

模型的固有复杂度和泛化能力与几何曲率的关系

吕子昂 罗四维 杨 坚 刘蕴辉 邹 琪

(北京交通大学计算机与信息技术学院 北京 100044)

摘 要 从微分几何角度考察与参数化形式无关的统计模型流形的固有复杂度,指出模型流形的 Gauss-Kroneker 曲率可以完全刻画模型流形在一点处的全部性质,进而分析了曲率与体积的关系;给出了基于参数估计量邻域附近的解轨迹方法的曲率计算方法;证明了用于衡量泛化能力的未来残差可以用模型的曲率来表示,由此给出一种新的以曲率度量模型复杂度的模型选择准则 GKCIC;对几何方法和统计学习理论进行了分析比较.在人工数据集和真实数据集上的比较实验结果表明了文中提出的方法的有效性.

关键词 模型选择; 泛化能力; 固有复杂度; 统计流形; Gauss-Kroneker 曲率

中图法分类号 TP18

The Relation Between Intrinsic Complexity and Generalization of a Model and the Geometric Curvature

LU Zi-Ang LUO Si-Wei YANG Jian LIU Yun-Hui ZOU Qi

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044)

Abstract The paper uses the conception of curvature from the point of view of differential geometry to explore the intrinsic model complexity that is free of reparametrization; and then through theoretical analysis, shows that the Gauss-Kroneker curvature can describe the whole properties of the statistical manifold, thus gives the relation between curvature and the volume of the manifold. An algorithm is proposed based on study of the solution locus in the neighborhood of the expectation of parameters to calculate the curvature of the model. This paper proves that the future residual that is qualified to measure the generalizability can be expressed by using the intrinsic curvature array of model, from which a new model selection criterion GKCIC is given. It not only considers the factors such as the number of parameters, sample size and functional form, but also with very clear and intuitive geometric understanding of model selection. The geometrical method of the statistical manifold is compared with the statistical learning theory, in particular, the VC dimension versus the Gauss-Kroneker curvature. By running the algorithm on synthetic and real datasets, the author argue that the GKCIC work efficiently.

Keywords model selection; generalizability; intrinsic complexity; statistical manifold; Gauss-Kroneker curvature

1 引 言

从给定的数据样本中发现规律,并将之应用于

对服从同一分布的未来数据的预测是机器学习乃至科学研究的核心问题,通常这一过程包括对数据建模和对模型进行选择两大方面.然而,从描述同一给定数据的几个可能的模型中选择一个好的模型往往

收稿日期:2005-03-17;修改稿收到日期:2007-04-19. 本课题得到国家自然科学基金(60373029)、教育部博士点基金(20050004001)和北京市重点学科共建项目基金资助. 吕子昂,男,1972年生,博士研究生,主要研究方向包括机器学习、神经网络、知觉学习等. E-mail: lvziang@tsinghua.org.cn. 罗四维,男,1943年生,博士,教授,博士生导师,主要研究领域包括网格、神经网络、计算机视觉等. 杨 坚,男,1970年生,博士,主要研究方向包括神经网络、模型选择等. 刘蕴辉,女,1976年生,博士,讲师,主要研究方向包括机器学习、计算机视觉等. 邹 琪,女,1980年生,博士,讲师,主要研究方向包括机器学习、计算机视觉等.

是困难的,这是因为模型选择问题具有很强的不稳定性.其不稳定性主要表现为:一方面,由于给定数据样本的数量是有限的,不足以使我们确定唯一的模型;另一方面,真实数据必然包含有噪声,如果过分的追求模型对样本数据的拟合,就会导致模型在噪声干扰下的不稳定.

模型选择的目的是获得在未来数据集上具有最佳性能即泛化能力的模型.然而,我们仅能通过减小模型在给定样本数据集上某种度量意义下的误差达到对样本数据的最佳拟合.问题是:如果有两个模型在某种度量意义下对样本数据的拟合程度相同,我们应当如何选择呢?事实上,模型的任何拟合行为都同时包含了对真实规律的逼近和对随机噪声的拟合^[1].模型拟合各种不同数据模式的能力即模型的复杂度.根据 Weierstrass 定理导出的通用逼近定理指出^[2]:具有单隐层的多层感知器可以在有限个点上无限逼近任意有限维连续函数.换言之,总是可以通过增加复杂度来提高拟合度.然而,模型越复杂,吸收随机噪声的能力就越强,最终模型拟合的可能是噪声而并非数据的内在规律.

那么,如何在拟合度和复杂度之间达到最佳平衡,从而获得最佳的泛化能力呢?换言之,对拟合度施加什么程度的复杂度惩罚或限制才可能得到最好的泛化能力?常用的模型选择方法都以达到这一平衡为目标. AIC(Akaike Information Criterion)^[3]复杂度只考虑了模型参数的个数.基于贝叶斯统计理论的 BIC(Bayesian Information Criterion)^[4],其复杂度考虑了参数个数和数据样本大小.随机复杂度 SC(Stochastic Complexity)^[5]通过模型参数的 Hessian 矩阵来考虑影响复杂度的各种因素.源于算法编码理论(algorithmic coding theory)的最小描述长度规则 MDL(Minimum Description Length Principle)^[6]认为描述数据的最佳模型是对其进行最大压缩的描述模型,该方法的复杂度包括参数个数、样本大小和函数形式.统计学习理论(Statistical Learning Theory, SLT)^[7-8]提出了函数集的 VC 维数作为统计模型复杂度的度量.这些方法在拟合度定义上区别不大,但在复杂度的形式上有较大的区别.

然而,以上方法都是从模型的解析形式角度考虑自由参数的个数、函数形式和参数取值范围,作为复杂度的度量.这种方法的缺点是不具备参数表示的不变性,不能反映模型内在的物理特性.从几何的角度,统计模型可以看作微分流形.文献[9]引入了几何复杂度,从几何角度对模型选择作了较深入的

分析.该文作者认为模型的复杂度正比于流形的体积,其含义为模型所包含的可区分的分布个数.文献[10]中作者认为模型推断的贝叶斯方法中那些比例于 $1/N$ 的项(N 是数据样本大小)实质上是为了惩罚模型的曲率.微分几何表明,在给定度量的流形上曲率具有内蕴性质,具有清楚和直观几何意义的曲率可以用来衡量一个模型的非线性程度.而且,从曲率出发,还可以得到更多的反映模型内蕴物理特性的几何量.

本文首先介绍统计模型的微分几何方法,重点分析曲率对统计模型流形的意义,指出高维统计模型流形上 Gauss-Kronecker 曲率反映了模型的复杂度,给出了基于参数估计量邻域附近的解轨迹方法的曲率的计算方法.然后与统计学习理论的进行比较分析,进一步讨论了统计流形局部与大范围的几何性质对模型复杂度的作用.最后给出了与其他模型选择方法的比较实验结果.

2 统计模型的几何方法

2.1 Gauss-Kronecker 曲率和统计模型的复杂度

一个统计模型是一族概率分布的集合,所有可能概率分布的集合形成一个概率空间流形,统计模型是嵌入到这个流形的有限维子流形.虽然真实信息过程所表示的概率分布,也是这个流形的子流形(可能退化为一个点),但真实分布可能不在模型内而是接近它.模型选择就是一个用统计模型去逼近一个真实信息过程的统计推断问题,为了评价这一过程,必须研究统计流形在它们的包容流形(可能概率分布空间)中的几何形状与相对位置等局部和整体的几何性质.

令 $S = \{p(x; \theta)\}$ 为一参数化概率分布的统计模型.其中 X 是样本空间, $x \in X$ 为随机变量; Θ 是 n 维实空间 R^n 的某个开子集中, $\theta \in \Theta$ 是 n 维实值参数; $p(x; \theta)$ 是被 θ 参数化的概率密度函数.

当 $p(x; \theta)$ 在 θ 处充分光滑,就自然导出在统计模型 S 中以 θ 为坐标系的一个可微结构.通过分析坐标可以得到 S 的几何性质,而有趣的是内蕴几何性质与坐标无关.

当样本数量较大时,可以构造全局渐进统计推断的理论.由于此时模型分布与真实分布非常接近,即使模型在整个空间中是弯曲的,我们仍可以进行局部线性化的处理,也就是在一点处用切空间逼近流形.然而,局部线性化只能描述模型的局部特性,

为了说明模型的大尺度属性,必须引入模型流形上两个不同点的切空间之间的仿射联络,这是微分几何的标准技术.有了仿射联络,还可以研究更多的局部非线性属性,例如曲率;更进一步的,可以通过不同点间的切空间的联系获得模型的全局属性.

微分几何方法在统计中的应用始于 Rao 把 fisher 信息量作为统计流形的 Riemannian 度量,并且计算了两个不同分布之间的测地距离^[11]; Chentsov 引入了统计流形的 alpha 联络^[12]; Efron 澄清了曲率的统计意义,并且指出了统计曲率在统计推断的高阶逼近理论中扮演着中心角色^[13].在此基础上,Amari^[14-15]进一步提出了一个微分几何的框架,详细讨论了指数族和混合族的曲率及其对偶性在统计推断中的重要作用.特别引入了对偶联络,发展了对偶平坦空间的大范围理论.

为了更好地理解曲率在刻画流形的性质上的重要性,我们首先介绍经典微分几何 R^3 中的曲面论中关于 Gauss 曲率的结论.

(1) 曲面上一点 p 处的 Gauss 曲率 K 定义为^[16]

$$K = k_1 k_2 = \frac{LN - M^2}{EG - F^2} \tag{1}$$

其中 k_1, k_2 是曲面在点 p 处所有方向的法曲率的最大值和最小值,是两个正交共轭主方向上的主曲率. E, G, F 是曲面 I 形式的系数, L, N, M 是曲面 II 形式的系数.

(2) Gauss 曲率具有重要的几何意义.

设曲面上包含点 p 的小区域 σ , σ 的弯曲程度可以用 σ 的 Gauss 映射的像 σ^* 的面积与 σ 本身的面积的比值来刻画,当 σ 收缩到点 p 时这个比值的极限就表示了曲面在点 p 处的弯曲程度,而这个极限就是点 p 处的 Gauss 曲率,即

$$|K_p| = \lim_{\sigma \rightarrow p} \frac{Area(\sigma^*)}{Area(\sigma)} \tag{2}$$

这一性质的重要性在于说明了 Gauss 曲率反映了曲面在一点处的弯曲程度.曲面在一点处的形状完全由 Gauss 曲率决定,那么,我们完全可以从曲率出发得到曲面更多的几何性质.

(3) 曲面论最重要的结论是 Gauss 绝妙定理,即: Gauss 曲率是内蕴的.

内蕴是指在等距变换下保持不变的性质,内蕴量具有相同的 I 形式,可以完全由度量决定.给定了度量,内蕴量可以不依赖所嵌入的外部空间而在曲面上计算获得.

综上所述, Gauss 曲率反映了曲面的性质,而它又是内蕴的,与参数变化或坐标变换无关.

在高维空间中,流形的 Riemannian 度量张量和仿射联络决定了 Riemannian 曲率张量, Riemannian 曲率张量可以完全确定流形的几何性质.然而, Riemannian 张量计算复杂,在统计模型的曲率研究中,我们选用更直观和便于计算的 Gauss-Kronecker 曲率.

在 n 维流形 M 上给定度量 g , 类似曲面的情况,在 M 上的一点 p 处,有 $n-1$ 个主方向和 $n-1$ 个主曲率 $k_1 \cdots k_{n-1}$, $K = \prod_{i=1}^{n-1} k_i$ 为流形的 Gauss-Kronecker 曲率.和曲面的情况类似, Gauss-Kronecker 曲率具有重要的几何意义^[17].我们定义 n 维空间中的 Gauss 超球映射为 $M \rightarrow U_{n-1}$, 它将 M 上的一点 p 通过平移过 p 点的单位法矢量映射为单位超球上的一点,则有

$$Kds = du_{n-1} \tag{3}$$

其中 ds 是流形 M 的体元, du_{n-1} 是 ds 的 Gauss 超球映射像的体元.说明了 Gauss-Kronecker 曲率反映了流形在一点处的弯曲程度,流形在一点处的形状完全由 Gauss-Kronecker 曲率决定.类似的,我们有高维流形上的 Gauss 绝妙定理: Gauss-Kronecker 曲率是内蕴的,即在坐标变换下保持不变,至多改变符号^[18].因此,我们完全可以从 Gauss-Kronecker 曲率出发得到流形其它的几何性质.

文献[9]从几何角度分析了文献[6]提出的随机复杂度的改进版本——基于最小描述长度的具有参数表示不变性的模型复杂度度量,认为两个不同的模型复杂度之比相当于流形的体积之比,而体积的含义为模型所包含的可区分的分布个数.

$$MDL = -\ln f(y | \hat{\theta}) + \frac{k}{2} \ln \left(\frac{N}{2\pi} \right) + \ln \int d\theta \sqrt{\det \mathbf{I}(\theta)} \tag{4}$$

其中 $\mathbf{I}(\theta)$ 是 Fisher 信息阵,定义为

$$I_{ij}(\theta) = -E_{\theta} \left[\frac{\partial^2 \ln f(y | \theta)}{\partial \theta_i \partial \theta_j} \right].$$

下面着重分析 Gauss-Kronecker 曲率与体积的关系.定义流形 M 的第 r 个平均曲率积分为^[17]

$$MK_r(M) = \binom{n-1}{r}^{-1} \int_M \{k_{i_1}, k_{i_2}, \dots, k_{i_r}\} ds \tag{5}$$

其中 $\{k_{i_1}, k_{i_2}, \dots, k_{i_r}\}$ 定义了主曲率的第 r 阶初等对称函数.例如 $\{k_{11}, k_{12}, \dots, k_{1r}\} = \sum_{i=1}^r k_i, \{k_{21}, k_{22}, \dots,$

$$k_{2r}\} = \sum_{i < j} k_i k_j.$$

流形上 Gauss-Kronecker 曲率的积分称为全绝对曲率,它等于流形球面映射像的体积^[17],也就是单位超球面被流形的 Gauss 超球映射覆盖的次数的平均值.

$$MK_{n-1} = \int_M \prod_{i=1}^{n-1} k_i ds = \int_M K ds = \int_M du_{n-1} \quad (6)$$

在这里我们可以看到,统计模型流形的体积是可以用 Gauss-Kronecker 曲率表示的.这是因为事实上 Gauss-Kronecker 曲率可以用度量矩阵表示^[18]

$$K = \frac{1}{2^{n/2} n!} \sum_{\substack{i_1, \dots, i_n \\ j_1, \dots, j_n}} R_{i_1 i_2 j_1 j_2} \dots R_{i_{n-1} i_n j_{n-1} j_n} \frac{\epsilon_{i_1 \dots i_n}}{\sqrt{\det(g_{ij})}} \frac{\epsilon_{j_1 \dots j_n}}{\sqrt{\det(g_{ij})}} \quad (7)$$

也就是说度量决定了曲率,反过来曲率也可以表示度量的积分——体积.

流形相对于欧式空间最重要的性质就是它的弯曲性.可以看到,曲率可以完全表示流形在一点处的复杂性,并且曲率是内蕴的,更重要的是,曲率的积分有着刻画流形全局性质的重要意义.下面讨论曲率的具体计算.

2.2 统计模型曲率的计算

2.2.1 基本概念和计算方法

以随机神经网络为例,可将统计模型用 p 维参数 θ 参数化为 $y_t = f(x_t, \theta) + \epsilon_t (t = 1, 2, \dots, n)$, 其中 θ 是一个广义参数,其分量对应了网络的权重和门限,与网络单元的作用函数形式也有关.换言之,取不同的作用函数(sigmoid、三角函数、样条函数、方向基函数^[19]等),网络的权重和门限值也会发生变化. n 是用于训练的观测样本维数, ϵ 为随机噪音, f 是网络中神经元的激活函数的某种整合.上式的向量形式为 $\mathbf{Y} = f(\mathbf{X}, \theta) + \epsilon$, 其中 $\mathbf{Y} \in R$, $\mathbf{X} \in R^k$, $\theta \in \Theta$, 通常假设 Θ 是 p -维欧氏空间 R^p 的一个开或紧的子集, f 是关于 θ 的至少 C^2 连续函数, $\epsilon \sim N(0, \sigma^2 I)$, 简称 ϵ 为 iidN.

曲率的直接计算是比较复杂的,为了便于表示和计算,我们使用如下记号^[20].

定义 1. 阵列及其方括号乘法.

一个阵列定义为三维数组 $\mathbf{X}_{npq} = (\mathbf{X}_{tij}) (t = 1, 2, \dots, n; i = 1, 2, \dots, p; j = 1, 2, \dots, q)$, 其中 t, i 和 j 分别为阵列的面、行和列指标. $\mathbf{Y} = [\mathbf{A}][\mathbf{X}]$ 即矩阵 \mathbf{A}_{st} 和阵列 \mathbf{X}_{tij} 的方括号乘法,表示 \mathbf{A} 与 \mathbf{X} 的面指标

的相乘,具体运算为 $\mathbf{Y}_{sij} = \sum_{t=1}^n \mathbf{A}_{st} \mathbf{X}_{tij}$. 更多细节请参考文献[20].

定义 2. 模型的导数和解轨迹流形.

模型 $f(\mathbf{X}, \theta) = f(\theta) = (f_t)$ 的 1, 2 阶导数为 $V(\theta)_{np} = \left(\frac{\partial f_t(\theta)}{\partial \theta_i} \right)$, $W(\theta)_{npp} = \left(\frac{\partial^2 f_t(\theta)}{\partial \theta_i \partial \theta_j} \right)$, $(t = 1, 2, \dots, n; i, j = 1, 2, \dots, p)$. 模型 $\eta = f(\mathbf{X}, \theta)$ 可以看成从 Θ 到 R^n 的映射. 该映射在 R^n 中定义了一个由所有向量 $\boldsymbol{\eta} = f(\theta)$ 的端点组成的流形,称为模型的解轨迹流形(solution locus),记为 π .

通过分析 π 的几何性质可以得到模型的统计特征. 如果模型是线性的,那么 π 是一个由 \mathbf{X} 的列向量生成的超平面(线性空间), θ 的统计特性完全由 π 决定. 对于一个非线性模型,假设噪音水平相对于 π 来讲很小(充分光滑),可以用切平面来逼近 π .

定义 3. 最小二乘估计(LSE).

对模型 $\mathbf{Y} = f(\mathbf{X}, \theta) + \epsilon$, 称 $\hat{\theta}(\mathbf{Y})$ 为 θ 的 LSE, 若估计量 $\hat{\theta} = \hat{\theta}(\mathbf{Y})$ 满足 $S(\hat{\theta}) = \inf_{\theta \in \Theta} S(\theta)$, 其中 $S(\theta) = \|\mathbf{Y} - f(\theta)\|^2 = (\mathbf{Y} - f(\theta))'(\mathbf{Y} - f(\theta))$ 为 \mathbf{Y} 与 $f(\theta)$ 间的距离.

残差向量 $\hat{\epsilon} = \mathbf{Y} - f(\hat{\theta})$ 在点 $\hat{\theta}$ 处与切空间正交, $f(\hat{\theta})$ 是 π 上最接近 \mathbf{Y} 的点,这一最短距离即为 $\|\hat{\epsilon}\|$, Jennrich 定理^[21]保证了 LSE 的存在性. 若 ϵ 为 iidN, 则 θ 的 LSE 就是 θ 的最大似然估计 MLE.

模型的非线性强度的度量最早是由 Beale 提出的^[22], 他提出了 4 种曲率度量,但尚未揭示模型的非线性本质. 在此基础上,受 Efron 统计曲率定义^[13]的启发, Bates 和 Watts 从微分几何观点定义了如下的关于模型的固有曲率和参数效应曲率^[20].

定义 4. 固有曲率和参数效应曲率.

$$\mathbf{K}_h^N = \frac{\|(h'Wh)^N\|}{h'V'h}, \quad \mathbf{K}_h^T = \frac{\|(h'Wh)^T\|}{h'V'h} \quad (8)$$

这里 h 代表 π 上的某方向(我们可以取其单位向量), $(h'Wh)^N$ 和 $(h'Wh)^T$ 分别表示 $(h'Wh)$ 的法向量和切向量. Bates 和 Watts 证明了 \mathbf{K}_h^N 与参数化形式无关,只由模型的内在性质决定;而 \mathbf{K}_h^T 不仅依赖于模型本身,而且还依赖于参数的选择. 事实上,在某一点 θ_0 处 \mathbf{K}_h^N 的最大值 $\mathbf{K}^N = \max_h \mathbf{K}_h^N$ 就是模型流形的 Gauss-Kronecker 曲率^[20].

为了避免对方向的依赖,采用 QR 分解方法选择一组标准正交基作为切空间的基:

$$\mathbf{V}_{np} = (\mathbf{Q} \quad \mathbf{N}) \begin{pmatrix} \mathbf{R} \\ 0 \end{pmatrix} = \mathbf{Q}_{np} \mathbf{R}_{pp} \quad (9)$$

其中 \mathbf{Q} 和 \mathbf{N} 的列向量分别是切空间和法空间的标准正交基, \mathbf{R} 是非退化的上三角矩阵. 变换后 $\mathbf{V}, \mathbf{W}, \mathbf{h}$ 分别成为 $\mathbf{Q}=\mathbf{V}\mathbf{L}$, $\mathbf{U}=\mathbf{L}'\mathbf{W}\mathbf{L}$ ($\mathbf{L}=\mathbf{R}^{-1}$) 和 $\mathbf{d}=\mathbf{R}\mathbf{h}$, 相应的得到固有曲率阵列 $\mathbf{I}_{(n-p)pp}=[\mathbf{N}'][\mathbf{U}]$ 和参数效应曲率阵列 $\mathbf{P}_{ppp}=[\mathbf{Q}'][\mathbf{U}]$.

2.2.2 曲率计算举例

我们来看一个简单的例子, 假设我们有数据集 $\mathbf{X}=\{2, 3\}, \mathbf{Y}=\{2.5, 10\}$, 回归模型 $E(\mathbf{Y})=\mathbf{X}^\theta$, 此时 $p=1, n=2$. 可以求得最小二乘估计 $\hat{\theta}=2.0537$.

此时对应的预测值为 $\hat{\mathbf{Y}}=\{4.1517, 9.5468\}$, 残差方差为 $\hat{\sigma}^2=[(2.5-4.1517)^2+(10-9.5468)^2]/1=2.9334$. 则

$$\mathbf{V}(\theta)_{np}=\left(\frac{\partial f_i(\theta)}{\partial \theta_i}\right)=(\mathbf{x}_i^\theta \ln \theta)=\begin{pmatrix} 4.1517 \ln 2 \\ 9.5468 \ln 3 \end{pmatrix}=\begin{pmatrix} 2.8777 \\ 10.4882 \end{pmatrix}.$$

QR 分解可得: $\mathbf{V}=\mathbf{Q}\mathbf{R}=(\mathbf{Q} \ \mathbf{N})\begin{pmatrix} \mathbf{R} \\ 0 \end{pmatrix}=\begin{pmatrix} -0.26460 & -0.96438 \\ -0.96438 & 0.26460 \end{pmatrix}\begin{pmatrix} -6.3500 \\ 0 \end{pmatrix}$, 则 $\mathbf{L}=\mathbf{R}^{-1}=-0.15748$.

$$\text{而 } \mathbf{W}(\theta)_{npp}=\left(\frac{\partial^2 f_i(\theta)}{\partial \theta_i \partial \theta_j}\right)=(\mathbf{x}_i^\theta \ln^2 \theta)=\begin{pmatrix} 4.1517 \ln^2 2 \\ 9.5468 \ln^2 3 \end{pmatrix}=\begin{pmatrix} 1.9947 \\ 11.5225 \end{pmatrix},$$

$$\mathbf{U}=\mathbf{L}'\mathbf{W}\mathbf{L}=-0.15748\begin{pmatrix} 1.9947 \\ 11.5225 \end{pmatrix}(-0.15748)=\begin{pmatrix} 0.04947 \\ 0.28576 \end{pmatrix}.$$

$$\text{则固有曲率阵列 } \mathbf{I}_{(n-p)pp}=[\mathbf{N}'][\mathbf{U}]=(-0.96438 \ 0.26460)\begin{pmatrix} 0.04947 \\ 0.28576 \end{pmatrix}=0.0279,$$

$$\text{参数效应曲率阵列 } \mathbf{P}_{ppp}=[\mathbf{Q}'][\mathbf{U}]=(-0.26460 \ -0.96438)\begin{pmatrix} 0.04947 \\ 0.28576 \end{pmatrix}=-0.28867.$$

对于一般的情况, 没有计算最大固有曲率的公式, 文献[23]给出了曲率计算的迭代算法, 我们在此基础上作了进一步的改进.

2.3 统计模型复杂度和泛化能力的曲率度量

2.3.1 正则条件

如引言中所分析的, 对模型泛化能力的评价应当是它在测试集上的性能, 也就是在统计理论中的未来残差. 下面我们主要讨论未来残差期望和方差的曲率表示形式.

为了研究模型的性质, 通常考虑参数 θ 及其各阶渐近矩的最小二乘估计的渐进特性, 如相合性和渐进正态性等. 这些性质的分析大多需要一定的正则条件约束. 现将所需的正则条件列举如下^[21, 23]:

(1) 模型 $\mathbf{Y}=f(\mathbf{X}, \theta)+\epsilon$ 中, 函数 f 是紧集 Θ 上关于 θ 的连续函数, 随机误差项 $\epsilon \sim N(0, \sigma^2 I)$.

(2) 当 $n \rightarrow \infty$ 时, 函数 $D_n(\theta, \theta')=\frac{1}{n} \sum_{i=1}^n [f_i(\theta)-f_i'(\theta)]^2$ 在 $\Theta \times \Theta$ 上一致收敛到函数 $D(\theta, \theta')$, 且 $D(\theta, \theta')$ 在 Θ 上存在唯一的极小值点.

(3) 函数 f 在 Θ 上关于 θ 存在一阶连续偏导数, 且在 θ_0 的某邻域内有 $\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{V}'(\theta) \mathbf{V}(\theta)=\mathbf{\Omega}(\theta)$, 其中 $\mathbf{\Omega}(\theta)$ 为正定矩阵, 这时记 $\mathbf{V}'(\theta) \mathbf{V}(\theta)=O(n)$, $\mathbf{V}(\theta)=O(n^{1/2})$.

(4) 函数 f 在 Θ 上关于 θ 存在二阶连续偏导数, 且在 θ_0 的某邻域内有 $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left[\frac{\partial^2 f_i(\theta)}{\partial \theta_i \partial \theta_j} \right]^2 = E_{ij}(\theta)$.

(5) 假定模型函数 f 在 Θ 上存在三阶连续导数, 并且下列有关导数的极限存在

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \frac{\partial f_i(\theta)}{\partial \theta_i} \frac{\partial^2 f_i(\theta)}{\partial \theta_j \partial \theta_k} = E_{ijk}(\theta) < +\infty,$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left[\frac{\partial^3 f_i(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k} \right]^2 = G_{ijk}(\theta) < +\infty.$$

(6) 当 $n \rightarrow \infty$ 时, $C(J_1, J_2)=O(1)$ 对任意指标集 J_1, J_2 成立. 其中 J_1, J_2 表示标号的指标集, $C(J_1, J_2)=\frac{1}{n} \sum_{i=1}^n f_{i,J_1} f_{i,J_2}=\frac{1}{n} (f_{i,J_1})'(f_{i,J_2})$. 例如 $J=(i)$ 则 $f_{i,J}=\frac{\partial f_i}{\partial \theta_i}$, 若 $J=(i, j, k)$ 则 $f_{i,J}=\frac{\partial^3 f_i}{\partial \theta_i \partial \theta_j \partial \theta_k}$. 依此类推.

这些正则条件由弱到强, 条件(1~4)是比较基本的约束, 其中条件(2)是保证相合性成立的重要条件, 条件(1~4)保证了渐近正态性. 条件(5)是较强的约束, 是对最小二乘估计作二阶随机展开的保证, 条件(6)是最强的约束, 是计算参数估计量的各阶矩的保证.

2.3.2 曲率度量

对模型 $\mathbf{Y}=f(\mathbf{X}, \theta)+\epsilon$, 当给定新数据 \mathbf{X}_0 时, 我们考虑 $\mathbf{Y}_0=f(\mathbf{X}_0, \hat{\theta})$ 的预测值.

引理 1. 若满足正则条件(1)~(5), 那么有^[23]

$$\hat{\theta}-\theta_0=\mathbf{L}\boldsymbol{\tau}+\mathbf{L}\left\{[\lambda'][\mathbf{I}]\boldsymbol{\tau}-\frac{1}{2}\boldsymbol{\tau}'\mathbf{P}\boldsymbol{\tau}\right\}+o_p(n^{-3/2}) \quad (10)$$

$$\hat{\epsilon} = \mathbf{N}\boldsymbol{\lambda} - \mathbf{Q}[\boldsymbol{\lambda}'][\mathbf{I}]\boldsymbol{\tau} - \frac{1}{2}\mathbf{N}(\boldsymbol{\tau}'\mathbf{I}\boldsymbol{\tau}) + o_p(n^{-1}) \quad (11)$$

$$\Delta f = \mathbf{Q}\boldsymbol{\tau} + \mathbf{Q}[\boldsymbol{\lambda}'][\mathbf{I}]\boldsymbol{\tau} + \frac{1}{2}\mathbf{N}(\boldsymbol{\tau}'\mathbf{I}\boldsymbol{\tau}) + o_p(n^{-1}) \quad (12)$$

其中 \mathbf{N} 的列向量是法空间的标准正交基, $\boldsymbol{\tau} = \mathbf{Q}'\boldsymbol{\epsilon}$, $\boldsymbol{\lambda} = \mathbf{N}'\boldsymbol{\epsilon}$, \mathbf{I} 和 \mathbf{P} 分别为固有曲率和参数效应曲率阵列, 均在 θ_0 处计算.

上述定理表明 $\hat{\epsilon}$ 和 $\Delta f = f(\hat{\theta}) - f(\theta_0)$ 都可以仅用固有曲率阵列 \mathbf{I} 表示, 因此与参数化(坐标变换)形式无关.

引理 2. 如果满足所需的正则条件(1)~(6), 那么 $\hat{\epsilon}$ 和 Δf 的期望和方差为^[24]

$$E\hat{\epsilon} = -\frac{\sigma^2}{2}\mathbf{N}tr[\mathbf{I}] + O(n^{-1}) \quad (13)$$

$$Var(\hat{\epsilon}) \approx \sigma^2\mathbf{P}_N + \sigma^4\mathbf{Q}\mathbf{V}_I\mathbf{Q}' + \frac{1}{2}\sigma^4\mathbf{N}\mathbf{V}_I^*\mathbf{N}' \quad (14)$$

$$E[\Delta f] = \frac{\sigma^2}{2}\mathbf{N}tr[\mathbf{I}] + O(n^{-1}) \quad (15)$$

$$Var(\Delta f) \approx \sigma^2\mathbf{P}_T + \sigma^4\mathbf{Q}\mathbf{V}_I\mathbf{Q}' + \frac{1}{2}\sigma^4\mathbf{N}\mathbf{V}_I^*\mathbf{N}' \quad (16)$$

其中 $\mathbf{V}_I = \sum_{t=1}^{n-p} \mathbf{I}_t^2$, \mathbf{I}_t 是 \mathbf{I} 的第 t 面; $\mathbf{V}_I^* = \sum_{k=1}^p \sum_{l=1}^p \mathbf{I}_{kl}\mathbf{I}_{kl}'$, \mathbf{I}_{kl} 是 (k, l) 处 \mathbf{I} 的 $(n-p)$ 维向量.

同理可得上式均与参数化形式无关, 因为它们仅依赖于固有曲率阵列 \mathbf{I} .

以上所讨论的是当前数据集上的残差的曲率度量, 如前所述, 为了获得最佳的泛化能力, 我们必须最小化未来数据集上的残差. 下面分析未来残差的曲率度量, 我们取由文献[25]定义的未来残差形式:

$$R_{\text{exp}} = E^*[E[\|\mathbf{Y}^* - f(\hat{\theta})\|^2]] \quad (17)$$

上式中, $E^*[\cdot]$ 和 $E[\cdot]$ 分别表示关于未来数据和当前数据的期望. 由于 $f(\hat{\theta})$ 由当前数据 \mathbf{Y} 决定, 等式(17)简化为

$$\begin{aligned} R_{\text{exp}} &= E^*[E[\|\mathbf{Y}^* - f(\theta_0)\| - \|f(\hat{\theta}) - f(\theta_0)\|^2]] \\ &= E^*[\|\mathbf{Y}^* - f(\theta_0)\|^2] + E[\|f(\hat{\theta}) - f(\theta_0)\|^2] \end{aligned} \quad (18)$$

假设 \mathbf{Y}^* 和 \mathbf{Y} 独立同分布, 则 $E^*[\|\mathbf{Y}^* - f(\theta_0)\|^2] = E[\|\mathbf{Y} - f(\theta_0)\|^2]$. 那么,

$$R_{\text{exp}} = E[\|\mathbf{Y} - f(\theta_0)\|^2] + E[\|f(\hat{\theta}) - f(\theta_0)\|^2] \quad (19)$$

而 $\hat{\epsilon} = \mathbf{Y} - f(\hat{\theta}) = \mathbf{Y} - f(\theta_0) - [f(\hat{\theta}) - f(\theta_0)] = \boldsymbol{\epsilon} - \Delta f$, 最后我们有

$$\begin{aligned} R_{\text{exp}} &= E[\|\boldsymbol{\epsilon}\|^2] + E[\|\Delta f\|^2] \\ &= E[\|\hat{\epsilon} + \Delta f\|^2] + E[\|\Delta f\|^2] \end{aligned} \quad (20)$$

至此, 我们可以利用前面的定理得到 R_{exp} 与曲率的关系如下.

定理 1. 如果满足所需的正则条件(1)~(6), 那么有

$$R_{\text{exp}} \approx (n+p)\sigma^2 + \frac{1}{4}\|\text{tr}[\mathbf{I}]\|^2 + \frac{3}{2}\sigma^4 \sum_{k=1}^p \sum_{l=1}^p \|\mathbf{I}_{kl}\|^2 \quad (21)$$

证明. 首先有

$$E\|\cdot\|^2 = \|E\cdot\|^2 + \text{tr}\{\text{Var}(\cdot)\} \quad (22)$$

其中 \cdot 代表向量.

由等式(15)有

$$\|E\Delta f\|^2 = \frac{1}{4}\sigma^4 \|\text{tr}[\mathbf{I}]\|^2.$$

由等式(16)有

$$\begin{aligned} \text{tr}\{\text{Var}(\Delta f)\} &\approx \sigma^2 \text{tr}(\mathbf{P}_T) + \sigma^4 \text{tr}(\mathbf{V}_I\mathbf{Q}'\mathbf{Q}) + \\ &\quad \frac{1}{2}\sigma^4 \text{tr}(\mathbf{V}_I^*\mathbf{N}'\mathbf{N}) \\ &\approx p\sigma^2 + \sigma^4 \text{tr}(\mathbf{V}_I) + \frac{1}{2}\sigma^4 \text{tr}(\mathbf{V}_I^*), \end{aligned}$$

而

$$\begin{aligned} \text{tr}(\mathbf{V}_I) &= \text{tr}\left(\sum_{t=1}^{n-p} \mathbf{I}_t^2\right) = \sum_{t=1}^{n-p} \text{tr}(\mathbf{I}_t^2) \\ &= \sum_{t=1}^{n-p} \sum_{k=1}^p \sum_{l=1}^p \mathbf{I}_{tkl}^2 = \sum_{k=1}^p \sum_{l=1}^p \|\mathbf{I}_{kl}\|^2, \\ \text{tr}(\mathbf{V}_I^*) &= \text{tr}\left(\sum_{k=1}^p \sum_{l=1}^p \mathbf{I}_{kl}\mathbf{I}_{kl}'\right) = \sum_{k=1}^p \sum_{l=1}^p \text{tr}(\mathbf{I}_{kl}'\mathbf{I}_{kl}) \\ &= \sum_{k=1}^p \sum_{l=1}^p \|\mathbf{I}_{kl}\|^2. \end{aligned}$$

将以上等式代入(22)可得

$$E\|\Delta f\|^2 \approx p\sigma^2 + \frac{3}{2}\sigma^4 \sum_{k=1}^p \sum_{l=1}^p \|\mathbf{I}_{kl}\|^2 + \frac{1}{4}\sigma^4 \|\text{tr}[\mathbf{I}]\|^2 \quad (23)$$

同样的计算容易得到 $E\|\boldsymbol{\epsilon}\|^2 = n\sigma^2$, 那么

$$\begin{aligned} R_{\text{exp}} &= E[\|\boldsymbol{\epsilon}\|^2] + E[\|\Delta f\|^2] \\ &\approx (n+p)\sigma^2 + \frac{1}{4}\|\text{tr}[\mathbf{I}]\|^2 + \frac{3}{2}\sigma^4 \sum_{k=1}^p \sum_{l=1}^p \|\mathbf{I}_{kl}\|^2. \end{aligned}$$

证毕.

由定理知未来残差只依赖于固有曲率阵列 \mathbf{I} , 与参数化无关. 对于给定的数据样本, 固有曲率阵列的几何意义就是 Gauss-Kronecker 曲率, 它真实地反映了模型的内在复杂程度, 我们可以把它作为模型的复杂度衡量标准.

2.4 基于 Gauss-Kronecker 曲率的模型选择准则

如引言所述,模型的泛化能力可以分解成对样本的拟合度和模型的固有复杂度,我们的目标是在二者间达到一个平衡.因此由定理得到如下基于 Gauss-Kronecker 曲率的模型选择准则(Gauss-Kronecker Curvature Information Criterion, GKCIC).在给定的模型中,最佳的选择是使下式最小化^[26]:

$$GKCIC = -\ln f(x | \hat{\theta}) + \frac{p}{2} \ln \left(\frac{n}{2\pi} \right) + GKC \quad (24)$$

其中 GKC 即拟合点 θ_0 处 K_h^N 的最大值 $K^N = \max_h K_h^N$,也就是模型流形的 Gauss-Kronecker 曲率.

这说明给定 n 个数据样本时,若模型的拟合程度相同,应当选择具有尽可能少的参数和在拟合点附近具有尽可能小的固有曲率的模型,换言之在情况允许的条件下倾向于选择尽可能逼近线性的模型.

3 统计模型的几何方法和统计学习理论比较

近年来, Vapnik 等人提出的统计学习理论^[7-8]和基于该理论的支持向量机(Support Vector Machine, SVM)算法在机器学习领域掀起了一场革命.和传统的统计模式识别方法相比,统计学习理论不再是在估计分布的基础上的演绎推理,而是强调在有限样本条件下的从数据到分布的归纳推理,其推理依据也不再是传统统计学的大数定律,而是可能近似正确(Probably Approximately Correct, PAC)框架^[27].统计学习理论认为,在大样本条件下,几乎所有的学习机器都可以通过经验风险最小化(Empirical Risk Minimization, ERM)达到最优的泛化能力.但在有限样本的条件下,必须考虑模型的复杂度,通过结构风险最小化(Structural Risk Minimization, SRM)对给定数据逼近的精度和逼近函数的复杂性之间取得折中. Vapnik 提出了函数集的 VC 维数作为对统计模型复杂度的度量,它是描述函数集或学习机器的复杂性或学习能力的一个重要指标.以指示函数集为例,如果存在 h 个样本能够被函数集中的函数按所有可能的 2^h 种形式分开,则称函数集能够把 h 个样本打散,函数集的 VC 维就是它能打散的最大样本数目 h .考虑一系列嵌套的函数子集 $S_1 \subset S_2 \subset S_3 \cdots$,随着函数集的扩大 VC 维随之增大,学习机器也就更复杂(容量越大),泛化能力将

随之下降.因此必须考虑最小化结构风险或者说综合考虑经验风险和置信区间以求达到最佳的学习效果.为了避免过学习,应当在保持经验风险的前提下,选择 VC 维较小的学习器.参见图 1. SVM 算法是建立在 SLT 的 VC 维理论和 SRM 原理基础上的实际应用,对于非线性问题其主要思路是通过构造核函数将问题通过非线性变换转换到高维的线性空间,在高维空间中构造线性判别函数来实现原空间中的非线性判别.然而,在实践中对于一个实际的学习机器的 VC 维的分析尚没有通用的方法,理论分析主要是对线性判别函数有结论 $h = n + 1$,对于非线性函数,除了一些特殊函数之外,主要是将其与线性判别函数进行类比.对于特定结构的神经网络也有一些研究成果.

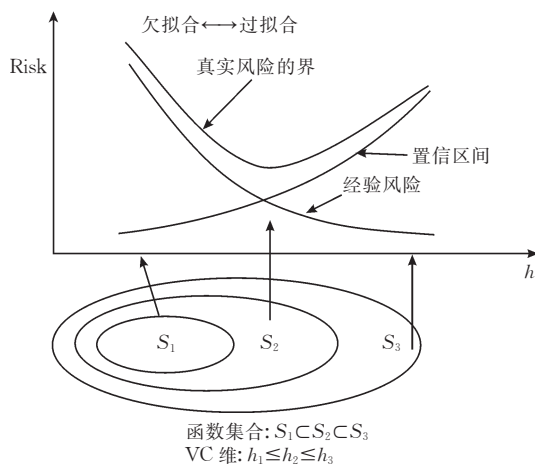


图 1 VC 维和 SRM^[7]

本文提出的基于 Gauss-Kronecker 曲率的模型选择的理论基础是 Amari 创立的信息几何(Information Geometry)理论^[14-15],该理论是在对概率分布流形几何特性研究的基础上,将流形上的微分几何方法引入到统计学.从信息几何的角度看,统计模型是嵌入在所有可能概率分布空间中的一个有限维分布流形,流形上的每个点表示一个分布.模型选择就是一个统计推断的过程,为了评价这一过程必须考察模型流形的几何性质,分析由具有微分结构和参数结构的概率分布族表示的信息系统的统计流形的内在几何结构,了解统计模型在所有可能概率分布空间中的位置和形状,从而揭示系统的信息处理能力.位置信息给出了分布流形与真实分布的接近程度,反映的是模型的几何拟合度,形状信息则反映了模型的内蕴几何复杂度.对形状信息最自然的度量就是流形的曲率,微分几何理论指出流形的黎曼曲率张量完全确定了流形的性质^[18].在实际应用

中,由于黎曼曲率计算较为复杂,高斯曲率是在流形上一点处的微元面积与其在该点处单位超球体上的高斯映射的像的面积的比值的极限^[17],反映了流形上该点处的弯曲程度.高斯曲率是流形的第二基本形式与第一基本形式的比,是不依赖于坐标选择的参数无关的内蕴几何量.从几何上看,固有曲率 K_h^N 是向量方向上期望曲面最好的近似圆的半径的倒数^[28],考虑样本特性因素后可以得到相对固有曲率,其几何意义就是高斯曲率.相对曲率反映了对数据的依赖,该方法在预测问题中的实际应用在于找到相对固有曲率最小的模型,对于参数效应曲率较大的模型,可以进行重参数化,在固有曲率保持不便的条件下减少参数效应曲率.由此可以看出,基于曲率的模型复杂度度量强调了数据集对模型复杂度的影响,即不同数据集上的同一模型会有不同的曲率.

一个有趣的结论是正弦函数族的 VC 维为无穷大^[7],对于这一问题,基于信息几何方法的模型曲率复杂度理论解释如下:对于给定值域区间 $[-1,1]$ 上的正弦函数族,在任意点处,由于其必然是正交的,可得其一阶导数的 QR 分解形式不变,即 $\mathbf{R}=\mathbf{I}$, $\mathbf{Q}=\mathbf{V}$,由此可得 $\mathbf{U}=\mathbf{W}$.而此时 N 的列向量生成该点处的法空间没有限制,可以为任意大的值.因此固有曲率阵列 $\mathbf{I}_{(n-p)pp}=[\mathbf{N}'][\mathbf{U}]$ 的值也可以为无穷大,这就是曲率方法对正弦函数族复杂度的解释.

通过以上分析可以看出,几何方法和统计学习理论都是统计学在机器学习中的应用.在模型选择方面,基于信息几何理论的模型固有复杂度的曲率度量方法和统计学习理论的 VC 维度量方法相比较,二者的理论出发点是一致的,都认为泛化能力是学习机器的优化目标,而且泛化能力是在拟合度与复杂度之间的折中.二者的主要区别在于具体的实现方法,SLT 的思路是非线性问题的线性化,将主要矛盾转移到核函数的选取;基于信息几何方法的非线性模型复杂度分析的研究思路主要是计算模型解轨迹的曲率,作为对模型非线性强度的评价,直接将非线性模型用于预测,该方法特别强调了模型对数据的依赖性.

4 实验和分析

为验证 GKCIC 方法的有效性,我们分别在人工数据集和真实数据上进行了实验.

4.1 人工数据

由于 VC 维没有统一的计算方法,而 AIC 显然

性能比 BIC 差,我们选择 BIC 和 MDL 作为对照.其准则函数如下:

$$\begin{aligned} BIC &= -\ln f(x|\hat{\theta}) + p\ln(n) \quad (25) \\ MDL &= -\ln f(x|\hat{\theta}) + \frac{p}{2}\ln\left(\frac{n}{2\pi}\right) + \ln\int \sqrt{\det\mathbf{I}(\theta)}d\theta \quad (26) \end{aligned}$$

其中 $\hat{\theta}$ 是模型参数的最大似然估计 MLE, p 是模型中参数个数, n 是样本大小, $\mathbf{I}(\theta)$ 是 Fisher 信息矩阵.以上各式中,第一项为拟合度,余项为复杂度.

- 实验方案如下:
- (1) 选用心理学研究中的两个物理-心理对照模型 Steven 模型($M1:y=ax^b+error$)和 Fecher 模型($M2:y=a\ln(x+b)+error$)^[9].
 - (2) 对于每个模型,在相同的 $\mathbf{X}=(1,2,3,4)$ 上分别采样生成容量为 200,600,1000 的样本集,误差区均值为 0,方差为 1 的随机噪声.模型参数为 $M1:a=2,b=2$; $M2:a=2,b=5$.

- (3) 对样本集中的每个样本进行模型选择,即对该样本计算用 M1 和 M2 拟合时的准则函数值,选择较小的一个为恢复模型.
- (4) 对样本集计算正确选择和错误选择的百分比.

选择准则的比较是基于它们多大程度恢复产生样本数据的真实模型的能力,即各准则选择真实模型的样本数占总样本数的百分比.从表 1 中数据可以看出:由于 BIC 方法没有考虑模型的函数形式对复杂度的影响,总是倾向于选择 M2 为恢复模型;而 MDL 和 GKCIC 都考虑了影响复杂度的样本容量、参数个数及函数形式.因此可以选择出正确的恢复模型.随着样本容量的增加,三种方法恢复正确模型的能力也随之增加. MDL 方法是目前公认的最有效的方法,我们的方法和 MDL 相比差别不大.但是从另一角度揭示了模型选择的内在本质,它使我们能够更直观、更清晰地理解模型选择的几何意义.

表 1 模型选择方法比较(恢复能力/%)							
样本 容量	拟合 模型	恢复能力					
		BIC		MDL		GKCIC	
		M1	M2	M1	M2	M1	M2
200	M1	92.5	70.0	85.3	10.7	85.0	10.0
	M2	7.5	30.0	14.7	89.3	15.0	90.0
600	M1	96.9	65.5	90.8	2.3	87.5	5.0
	M2	3.1	34.5	9.2	97.7	12.5	95.0
1000	M1	99.0	60.3	98.6	0.8	96.7	1.0
	M2	1.0	39.7	1.4	99.2	3.3	99.0

4.2 GKCIC 在知觉组织中的应用

我们将 GKCIC 模型选择方法应用于知觉组

织,实现了基于自然图像统计特性的格式塔规则模型的选择^[29].以接近律为例,需要进行模型选择的问题可描述为“已知两条边缘间的接近程度,这两条边缘属于同一轮廓编组的概率服从怎样的分布函数”.

模型选择算法的流程如下:

(1)产生样本数据,即边缘属于同一轮廓编组时接近律的取值.对某一类自然图像形成重要边缘图,测量某两条属于同一轮廓编组的边缘之间的距离作为一个样本.对该类中不同的自然图像采样,当样本数量达到预定规模时停止采样;

(2)作出样本分布图,根据总体分布走势,选择若干个含有待定参数的函数作为候选模型;

(3)根据 GKCIC 模型选择准则,从候选模型中选出最适合的函数并确定待定参数,作为该条格式塔规则在该类自然图像下的统计模型.

根据这一准则,我们确定最佳的高斯函数、拉普拉斯函数和幂函数分别为

$$f(x) = \frac{2}{2.2\sqrt{2\pi}} \exp\left\{-\frac{(x-1.4)^2}{2 \times 2.2^2}\right\}, x \geq 1.4 \quad (27)$$

$$f(x) = 0.48 \exp\{-(0.82 |x-1.4| / 1.91)^{1.63}\}, x \geq 1.4 \quad (28)$$

$$f(x) = 0.68x^{-1.56}, x \geq 1.4 \quad (29)$$

根据这 3 个函数在 GKCIC 准则下的适应值,我们选取幂函数式(29)为最佳模型.

5 总 结

模型选择是机器学习理论研究中重要而又困难的一个领域,Leo Breiman 指出的机器学习有 3 个基本问题,其中模型多样性的 Rashomon 问题和简单性与精度的冲突的 Occam 问题都与模型选择相关^[30].模型选择的任务就是选择具有最佳泛化能力的模型,为此必须做到拟合度与复杂度的平衡.本文从几何角度对曲率在刻划统计模型流形性质中的作用进行了深入研究,指出了与参数化无关的模型的固有特性度量,即 Gauss-Kroneker 曲率的复杂度如何影响模型的泛化能力.提出了 GKCIC 模型选择准则并将至应用于实际任务,对于模型选择的 Occam 规则给出了清晰而直观的理解.正如文献[1]中所说:现有的模型选择方法只是一些协助了解事实过程的工具,它们不是事实的仲裁者,但当考虑了其它关于选择的因素时我们能够使之更为有用.

参 考 文 献

- [1] Myung I J, Pitt M A. Model comparison methods//Brand L, Johnson M L eds. Proceedings of the Methods in Enzymology. Elsevier, 2004: 351-366
- [2] Haykin S. Neural Networks: A Comprehensive Foundation. 2nd Edition. Upper Saddle River, NJ: Prentice Hall, 1999
- [3] Akaike H. Information theory and an extension of the maximum likelihood principle//Proceedings of the 2nd International Symposium on Information Theory. Budapest, 1973: 267-281
- [4] Schwarz G. Estimation the dimension of a model. Annals of Statistics, 1978, 7(2): 461-464
- [5] Rissanen J. Universal coding, information, prediction, and estimation. IEEE Transactions on Information Theory, 1984, 30(4): 629-636
- [6] Rissanen J. Fisher information and stochastic complexity. IEEE Transactions on Information Theory, 1996, 42(1): 40-47
- [7] Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1999
- [8] Vapnik V N. Statistical Learning Theory. New York: John Wiley & Sons, 1998
- [9] Myung I J, Balasubramanian V, Pitt M A. Counting probability distributions: Differential geometry and model selection. Proceedings of National Academy of Science, 2000, 97(21): 11170-11175
- [10] Balasubramanian V. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. Neural Computation, 1997, 9(2): 349-368
- [11] Rao C R. Information and accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society, 1945, 37: 81-91
- [12] Chentsov N N. Statistical decision rules and optimal inference. Providence, R.I.: AMS, 1982
- [13] Efron B. Defining the curvature of a statistical problem. Annals of Statistics, 1975, 3(3): 1189-1242
- [14] Amari S. Differential-Geometrical Methods in Statistics. New York: Springer-Verlag, 1985
- [15] Amari S. Methods of Information Geometry. New York: Oxford University Press, 2000
- [16] Carmo M P D. Differential Geometry of Curves and Surfaces. Beijing: China Machine Press, 2004
- [17] Santalo L A. Integral Geometry and Geometric Probability. Reading, Mass: Addison-Wesley, 1979
- [18] Spivak M. A Comprehensive Introduction to Differential Geometry. Wilmington: Publish or Perish, 1979
- [19] Wang S, Shi J, Chen C, Li Y. Direction-basis-function neural networks//Proceedings of the International Joint Conference on Neural Networks. Washington DC, USA, 1999: 1251-1254

[20] Bates D. Relative curvature measures of nonlinearity. Journal of the Royal Statistical Society, 1980, 42(1): 1-25

[21] Jennrich R I. Asymptotic properties of nonlinear least squares estimators. Annals of Math, 1969, 40(3): 633-643

[22] Beale E M L. Cofidence regions in nonlinear estimation. Journal of the Royal Statistical Society, 1960, B22(1): 41-88

[23] Ratkowsky D A. Nonlinear Regression Modeling. New York, Marcel Dekker, Inc, 1983

[24] Wei B C. Modern Nonlinear Regression Analysis. Nanjing: Southeast University Press, 1989

[25] Kanatani K. Geometric information criterion for model selection. International Journal of Computer Vision, 1998, 26(3): 1-21

[26] Lv Z A. A new geometric approach to the complexity of model selection//Proceedings of the IEEE International Conference on Cognitive Informatics. Beijing, China, 2006: 268-273

[27] Valiant L G. A theory of learnability. Communications of the ACM, 1984, 27(11): 1134-1142

[28] Bates D, Watts D G. Nonlinear regression analysis and its applications. John Wiley & Sons, Inc. , 1988

[29] Zou Q. Research on computational model of contour grouping and attention model[Ph D dissertation]. Beijing Jiaotong University, Beijing, 2006

[30] Breiman L. Statistical modeling: The two cultures. Statistical Science, 2001, 16(3): 199-231



LU Zi-Ang, born in 1972, Ph. D. candidate. His research interests include machine learning, neural network, and perceptual learning etc.

LUO Si-Wei, born in 1943, professor, Ph. D. supervi-

sor. His research interests include grid, neural network, and computer vision etc.

YANG Jian, born in 1970, Ph. D. His research inter-ests include neural network, and model selection etc.

LIU Yun-Hui, born in 1976, Ph. D. Her research inter-ests include machine learning, and computer vision etc.

ZOU Qi, born in 1980, Ph. D. Her research interests include machine learning, and computer vision etc.

Background

The work is supported by the National Natural Science Foundation of China under grant 60373029, Research Fund for the Doctoral Program of Higher Education of China (20050004001) and Co-construction Project of Key Subject of Beijing.

Neural computation is an intercrossed science involving cognitive science, artificial intelligence, neural network and so on. Among which, effective coding is an important theory for comprehending neural system function. The theory was brought forward in 1961, but it is not developed extensively until man understands more deeply toward neural system in 1990's. Therefore, it is of great sense tracing research in this field.

Using mathematic tools as differential geometry and statistics, the project is to advance new cognitive model fitting

theoretical framework of effective coding in human vision cognitive system better. The model is built from viewpoints of differential manifold, information geometry and statistics. Through analyzing cooperation mechanism among multi visual areas, the authors construct a hierarchical and parallel effective coding model, which can realize side feedback and side inhibition to some extent. In this way, the model approaches information processing mechanism of human brain more closely. The research results of above theory and model are to be validated by being applied into practical system.

The project adopts the method combining cognitive science and neural science. With deeper exploration of human behavior, it is sure to have broad application space.

This project has been studied for about 4 years, and of more than 30 papers has been published.