

超核函数支持向量机

贾磊 廖士中

(天津大学计算机科学与技术学院 天津 300072)

摘要 支持向量机是当前机器学习、模式识别和数据挖掘等领域的重要学习方法,核函数的构造是研究和应用支持向量机的关键问题。针对这一问题,提出了核函数构造的组合理论,定义了超核函数概念,并通过多项式组合现有核函数构造出一类超核函数。具体地,首先分析了一般核函数存在的过学习和欠学习现象,然后证明了组合理论构造的核函数的 Mercer 性质,并通过在仿真数据集和标准数据集上的对比实验,验证了超核函数的性能。理论分析和实验结果阐明了所提出的超核函数组合构造理论的合理性和有效性,开拓了模型选择组合方法的研究途径。

关键词 核方法,超核函数,多项式组合,支持向量机

Support Vector Machines with Hyperkernel Functions

JIA Lei LIAO Shi-zhong

(School of Computer Science and Technology, Tianjin University, Tianjin 300072, China)

Abstract Support Vector Machines (SVM) provides an important approach to machine learning, pattern reorganization and data mining. Construction of kernel functions is the key both in SVM research and application. We proposed a combinatorial construction of hyperkernel functions for SVM, which is constructed by a polynomial combination of common kernels. We first analyzed the over-learning and under-learning phenomena of common kernel function. Then, we proved the Mercer condition of hyperkernels constructed by the proposed method. Finally, we experimented both on simulation data and benchmark data to demonstrate the performance of hyperkernels of SVM. The theoretical proofs and experimental results illuminate the validity and feasibility of hyperkernels, and brighten the combinatorial approach to model selection.

Keywords Kernel method, Hyperkernel function, Polynomial combination, Support vector machines

1 引言

自上世纪 90 年代以来,支持向量机(Support Vector Machines, SVM)^[1,2]的研究不仅取得了众多理论成果,如快速收敛算法^[3]、超大数据集训练^[4]等,而且在机器学习、模式识别和数据挖掘等实际问题中得到了广泛的应用。核函数选择是支持向量机核方法研究和应用的关键,直接决定最终的效率和性能,成为支持向量机核方法研究的重要问题。

核函数的选择涉及两个方面:

(1)核函数的构造。实际应用中,往往结合问题的先验知识或领域知识来设计核函数。核函数的构造理论目前尚不成熟。曾有基于黎曼几何分析^[5]和 Boosting 法^[6]的研究工作。近几年,应用机器学习的观点来研究核函数构造理论逐渐成为主流,如正则化学习^[7]、半定规划学习^[8]、超核学习^[9]等。

(2)核参数的选择。核参数影响着核函数的性质。核参数选择理论研究工作一直受到广泛关注,提出了多参数学习^[10,11]、交叉验证、元学习和进化学习^[12,13],以及参数组合学习^[14]等方法。

核函数的构造和核参数的选择这两方面相辅相成,是结构与内容的关系。没有恰当的选择参数,再好的核函数也无法发挥功效;同样,没有正确的核函数形式,核参数选择也失去了意义。

本文研究核函数构造理论,提出一种核函数构造的组合

理论,将旋转不变核与平移不变核进行多项式组合,从而设计出一类超核,为支持向量机核函数的构造提供了一种新的途径。

2 单一核的局限性

支持向量经常用到两类核函数,一类为平移不变核,形式为

$$K(u, v) = r(\|u - v\|)$$

其中 $r(\cdot)$ 为任意函数,这类核函数包括高斯核、多面核、薄板样条核等;另一类为旋转不变核,形式为

$$K(u, v) = d(u \cdot v)$$

其中 $d(\cdot)$ 为任意函数,这类核函数包括多项式核、两层神经网络核等。表 1 列出了常见的核函数。使用单一核函数,支持向量机一般会陷入两种困境。其一,无法完成有效的非线性映射,如图 1 所示;其二,过学习或欠学习。下面以高斯核为例予以说明。

定理 1 当 γ 较大时($\gamma \rightarrow +\infty$),高斯核支持向量机存在过学习现象。

证明:当 γ 较大时有

$$\lim_{\gamma \rightarrow +\infty} K(u, v) = \begin{cases} 1: u = v \\ 0: u \neq v \end{cases}$$

假设训练集规模为 l ,其中,类别号为 $y = +1$ 和 $y = -1$ 的样本数分别为 l^+ 和 l^- 。又假设拉格朗日乘子

$$\lambda = \begin{cases} \lambda^+ : y = +1, \\ \lambda^- : y = -1, \end{cases} \quad 0 < \lambda^+, \lambda^- < C$$

下面求解 λ 。

由

$$\sum_{i=1}^l \lambda_i y_i = 0$$

得

$$\lambda^+ l^+ - \lambda^- l^- = 0 \quad (1)$$

根据 KKT 条件可得

$$\lambda_i [\gamma y_i (w \cdot \Phi(x_i) + b) - 1] = 0$$

$$w \cdot \Phi(x_i) + b = y_i$$

即

$$\sum_{j=1}^l \lambda_j y_j K(x_j, x_i) + b = y_i$$

令 $\gamma \rightarrow +\infty$, 可得到方程组

$$\begin{cases} \lambda^+ + b = 1 \\ \lambda^- - b = -1 \end{cases} \quad (2)$$

联立式(1),(2)解得

$$\lambda^+ = \frac{2l^-}{l}, \lambda^- = \frac{2l^+}{l}, b = \frac{l^+ - l^-}{l}$$

令

$$C > \max \left\{ \frac{2l^-}{l}, \frac{2l^+}{l} \right\}$$

则 λ 存在, 这时所有样本成为支持向量。故高斯核支持向量机存在过学习现象。 $\square$

表 1 支持向量机中常用核函数

高斯核	$K(u, v) = \exp(-\gamma \ u - v\ ^2)$
多面核	$K(u, v) = -\sqrt{\ u - v\ ^2 + \gamma^2}$
薄板样条核	$K(u, v) = \ u - v\ ^2 \ln \ u - v\ $
多项式核	$K(u, v) = (\gamma u \cdot v + c)^d$
两层 NN 核	$K(u, v) = \tanh(u \cdot v + \gamma)$

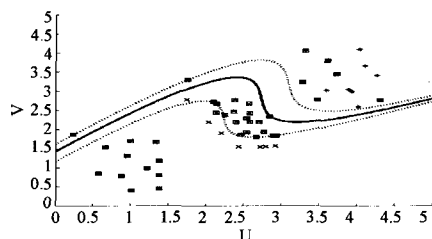


图 1 多项式核未能有效实现非线性映射

定理 2 当 γ 较小时 ($\gamma \rightarrow 0$), 高斯核支持向量机存在欠学习现象。

证明: 当 γ 较小时有, $\lim_{\gamma \rightarrow 0} K(u, v) = 1$, 此时分类面为

$$f(x) = \sum_{i=1}^l \lambda_i y_i + b = b \quad (3)$$

所有样本都被归为同一类, 故高斯核支持向量机存在欠学习现象。 $\square$

定理中讨论了 γ 的极限情况, 实际上 γ 只要达到一定大小, 过学习或欠学习现象就会产生, 图 2, 3 所示两类样本分别表示为“ $\times$ ”和“ $*$ ”, 支持向量以“ $\square$ ”标出, 理想的学习结果应为图 4 所示形式。

为避免单个核函数学习过程中出现这些问题, 我们提出了构造核函数的组合理论。

3 超核函数

在描述超核函数之前, 首先简要回顾一下 Mercer 定理,

它是判定核函数的充要条件。

令 X 是 R^n 的紧子集。假定 K 是连续对称函数, 存在积分操作数 $T_K: L_2(X) \rightarrow L_2(Y)$, 使得 $(T_K f)(\cdot) = \int_X K(\cdot, u) f(u) du$ 是正的, 也就是: $\int_{X \times X} K(u, v) f(u) f(v) du dv \geq 0$, $\forall f \in L_2(X)$ 扩展 $K(u, v)$ 到一个一致收敛的序列 (在 $X \times X$ 上), 这个序列由 T_K 的特征函数 $\Phi_i \in L_2(X)$ 构成, 归一化使得 $\|\Phi_i\|_{L_2} = 1$, 并且 $\lambda_i \geq 0$, 则有:

$$K(u, v) = \sum_{i=1}^{\infty} \lambda_i \Phi_i(u) \Phi_i(v) \quad (4)$$

定义 1 设 K 为核函数, 称 $\bar{K}$ 为超核函数, $\bar{K}(u, v) = p$

$(K(u, v), \beta, \theta)$, 其中 $p(x, \beta, \theta) = \sum_{i=1}^m \beta_i x^{\theta_i}$, $\beta, \theta \geq 0$ 。

为了保证超核 $\bar{K}(u, v)$ 是良定义的, 我们需要验证 $\bar{K}(u, v)$ 的 Mercer 性质。

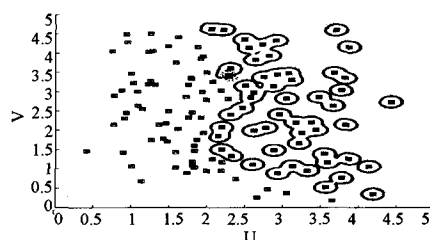


图 2 $\gamma=27$ 时高斯核支持向量机过学习, 所有样本点都成为支持向量

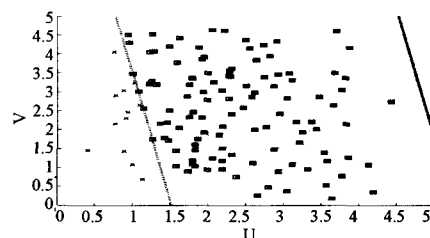


图 3 $\gamma=2^{-6}$ 时高斯核支持向量机欠学习, 几乎所有样本点在分类面同侧, 都被归为一类

引理 1 Mercer 核的非负线性组合仍为 Mercer 核。

证明: 设 $K_i, i=1, \dots, m$ 是 Mercer 核, 令

$$K(u, v) = \sum_{i=1}^m \beta_i K_i(u, v) \quad (5)$$

其中系数 $\beta_i \geq 0$ 为非负常数。因为 K_i 为 Mercer 核, 有

$$\int K_i(u, v) f(u) f(v) du dv \geq 0, \forall f \in L_2(X)$$

则对于线性组合 $K(u, v)$ 有,

$$\begin{aligned} & \int K(u, v) f(u) f(v) du dv \\ &= \int \sum_{i=1}^m \beta_i K_i(u, v) f(u) f(v) du dv \\ &= \sum_{i=1}^m \beta_i \int K_i(u, v) f(u) f(v) du dv \geq 0 \end{aligned} \quad (6)$$

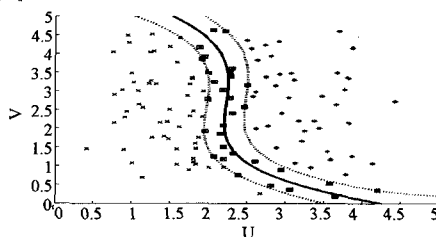


图 4 较理想的学习结果

特别地,当 $\sum_{i=1}^m \beta_i = 1$ 时 $K(u, v)$ 为 $K_i(u, v)$

凸组合 Mercer 核函数。

引理 2 Mercer 核的乘积仍为 Mercer 核。

证明:设 $K_i, i=1, \dots, m$ 是 Mercer 核,

令

$$K(u, v) = \prod_{i=1}^m \beta_i K_i(u, v) \quad (7)$$

其中系数 $\beta_i \geq 0$ 为非负常数。则对于线性组合 $K(u, v)$ 有,

$$\begin{aligned} & \int K(u, v) f(u) f(v) du dv \\ &= \int \prod_{i=1}^m \beta_i K_i(u, v) f(u) f(v) du dv \\ &= \prod_{i=1}^m \int \beta_i K_i(u, v) f(u) f(v) du dv \geq 0 \end{aligned}$$

由引理 1, 2 不难得到下面的定理。

定理 3 如果 K 为 Mercer 核, 则超核函数 $\bar{K} = p(K, \beta, \theta)$ 也为 Mercer 核。

定理 3 告诉我们, Mercer 核的多项式组合仍为 Mercer 核。因此, 以定理 3 为基础, 我们可以充分利用现有核函数 (如表 1 中常用核) 构造一类超核 $\bar{K}(u, v)$, 这类核将同时具有平移不变性和旋转不变性, 能适用于各种样本集的学习。除了原有核参数, θ 也需要选择, 可以利用现有参数选择方法进行调节。

4 实验及分析

下面通过两个分类实验来验证超核函数在支持向量机中的有效性。

实验 1 设置两类样本点作为实验对象, 标识为“×”和“*”, 各取 40 个作为训练数据, 又各取 20 个作为测试数据。图 5, 6, 7 分别给出了立方多项式核、高斯核和超核支持向量机在训练数据上学习得到的最优分类面, 支持向量加“○”标出。

$$\bar{K}(u, v) = \beta_1 \exp(-\gamma \|u - v\|^2) + \beta_2 (u \cdot v)^d \quad (8)$$

实验中的超核采用式(8)形式, 即高斯核和多项式核的线性组合, 有 4 个参数需要确定。表 2 列出了 3 种分类器的参数个数、支持向量个数、训练和测试正确率, 对比看出超核支持向量机具有更好的学习效果。

实验 2 为了验证超核支持向量机在大数据集上的表现, 我们选取 UCI 标准数据库中的 Titanic 数据集作实验对象。Titanic 数据集共有 2201 个三维样本, 三个维度分别表示地位、年龄和性别, 分类标号表示生还与否。我们取 2000 个样本作为训练集, 剩余 201 个用于测试。分别用高斯核、多项式核和超核支持向量机学习最优分类面, 超核采用形式, 共有 6 个参数需要确定, 实验中我们采用文献[14]中的方法调节各参数。表 3 列出了三种分类器的训练和测试结果。

表 2 实验 1 对比结果

核函数	#参数	#SVs	训练率	测试率
多项式	1	56	0.80	0.71
高斯	1	18	0.99	0.92
超核	4	14	0.96	0.94

$$\bar{K}(u, v) = \beta_1 \exp(-\gamma \|u - v\|^2) + \beta_2 (u \cdot v + a)^d + \beta_3 \tanh(u \cdot v + b)$$

直观上, 超核函数比单核函数有更多的参数值需要确定, 与单核比较需要花费更大的计算代价。其实不然, 因为目前

已有的参数选择方法均不是逐个调节参数的, 而是将参数向量作为调节对象, 对所有参数同时调节, 所以, 超核函数多出的几个参数只为参数向量增加了常数长度, 并不影响计算代价。

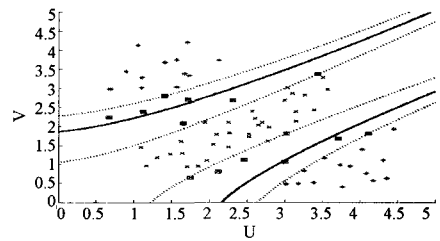


图 5 立方多项式核支持向量机最优分类面

表 3 实验 2 对比结果

核函数	参数	SVs	训练率	测试率
多项式	1	1387	0.78	0.77
高斯	1	947	0.79	0.79
超核	7	900	0.81	0.80

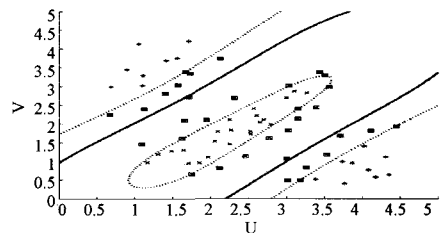


图 6 高斯核支持向量机最优分类面。

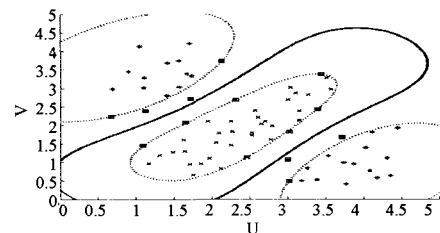


图 7 超核支持向量机最优分类面

结束语 本文提出了一种核函数构造的组合理论, 具体地构造了一类超核函数, 并讨论了超核在支持向量机中的应用。超核函数由常见核进行多项式组合而成, 具有平移不变性和旋转不变性, 能更好地适应不同问题的需要。超核函数理论上满足 Mercer 定理的要求, 在实验中也表现出了良好的性能。超核函数不仅提供了一类新的核, 也开拓了一条核函数构造的新途径, 把具有不同特性的核函数, 按照某种形式组合起来, 可得到各个单核所不具有的功能。如何将超核函数构造理论应用于其他核方法问题是进一步工作的重点。

参考文献

- [1] Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995
- [2] Burges C J C. A Tutorial on Support Vector Machine for Pattern Recognition. Data Mining and Knowledge Discovery, 1998, 2: 121-167
- [3] Platt J, Burges C J C. Fast Training of Support Vector Machines Sequential Minimal Optimization // Shölkopf B, Smola A J, eds. MIT Press, 1998

(下转第 166 页)

显示出利用两种补全策略共同启发起到了正面作用。由此可以推测出 GS 和 EM 数据补全策略共同启发的策略起到了良好作用,能够相对促使 SGS-EM-PACOB 算法平稳地得到较优网络结构。

4.3 实验小结

我们还对 INSURANCE 等数据集进行了实验,其结果与本节反映的结论基本一致。限于篇幅,只列举了 ALARM 数据集的情况。

综上所述可以得出由 GS 和 EM 两种数据补全策略联合的 SGS-EM-PACOB 算法无论在数据拟合度上,还是网络结构上都表现出较好的特性。更重要的是该算法能够平稳地收敛到理想结果,而不会像 SGS-PACOB 和 SEM-PACOB 算法那样较高概率地收敛到局部最优。

结束语 本文建立了具有数据缺失学习贝叶斯网混合启发模型的算法,把 GS 和 EM 两种数据补全策略较好地联合起来共同启发,能够使 SGS-EM-PACOB 算法平稳地收敛到理想结果,相对于单一数据补全策略的算法取得了改进结果。但同时也注意到 EM 补全策略计算量相当庞大,当面对大规模数据缺失或者大规模例子训练数据集时是一个瓶颈。未来的研究中,可以找一计算量相对较少的数据补全策略进行更深更紧的联合。比如 C. Riggelsen 提出的 MBP 算法中补全策略就是一个不错的选择。

参考文献

- [1] Heckerman D. Bayesian networks for data mining. Technical Report, MSR-TR-97-02. Microsoft Research, Redmond, 1997
- [2] 张连文,郭海鹏. 贝叶斯网络引论. 北京:科学出版社,2007
- [3] Tanner M, Wong W. The calculation of posterior distributions by data augmentation. J. of the Am. Stat. Assoc, 1987, 82(398): 528-540
- [4] Friedman N. Learning belief networks in the presence of missing values and hidden variables // Proc. of the 14th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997: 125-133
- [5] Friedman N. The Bayesian Structural EM Algorithm // Proc. of the 14th Int'l Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1998: 129-138
- [6] 王双成,苑森森. 具有丢失数据的贝叶斯网络结构学习研究. 软件学报, 2004, 15(7): 1042-1048
- [7] Riggelsen C, Feelders A. Learning Bayesian network models from incomplete data using importance sampling // Cowell R G, Ghahramani Z, eds. Proc. of Intelligence and Statistics. 2005: 301-308
- [8] Riggelsen C. Learning Bayesian Networks from Incomplete Data: An Efficient Method for Generating Approximate Predictive Distributions. Department of Information and Computing Sciences, Universiteit Utrecht, 2007
- [9] Geman S, Geman D. Stochastic relaxation, Gibbs distributions and Bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1984, 6(6): 721-742
- [10] Heckerman D. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. MSR-TR-94-09. Microsoft Research, 1995
- [11] 吕强,高彦明,钱培德. 共享信息素矩阵:一种新的并行 ACO 方法. 自动化学报, 2007, 33(4): 418-421
- [12] 潘吉斯,吕强,王红玲. 一种并行蚁群 Bayesian 网络学习的算法. 小型微型机计算机系统, 2007, 28(4): 651-655
- [13] de Campos L M, Fernandez-Luna J M, Gamez J A, et al. Ant colony optimization for learning Bayesian networks. Int. J. Approx. Reasoning, 2002, 31(3): 291-311
- [14] Cooper G F, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 1992, 9(4): 309-348
- [15] Bouckaert R R. Probabilistic Network Construction Using the Minimum Description Length Principle // Lecture Notes in Computer Science. 1993
- [16] Kruse R, Borgelt C. An empirical investigation of the k2 metric // Proc. 6th European Conf. on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. 2001: 240-251
- [17] 廖学清,吕强. 试析数据缺失下学习贝叶斯网中初始网络的选择. 计算机科学专刊录用
- [18] with Hyperkernels. Journal of Machine Learning Research, 2005, 6: 1043-1071
- [10] Chapelle O, et al. Choosing Multiple Parameters for Support Vector Machines. Machine Learning, 2002, 46(11): 131-159
- [11] Chapelle O, Vapnik V N. Model selection for support vector machines // Proceedings of Thirteenth Annual Conference on Neural Information Processing Systems. Cambridge, MA, MIT Press, 1999: 230-236
- [12] Frauke F, Christian I. Evolutionary tuning of multiple SVM parameters. Neurocomputing, 2005, 64: 107-117
- [13] Soares C, Brazdil P B, Kuba P. A Meta-Learning Method to Select the Kernel Width in Support Vector Regression. Machine Learning, 2004, 54: 195-209
- [14] Liao S, Jia L. Simultaneous Tuning of Hyperparameter and Parameter for Support Vector Machines // Proceedings of the Eleventh Pacific-Asia Conference on Knowledge Discovery and Data Mining. New York. Springer-Verlag, 2007: 162-172

(上接第 150 页)

- [4] Collobert R, Bengio S. SVMtorch: Support Vector Machines for Large-Scale Regression Problems. Journal of Machine Learning Research, 2001, 1: 143-160
- [5] Amari S, Wu S. Improving Support Vector Machine Classifier by Modifying Kernel Function. Neural Networks, 1999, 12(66): 783-789
- [6] Crammer J K, Elisseeff A, Shawe-Taylor J. Kernel Design using Boosting // Proceedings of Sixteenth Annual Conference on Neural Information Processing Systems. Cambridge, MA, MIT Press, 2002: 537-544.
- [7] Charles A M, Pontil M. Learning the Kernel Function via Regularization. Journal of Machine Learning Research, 2005, 6: 1099-1125
- [8] Lanckriet G, et al. Learning the Kernel Matrix with Semidefinite Programming. Journal of Machine Learning Research, 2004, 5: 27-72
- [9] Cheng S O, Smola A J, Williamson R C. Learning the Kernel