HSC 分类法及其在海量数据分类中的应用

任力安[1,2],何 清[1],史忠植[1]
(1. 中科院计算技术研究所智能信息处理重点实验室,北京 100080;2. 中国科技大学研究生院计算机学部,北京 100039)

摘 要: 使用支持向量机对非线性可分数据进行分类的基本思想是将样本集映射到一个高维线性空间使其线性可分.本文则基于 Jordan 曲线定理,提出了一种通用的基于分类超曲面的分类方法,简称 HSC 分类法,它是通过直接构造分类超曲面,根据样本点关于分类曲面的围绕数的奇偶性进行分类的一种新分类判断算法,与 SVM 方法相比,不需要考虑使用何种核函数,不需要做升维变换,直接解决非线性分类问题.对数据分类应用的结果说明:HSC 可以有效地解决非线性数据的分类问题,并能够提高分类效率和准确度.

关键词: 支持向量机; 分类超曲面; Jordan 曲线定理; HSC 分类法
中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112 (2002) 12-1870-03

HSC Classification Method and Its Applications in Massive Data Classifying

REN Li-an[1,2], HE Qing[1], SHI Zhong-zhi[1]
(1. *The Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080, *China*;
2. *Graduate College of University of Science and Technology of China, Beijing* 100039, *China*)

Abstract: The main idea of SVM used for classifying nonlinear separable data is to map the data into higher dimension linear space in which the data can be separated by a hyper plane. Based on Jordan Curve Theorem, a universal classification method based on hyper surface, which is called HSC classification, is put forward in this paper. The classification hyper surface is directly made to classify massive data according to whether the wind number is odd or even. It is a novel approach that does not need to make mapping from lower dimension space to higher dimension space and considering kernel function too. It can directly solve the nonlinear multiclasses classifying problem. The test reports show that HSC can efficiently and accurately classify large data.

Key words: support vector machine; separating hyper surface; Jordan Curve Theorem; HSC classification method

1 引言

机器学习研究获取新知识、新技巧,重组已经发现的知识的计算方法,它是人工智能中的基本问题,其中分类问题是许多其它问题的基础和核心.Vapnik 等人提出的支持向量机(Support Vector Machine,SVM)方法对小样本、非线性和高维特征具有很好的分类性能.SVM 的基本思想是通过内积函数定义的非线性映射(核函数)将非线性可分样本集映射到一个高维线性空间,在计算上,借助二次规划求解支持向量需要反复计算一个 m 维的内积矩阵(其中 m 是样本个数),所需要的计算开销是相当大的,在 PC 机上其处理样本规模一般为 4000 个样本[6].因而对海量数据进行分类几乎是不可能的.1999 年,我国学者张铃与张钹教授提出二次规划优化函数的几何方法,采用球面投影函数作为非线形映射,完成样本点的分类问题,即将计算分类超平面的问题转换为计算样本点两两之间距离所构成的距离空间上的覆盖问题[2,3].基于邻域的方法在计算样本之间内积的同时,判断哪些样本可以删除,每删除一个样本就意味着使得内积矩阵降低一维,因此,这个考虑特别适合内积矩阵阶数过大的情形.在文献[6]中,张文生,丁辉,王珏对邻域方法作了详尽的数学分析和几何解释,并给出了三种典型的求支持向量的邻域算法.

Vapnik 认为九十年代以他为代表的研究只是返回到感知机年代.感知机的线性特性,虽然使其不能解决非线性函数的优化问题,但是,其算法却相对简单得多.是否可以使用感知机原理解决非线性优化问题呢? 在历史上,为解决这个问题,在技术上曾经有过多次尝试,六十年代 Widrow 与 Hoff 提出的自适应线性元件神经网络 Adaline,以及由多个 Adaline 组成的 Madaline 就是这种尝试之一[11],他们试图使用多个超平面的划分来解决非线性划分问题,但如何求出这些自适应线性元件却是一个一直未解决的问题.

实际上,在解决非线性问题时,支持向量机张铃与张钹教授基于邻域的空间划分方法是在向高维空间作升维变换,最终构成分类超平面,如果这时考虑这个过程的逆变换-降维变换,则分类超平面就变形为分类超曲面了;因此,在低维空间来看,他们工作的本质是在找分类超曲面.

是否能找到一种方法,不通过向高维空间作升维变换,而直接地解决非线性分类问题呢? 本文提出的 HSC 分类方法

收稿日期:2001-08-20;修回日期:2002-04-26
基金项目:国家自然科学基金(No.60173017,90104021);北京市重点自然科学基金(No.4011003)

则对此作了一种新的尝试.

2 基于分类超曲面的分类判别方法的理论基础

HSC 分类判别方法基于拓扑学中的 Jordan 曲线定理[12],定理如下.

Jordan 曲线定理　设 $X \subset R^3$ 是闭子集,X 同胚于球面 S^2,那么它的余集 $R^3 \setminus X$ 有两个连通分支,一个是有界的,另一个是无界的,X 中任何一点的任何邻域与这两个连通分支均相交.

上述定理可推广到高维空间.

定理　(高维空间的 Jordan 曲线定理)若 $X \subset S^n$ 同胚于球面 S^m,那么 $m \leq n$ 否则 $X = S^n$.若 $m < n$,余集的同调群为

$$H_k(s^n \setminus X) \cong \begin{cases} Z \oplus Z, \text{若 } m = n-1 \text{ 且 } k = 0, \\ Z, \text{若 } m < n-1 \text{ 且 } k = 0, \\ 0, \text{其余} \end{cases}$$

特别地当 $m = n-1$ 时 $S^n \setminus X$ 由两个连通分支组成,当 $m < n-1$ 时,只有一个连通分支.

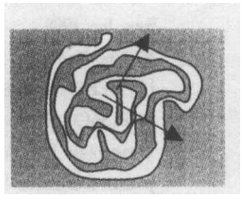

图 1　分类判别图示

Jordan 曲线定理表明任何由 $n-1$ 维球面经连续变形得到的双侧闭曲面都把 n 维空间分成两个区域——一个外部和一个内部,这种曲面可用于分类,我们称之为分类超曲面.给定一个点 x,如何判断它在分类超曲面 X 的内部,还是在外部呢?判断方法是:$x \in X$ 的内部$\Leftrightarrow$自 x 引出的射线与 X 的相交数(即 X 关于 x 的围绕数)为奇数,$x \in X$ 的外部$\Leftrightarrow$自 x 引出的射线与 X 的相交数为偶数.如图 1 所示.现在问题的关键在于如何获得分类超曲面.

3 HSC 分类方法的实现

根据上节所述基本思想,我们设计出下面的 HSC 分类方法的训练算法及分类算法.下述算法是针对三维数据的,关于二维数据的处理在文献[13]中作了详细阐述,本文主要针对三维数据特点给出算法,由于有高维空间的 Jordan 曲线定理作理论基础,上述方法有望推广到更高维数据的处理.

(1)学习算法

设样本空间为归一化立方体.

第 1 步,将此区域等分,其中任意一单元区域内至多包含一个训练样本点;

第 2 步,根据所含样本类别,将各单元区域表示为以下结构:单元区域:区域标号、类别标志、边界面链表;

第 3 步,找出所有相邻同类单元区域,以链表形式存放;

第 4 步,对相邻同类单元区域作边界相交操作,即消去公共边界;

第 5 步,以链表形式存储各连通区域完整边界链表——分类超曲面.

(2)细化方案

若一单元区域内存在多个训练样本;

第 1 步,设计训练样本分层链表结构.训练样本:同层样本链表、层次标志、下层样本链表;

第 2 步,将下层样本所在单元区域细化,并进行归一化操作;

第 3 步,转至学习算法第 1 步,继续标注,再合并边界,存放边界链表;循环完成此训练过程.

(3)分类算法

对一待识别的样本进行分类;

第 1 步,导入分类曲面链表;

第 2 步,在分类空间中,由样本向空间任意选定的一维方向引射线,分别记录与各类分类曲面的相交数;

第 3 步,根据与分类曲面相交数的奇偶性,判断出此样本所属类别.

如果样本所在单元区域已经细化,则将样本坐标单位化,放入细化区域,继续进行上述的分类过程.

4 实验与结果分析

4.1 构造测试数据

双螺旋分类问题[3]:两条螺旋线 K_1 和 K_2(极坐标形式)

$$K_1: \rho = \theta$$

$$K_2: \rho = \theta + \pi, \frac{\pi}{2} \leq \rho \leq 8\pi$$

$$Z = \rho$$

构造三维训练样本集合及测试样本集合如图 2 所示.

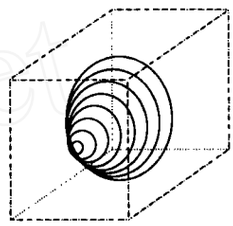

图 2　数据视图

4.2 实验过程

使用双螺线两类样本数据,测试 HSC 分类方法在三维数据分类中的性能*,所获得分类超曲面及其对样本点的覆盖情况如图 3 所示.

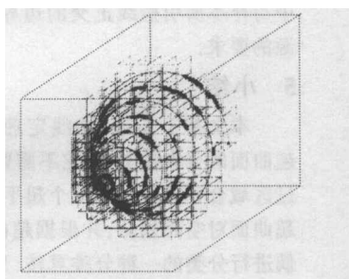

图 3　三维测试数据与结果

4.2.1 大规模样本实验结果　(1)三类测试结果见表 1;(2)两类大规模分类测试结果见表 2.

4.2.2 两类小样本训练,大样本分类测试结果　见表 3

表 1　三类样本测试结果

训练样本点个数	测试样本点个数	分类所需时间	分类正确率(%)
13503	1350003	18m 29s	99.80
13503	270000	3m 29s	99.80

* 本节所有数据均在以下测试环境中获得:奔腾 III,733MHz;内存:256M;操作系统 Microsoft Windows 2000 Server, Service Pack 2;数据库 Microsoft Access 2000;编译环境 Visual C++ 6.0, Service Pack 4.

表 2 两类大规模样本测试结果

训练样本点个数	测试样本点个数	分类所需时间	分类正确率(%)
5,400,000	10,800,000	2h 35m 48s	100.00
10,800,000	22,500,002	5h 14m 51s	100.00
22,500,002	60,000,000	14h 25m 8s	100.00

表 3 两类小样本训练,大样本分类测试结果

训练样本点个数	测试样本点个数	分类所需时间	分类正确率(%)
5,402	54,002	45s	99.82
5,402	540,000	7m 42s	99.81
27,002	540,000	7m 34s	99.98
54,002	540,000	7m 33s	100.00
54,002	5,400,000	1h 15m 59s	100.00
54,002	22,500,002	5h 15m 19s	100.00

当同类样本点在有限个连通分支分布时,学习算法与分类算法的算法复杂度都是多项式的.对于海量多类数据(10^7),HSC 分类方法可得到较高的计算速度和准确度,同时对计算机资源要求很低,而传统的 SVM 不具备有这种优点.另外表 3 测试样本点是由同一公式构造的另一样本集合(样本数量为训练样本的 10 倍以上),小样本训练大样本测试结果表明 HSC 分类方法的泛化能力较好.为保证分类曲面的连续性,在实际学习算法中,链表需同时记录同区域内同类训练样本,但在样本集规模很大的情况下,可对细化的层次进行控制,即在训练过程中将不影响分类曲面生成的样本删除,可保证计算速度.此外,在记录分类曲面时,只需存储对分类过程中与样本所引射线正交的边界面,可进一步减少对计算机资源的要求.

5 小结

本文基于 Jordan 曲线定理,提出了一种通用的基于分类超曲面的分类法 HSC,它不需要考虑使用何种核函数,旨在通过区域合并计算获得多个超平面组成的双侧闭曲面作为分类超曲面对空间划分,并根据超曲面关于样本点的围绕数的奇偶进行分类的一种分类算法.所获得的分类超曲面在一定意义下可以看作以超平面为自适应线性元件的神经网络.这种方法使得基于非凸的超曲面的分类判别变得直接、简便、易行,同时避免使用 SVM 方法向高维空间的升维变换,该法对噪声的干扰虽不能完全排除,但可以把噪声影响限制在局部小范围,最近的文献采用 SV 的近邻算法处理了二维双螺旋数据 10^4 个样本,本文采用 HSC 方法处理了 6×10^7 个三维双螺旋样本点,实验结果证明:HSC 分类方法可以有效地解决三维数据大量的分类问题,并能够在一定程度上提高分类效率和准确度.应当指出,本文所讨论方法是对直接解决非线性分类问题的一种尝试,此方法的一个前提是同类样本点应具有在有限个连通分支分布的特点,但与连通分支的形状无关.此种方法在处理如此分布的数据集时,有很好的效果.

参考文献:

[1] Vapnik V N. Support vector method for function approximation, regression estimation and signal processing [J]. Neural Information Processing Systems, 1996, 9: 281 - 287.

[2] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag, 1995.

[3] Ling Zhang, Bo Zhang. A geometrical representation of McCulloch-Pitts neural model and its applications [J]. IEEE Transactions on Neural Networks, 1999, 10(4): 925 - 929.

[4] 张学工.关于统计学习理论与支持向量机[J].自动化学报, 2000, 26(1): 32 - 42.

[5] 张学工译.统计学习理论的本质[M].北京:清华大学出版社, 2000.

[6] 张文生,丁辉,王珏.基于邻域原理计算海量数据支持向量的研究[J].软件学报, 2001, 12(5): 711 - 720.

[7] 边肇祺等.模式识别(第二版)[M].北京:清华大学出版社, 2000.

[8] Vapnik V N. Statistical Learning Theory [M]. New York: John Wiley & Sons, Inc, 1998.

[9] Widrow B, Winter R G. Layered neural nets for pattern recognition [J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1988, 36(3): 1109 - 1118.

[10] Christopher J C Burges. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121 - 167.

[11] Widrow B, Hoff M. Adaptive switching circuits [J]. IRE Wescon Convension Record, 1960, 4: 96 - 104.

[12] William Fulton, Algebraic Topology: A First Course [M]. New York: Springer-Verlag, 1995.

[13] Qing He, Zhong-Zhi Shi, Li-An Ren. The classification method based on hyper surface [A]. IEEE International Joint Conference on Neural Network (IJCNN) [C]. Hawaii: 2002 World Congress on Computational Intelligence, 2002. 206 - 211.

作者简介:

任力安 男,1975 年生于陕西西安,硕士研究生,主要研究方向:人工智能、模式识别、专家系统.

何 清 男,1965 年生于河北深泽,副研究员,博士后,主要研究方向:模糊集理论、人工智能、数据挖掘、机器学习.

史忠植 男,1941 年生于江苏,研究员,博士生导师,主要研究方向:人工智能、智能软件、神经计算.