分类超曲面方法在海量数据分类中的应用*)

A Method Based on Separating Hyper Surface and Its Applications in Massive Data Classifying

任力安[1] 何 清[2] 史忠植[2]

(中国科技大学研究生院计算机学部 北京 100039)[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]

Abstract It is quite difficult to classify large data by using the support vector machine. To solve the problem, based on Jordan Curve Theorem, a universal classification method based on hyper surface 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 need not make mapping from lower dimension space to higher dimension space and need not consider kernel function too. It can directly solve the nonlinear classifying problem. The experimental results show that the separating hyper surface method can effectively solve the problem of classification of huge data and it is clear that the classifying efficiency and accuracy have been improved by using the method.

Keywords Machine Learning, Separating hyper surface, Jordan curve theorem, Massive Data

1 引言

人的智慧中一个很重要的方面是从实例学习的能力,通过对已知事实的分析总结出规律,预测不能直接观测的事实。在人们对机器智能的研究中,用机器(计算机)来模拟这种学习能力,这就是我们所说的基于数据的机器学习问题,它是现代智能技术中的重要方面,其研究从观测数据(样本)出发寻找规律,利用这些规律对未来数据或无法观测的数据进行预测(分类)。统计机器学习理论为机器学习问题建立了一个较好的理论框架,也发展了一种新的通用学习算法--支持向量机(SVM),其关键思想是将在低维空间非线性可分的数据通过非线性函数(核函数)映射到一个非常高维的特征空间,并在这个新的线性空间构筑分类超平面[1]。这一结果相应于原始空间就是通过分类超曲面进行分类判别。

在计算上,SVM 通过求解二次规划所需要的计算开销是相当大的,同时,给定一个样本集(函数),通过计算寻找一个合适的非线性变换也不是一件轻松的事情。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?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 Aadaline,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,这个考虑是重要的,但是,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。

2 基于分类超曲面的分类判别方法

实际上,在解决非线性问题时,支持向量机是在向高维空间做升维变换,最终构成分类超平面。是否能找到一种方法,不通过向高维空间做升维变换,而直接地解决非线性分类问题呢?本文提出的基于分类超曲面的分类方法对此做了一种新的尝试。

2.1 理论基础

基于分类超曲面的直接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 $R^3\backslash X$ 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。

在三维状态下,Jordan 曲线定理表明任何由球面经连续变形得到的双侧闭曲面都把三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,这就是本文中要研究的分类超曲面。给定一个点,如何判断它在分类曲面的内部还是在外部呢?

分类判别定理 设 $X\subset R^3$ 是平面的闭子集,X 同胚于球面 S^2,那么它的余集 $R^3\backslash X$ 有两个连通分支,一个是内部,另一个是外部,任取 $x\in R^3\backslash X$,则:

$x\in X$ 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。

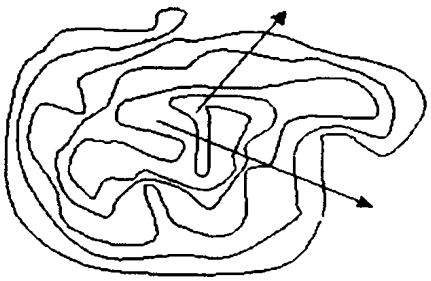

图 1 分类判别定理图示

上述定理可推广到更高维空间。

基于拓扑学中的 Jordan 曲线定理,通过与球面同胚的双

*)本文得到国家自然科学基金资助(批准号:60173017,90104021)与北京市自然科学基金资助(课题号 4011003)。任力安 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 博士后。主要研究方向:模糊集理论、人工智能、数据挖掘。史忠植 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。

侧闭曲面作为分类超曲面(Separating Hyper Surface)对空间进行划分。分类超曲面可以由多个超平面构成,而点属于超曲面内部还是外部取决于该点引出的射线与超曲面相交为奇数还是偶数,该判别方法使得基于非凸的超曲面的分类判别变得更直接、简便。这样,分类超曲面的构造方法就是主要问题了。

2.2 基于分类超曲面的构造与分类基本过程

根据上述定理我们提出如下基于分类超曲面的分类法,整个过程如下:

第 1 步,设已给的样本点落在一个长方体区域中;

第 2 步,将此区域划分成若干小区域,使每个小区域至多含一个样本点;

第 3 步,根据样本点的类别对每个含样本点的小区域边界进行标定,构成含类别分量的边界向量链表;

第 4 步,合并相邻同类区域边界,获得若干小平面封闭组成的分类超曲面,并以链的形式存储分类超曲面;

第 5 步,输入新样本点,计算该点关于以上分类超曲面的围绕数,根据围绕数判定该样本点所在的类;另一种简便方法是选择适当的由待定点出发的射线,通过射线与分类超曲面的相交数(即分类超曲面关于样本点的围绕数)的奇偶性判断样本点所在的类;若不能判断,就围绕该点做一个小矩形,并对边界进行标定,之后转入第 4 步。

3 数据分类

在双螺旋公式的基础上,构造三维训练样本集合及测试样本集合(参数方程):

$$K_1:\begin{cases}x=\rho\cos\rho\\ y=\rho\sin\rho\\ z=\rho\end{cases}\quad K_2:\begin{cases}x=\rho\cos(\rho+\pi)\\ y=\rho\sin(\rho+\pi)\\ z=\rho\end{cases}$$

其中,$\frac{\pi}{2}\leqslant\rho\leqslant 8\pi$。

本文主要针对三维数据特点给出算法,下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(如图 2)

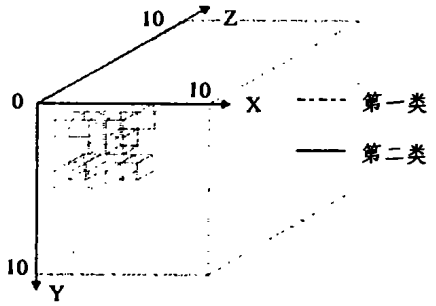

图 2 三维实现

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做边界相交操作,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。训练所得分类链表及其对样本点的覆盖情况,见图 3。

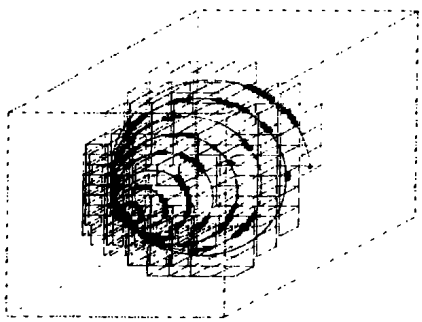

图 3 分类链表及其对样本点的覆盖情况

(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。

实验样本测试结果见表 1、表 2、表 3 所示。

表 1 大规模样本训练结果

训练样本点个数	训练所需时间	查全测试用时	查全测试正确率(%)
5,400,000	1h 2m 39s	1h 17m 53s	100.00
10,800,000	2h 6m 23s	2h 34m 45s	100.00
22,500,002	4h 23m 18s	5h 22m 26s	100.00

表 2 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 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①	分类所需时间	分类正确率(%)
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

①测试样本点由双螺线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)。

4 算法与实验结果分析

根据分类超曲面的思想，我们给出以上算法实现过程，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。通过实验我们可以看到基于分类超曲面的海量数据分类法，对解决非线性分类问题是有效的，具有较好的通用性，从实验结果看分类准确率高，特别是查全测试准确率高。同时，采用小规模样本构造超曲面，对大规模样本进行分类的结果表明，基于分类超曲面的直接分类方法有很好的推广能力。对噪声的干扰此方法虽然不能消除，但可以把噪声控制在局部范围。事实上，这种分类方法能有效地解决在有限区域分布的海量数据的非线性分类问题，而实际数据往往具备在有限个连通分支分布的特征。

结论 本文基于 Jordan 曲线定理，提出了一种通用的基于超曲面的直接分类方法，并由此提出了分类超曲面的思想。通过实验，可以证明采用基于分类超曲面的分类方法，对非线性数据进行分类是完全可行的，而且在处理大规模样本数据（10^7）时，分类速度和正确率都可以得到保证。另外，无须考虑矩阵的复杂计算，因而可以大大节省计算资源，有效提高了分类效率。

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样本点的分布具有保证数据集在有限个连通分支分布的特征。因此，这种方法在处理如此分布的海量数据集时，有较好的效果。

参 考 文 献

1 Vapnik V N. Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. Neural Information Processing Systems, Vol. 9. MIT Press, Cambridge, MA

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

3 Zhang Ling, Zhang Bo. A Geometrical Representation of McCulloch-Pitts Neural Model and Its Applications. IEEE Trans. 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. 自动化学报, 2000, 26(1)

5 张学工译. 统计学习理论的本质. 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. 软件学报, 2001, 12(5)

7 边肇棋等. 模式识别(第二版). 北京：清华大学出版社, 2000

8 Vapnik V N. Statistical Learning Theory. J. Wiley, New York, 1998

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

10 Burges C J C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2(2)

11 Widrow B, Hoff M. IRE Wescon Convension Record. Part 4, Institute of Radio Eng., New York, 1960. 96～104

12 Fulton W. Algebraic Topology A First Course, Springer-Verlag 1995

（上接第 29 页）

程。应用程序通过调用 tlqRegisterEvent()API 提出注册请求，EMS 核心会自动负责为新注册的事件分配一个类型号（事件类型）；随后的应用可以通过这个类型号或事件名称（在注册事件时提供）对事件进行发布和订阅。

对事件的注销是一个相反的过程。事件被注销后，EMS 会试图通知所有订阅了该事件的应用程序，然后清除同该类事件相关的资源。

EMS 实现了前述 GSRA 算法，EMS 将 GSRA 算法中的标记作为一个特殊的系统事件，当某个 EMS 按系统请求需要对系统状态进行快照时（比如一个交易的开始），它将启动 GSRA 算法。EMS 负责对这个特殊事件的响应。系统中任何一个 EMS 接收到 GSRA 标记时，将按照 GSRA 算法记录自己的状态，然后发送标记到其它相关节点。需要注意的是在两个 EMS 系统之间存在通讯通道是双向的，因此它实际代表了 GSRA 中的两个通道。本文将不对 TongLINK/Q EMS 具体实现细节进行详细讨论，结果表明，由于 TongLINK/Q EMS 采用 GSRA 算法，系统开销小，灵活方便。由于 TongLINK/Q EMS 只是在参照分布式事件规范、Java 消息规范中的订阅/发布模型后针对自身需要而设计的，严格来讲 TongLINK/Q EMS 还不是一个完全的分布式系统，也不是完全满足 GSRA 算法的理想条件，例如，在 TongLINK/Q 中，作为通讯使用的通道实际上容量有限，也不能保证无错缓冲，因此，在采用 GSRA 算法时还有一定限制。

参 考 文 献

1 Bernstein P A. Middleware: A Model for Distributed Services. Communications of the ACM, 1996, 39(2): 86～97

2 Stephen Williams Global State Recording Algorithm: GSRA from http: // courses. cs. vt. edu/～cs5204/fall00/Summaries/GlobalState/global_state. html

3 Eckerson W W. Three Tier Client/Server Architecture: Achieving Scalability, Performance, and Efficiency in Client Server Applications. Open Information Systems, 1995, 3(20)

4 Richard S. Middleware Demystified. Datamation, 1995, 41(6): 41～45

5 Distributed Event Specification, Revision 1. 0 Beta. Released on: July 17, 1998, Copyright 1998 Sun Microsystems, Inc. 901 San Antonio Road, Palo Alto, CA 94303, U. S. A

6 Java™ Message Service Specification ("Specification"), Version: 1. 0. 2. Released on: 12/17/99, Copyright 1999 Sun Microsystems, Inc. 901 San Antonio Road, Palo Alto, CA 94303, U. S. A

7 RFC 1862. Report of the IAB Workshop on Internet Information Infrastructure. Oct. 12-14, 1994

8 RFC 2543. SIP: Session Initiation Protocol

9 RFC 2768. Network Policy and Services: A Report of a Workshop on Middleware

10 RFC 2969. Wide Area Directory Deployment - Experiences from TISDAG

11 TUXEDO System 6 Course for Application Developers. BEA Systems, Inc., Dec. 1996

12 MQSeries Family June Product Announcement White Paper. IBM Inc, June 1999

13 Stevens W R. Advanced Programming in the UNIX Environment. 机械工业出版社, 1999

14 东方通科技公司. TongLINK/Q 技术白皮书. 1999