一数据点进行标记时，也可通过自身采用贪婪过程确定能最优地扩充当前查询集的查询。已知初始查询集 $s\subseteq\{1,2,\mathrm{L},N\}$，要找到最佳查询 i 加入 s，$i\in\{1,2,\cdots,N\}\backslash s$，以使其余变量中的熵 $H(\boldsymbol{y}_u\backslash i\,|\,\boldsymbol{y}_{\{i\}\cup s})=H(\boldsymbol{y}_u\,|\,\boldsymbol{y}_s)-H(\boldsymbol{y}_i|\,\boldsymbol{y}_s)$ 最小，就需找到 $H(\boldsymbol{y}_i\,|\,\boldsymbol{y}_s)$ 为最大值时的 $i\in u$，用 $\boldsymbol{M}_{uu}$ 的 Cholesky 分解计算这些熵值。此分解也可用于从 $\boldsymbol{y}_s$ 中预测 $\boldsymbol{y}_u$。在此情况下，$\boldsymbol{M}_{uu}$ 的 Cholesky 分解是不可用的(因为 $\boldsymbol{y}_u$ 不可预测或不用 Cholesky 分解预测)，而可采用 $\boldsymbol{M}$ 分解。由熵的链规则，有 $H(\boldsymbol{y}_u\backslash i\,|\,\boldsymbol{y}_{\{i\}\cup s})=H(\boldsymbol{y})-H(\boldsymbol{y}_{\{i\}\cup s})$。将找到的最佳查询加入 s 中，发现与 s 一起的查询 i 在域中有最大熵值 $H(\boldsymbol{y}_{\{i\}\cup s})$。熵值 $H(\boldsymbol{y}_{\{i\}\cup s})$ 能从 $\boldsymbol{M}$ 的 Cholesky 因子中算出。

3 实验结果与分析

本文用计算机图形仪柔化了 660 幅脸部图像，对脸部从不同方向进行透视和照明，且所有图像的透视方向和照明方向是可用的。任务是预测无监督图像的透视方向和照明方向。实验使用的图像中有一些是有监督的。可以通过随机或主动学习方法选择监督图像。预测值与实际值间的方差为无监督点与不同响应变量的均值，所有实验结果都是 20 个实验的平均值。

3.1 查询选择

从图 1 中可以看到采用随机查询选择(实线)和主动查询选择(点线)得到的关于模型中邻居数 k 的均方差(刻度为 log 值)。不同的线(从上到下)表示了 10 个、20 个和 100 个已标记点的结果。本文采用贪婪查询选择法进行查询选择。

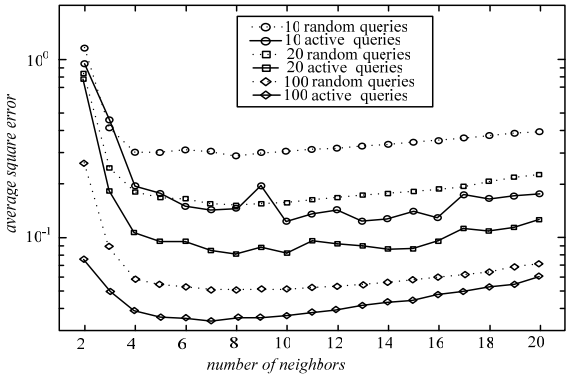

图 1 无监督点邻居数的误差

对有不同连通性的近邻图进行了实验：k=2, 3, …, 20。在所有 k 及大量已标记点(监督点)的前提下，在 57 个合并中除了 6 个外，主动学习算法要明显优于随机查询，在主动学习中标准差更小。

3.2 模型选择

查询选择依赖于要使用的邻居数目，(利用 MAP 标准)模型选择依赖于监督查询。为用查询选择合并模型选择，本文使用了一个贪婪过程：首先标记一个随机查询，然后轮流使用已标记点选择 k，选择依赖于当前 k 和已标记点的新查询。若选择随机查询，则直接使用已标记点进行模型选择。

表 1 为采用 20 个实验的最优 k 值和通过模型选择 k 值的平均结果。当加入主动和随机查询时，k 值为获得最小预测差的邻居数。

表 1 模型选择与最优 k 值比较

均方差	已标记点数		
	10	20	100
随机查询最优 k	0.25	0.14	0.06
随机查询模型选择	0.30	0.15	0.06
主动查询最优 k	0.13	0.09	0.04
主动查询模型选择	0.31	0.10	0.04

使用随机查询，模型选择接近于最优：最佳值 k 不会给出比自动选择的 k 更好的结果。然而，对于主动查询，在仅有 10 个监督点时，最佳 k 值的结果明显更好，因为对于少量的监督点，模型选择了过小的 k，使查询选择过程执行得较差。比较图 1 中主动和随机查询可见，对于 10 个查询，当 k 较小时，如果使用模型选择，主动查询和随机查询间没有太

大的不同。对于更多的查询，主动查询显得好一些。

3.3 GF 法与半监督分类学习法比较

将 GF 法与半监督分类的阶段学习法进行比较：使用与前面实验相同的数据集，利用随机选择的监督点重复 20 次，计算不同数量监督点及图中近邻数为 k 时剩余无监督点的残差。对于阶段学习法，需设置要用到的特征向量数 m(斜线后数字)。

表 2 给出特征向量最佳数结果 $m \in \{1,2,\cdots,20\}$，即导致最小平均误差的数量。若性能上有明显的不同则结果用黑体表示。当 $n_s \in \{10, 20\}$时，2 种方法间性能的差异不很明显。当 n_s=100 个监督点时，GF 法明显更好一些。GF 法无须选择特定特征向量数，可以少用一个参数。

表 2　使用 GF 和 SL 法获得的残差

监督点数	方法	k 值				
		6	10	12	16	20
10	GF	0.45±2.0	0.37±1.3	0.43±2.1	0.41±2.6	0.32±1.5
	SL	0.63/4±4.0	0.33/4±1.8	0.29/4±1.3	0.33/4±2.4	0.44/4±10.1
20	GF	**0.22**±0.5	0.19±0.5	0.20±0.8	0.18±0.6	0.18±0.4
	SL	0.31/7±1.2	0.16/7±0.6	0.29/4±0.3	0.33/4±0.9	0.44/4±0.6
100	GF	**0.08**±0.1	**0.08**±0.1	**0.07**±0.1	**0.07**±0.1	**0.08**±0.1
	SL	0.10/17±0.1	0.09/8±0.1	0.09/7±0.0	0.09/11±0.1	0.10/9±0.1

4　结束语

本文说明了基于稀疏近邻图的 GF 用于半监督回归的方法。提出了 MAP 标准以及用于主动学习的最大熵值查询选择过程，下一步工作要研究如何将半监督学习、主动学习与流形学习结合起来设计一种分类器。

参考文献

[1] Belkin M, Niyogi P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation[J]. Neural Computer, 2003, 15(6): 1373-1396.

[2] Tenenbaum J B, de Silva V, Langford J C. A Global Geometric Framework for Nonlinear Dimensionality Reduction[J]. Science, 2000, (290): 2319-2323.

[3] Zhu Xiaojun, Lafferty J, Ghahramani Z. Semi-supervised Learning Using Gaussian Fields and Harmonic Functions[C]//Proceedings of the International Conference on Machine Learning. California, USA: AAAI Press, 2003.

[4] Demsar J. Statistical Comparisons of Classifiers over Multiple Data Sets[J]. Journal of Machine Learning Research, 2006, 7(1): 532-543.

编辑　张正兴

（上接第 184 页）

4　结束语

本文针对 AFSA 存在的不足，利用公告板中的历史最优鱼和高斯变异的优点，提出基于自适应高斯变异的人工鱼群算法，用不同的测试函数对 AGMAFSA 和 AFSA 进行仿真试验、实例应用和理论分析，结果表明 AGMAFSA 中人工鱼能有效摆脱局部极值的束缚，同时也能加快搜索速度，使算法最终快速收敛于全局极值，且得到高精度的解，算法更加稳定。下一步研究方向是将算法有效地应用到实际工程中。

参考文献

[1] 李晓磊, 邵之江, 钱积新. 一种基于动物自治体的寻优模式: 鱼群算法[J]. 系统工程理论与实践, 2002, 22(11): 32-38.

[2] Bonabeau E, Theraulaz G. Swarm Smarts[J]. Scientific American, 2000, 282(3): 72-79.

[3] 李晓磊, 路　飞, 田国会, 等. 组合优化问题的人工鱼群算法应用[J]. 山东大学学报: 工学版, 2004, 34(5): 64-67.

[4] 左兴权, 李士勇. 一种新的免疫进化算法及其性能分析[J]. 系统仿真学报, 2003, 15(11): 1607-1609.

[5] 郑小霞, 钱　锋. 一种改进的微粒群优化算法[J]. 计算机工程, 2006, 32(15): 25-27.

[6] 吴红亮, 王耀南, 周少武, 等. 双群体伪并行差分进化算法研究与应用[J]. 控制理论与应用, 2007, 24(3): 453-458.

[7] 何献忠, 李　萍, 黄航汗, 等. 优化技术及其应用[M]. 2 版. 北京: 北京理工大学出版社, 1995.

[8] 张梅凤, 邵　诚, 甘　勇, 等. 基于变异算子与模拟退火混合的人工鱼群优化算法[J]. 电子学报, 2006, 34(8): 1381-1385.

编辑　金胡考

（上接第 186 页）

表 1　与其他算法的结果比较

算法	样本数目(训练集+测试集)	识别率	可靠性
文献[2]算法	10 000	0.918 1	0.938 5
文献[3]算法	12 000	0.928 0	0.928 0
文献[4]算法	14 650(11 000+3 650)	0.932 6	0.948 8
文献[5]算法	10 000(7 000+3 000)	0.901 2	0.994 0
本文方法	10 000(7 000+3 000)	0.976 3	0.987 7

5　结束语

本文采用方向线素特征辅以加权的特征点特征作为孟加拉数字图像的提取特征，以 BP 神经网络作为分类器，对孟加拉数字进行识别研究。以实际采集的信封图像上的邮政编码作样本库，实验结果表明，识别率达到 97.63%，可靠性为 98.77%，识别效果较好。

参考文献

[1] Dutta A K, Chaudhuri S. Bengalil Alpha-numeric Numeral Recognition Using Curvature Features[J]. Pattern Recognition, 1993, (26): 1757-1770

[2] Pal U, Chaudhuri B B. Automatic Recognition of Unconstrained Off-line Bangla Hand-written Numerals[J]. Lecture Notes on Computer Science, 2000, 1948: 371-378.

[3] Pal U, Belaid A, Chaudhuri B B. A System for Bangla Handwritten Numeral Recognition[J]. IETE Journal of Research, 2006, 52(1): 27-34.

[4] Roy K, Pal T, Pal U, et al. A System for Bangla Handwritten Numeral Recognition Based of Directional Feature[C]//Proc. of the 4th International Conference on Cognition and Recognition. Gold Coast, Australia: [s. n.], 2005.

[5] Xiang Jianying, Sun Shiliang, Lu Yue. A High Reliability Classifier Using Decision Trees and AdaBoost for Recognizing Handwritten Bangla Numerals[C]//Proc. of ICWAPR'07. Beijing, China, [s. n.], 2007.

[6] Kato N, Suzuki M, Omachi S, et al. A Handwritten Character Recognition System Using Directional Element Feature and Asymmetric Mahalanobis Distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(3): 258-262.

编辑　金胡考