

基于覆盖的分类算法研究进展

何 清 史忠植

(中国科学院计算技术研究所智能信息处理重点实验室 北京 100080)

摘 要 理解数据与感知数据密切相关. 覆盖学习算法在低维空间往往能模拟人的视觉感知来表示数据分布. 文中综述了基于覆盖的分类算法的研究进展, 特别对基于超曲面的覆盖分类算法进行了详细阐述和分析, 并指出了基于超曲面的分类算法进一步研究的方向.

关键词 覆盖算法; 基于超曲面的分类方法; 极小一致集; 机器学习

中图法分类号 TP181

The Advances in the Covering Based Classification Algorithms

HE Qing SHI Zhong-Zhi

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

Abstract The understanding of data is highly relevant to how one senses and perceives them. The covering learning algorithms can always simulate human visual cognition to represent the data distribution in the low dimension space. The advances in the area of covering based classification algorithms are summarized. Specially, the Hyper Surface Classification is introduced and analyzed in detail. Moreover, the future research directions are pointed out.

Keywords covering algorithm; hyper surface classification; minimal consistent subset; machine learning

1 引 言

分类算法研究是机器学习的核心研究内容, 分类能力是人类智能的最显著特征之一. 机器学习在最近三十多年取得了很大进展. 1971年, 两位前苏联数学家 Vapnik 与 Chervonenkis 提出了一种基于 VC 维度的对空间划分的理论^[1]. 1984年 Valiant 提出了可学习理论 PAC (Probability Approximately Correct), 并将可学习与计算复杂性联系在一起^[2]. 在 Valiant 学习理论中, 有两种学习复杂性测度. 一是样本复杂性. 这是指随机实例的数目, 用以产生具有高的概率和小的误差的结果. 二是

计算复杂性. 它被定义为最坏情况下给定数目的样本产生假设所要求的计算时间. 这两种复杂性在对分类算法的研究中起着重要作用. 1986年 Blumer 等人证明了 VC 维度与 Valiant 的“可学习理论”(PAC)之间的联系^[3]. 关于 PAC 的研究派生出被称为“计算学习理论 COLT”(Computational Learning Theory)的学派, 现这方面的国际会议已定期召开^[4]. 1995年 Vapnik 出版了《统计学习理论》(The Nature of Statistical Learning Theory)一书. 在理论上, 这是继 Duda 等人在 20 世纪 60 年代奠定统计模式识别理论之后^[5] (第二版中译本是文献^[6]), 对统计模式识别最为完整的研究. 这个理论的基础之一是 VC 维度 (Vapnik-Chervonenkis

收稿日期: 2007-01-15; 修改稿收到日期: 2007-06-15. 本课题得到国家自然科学基金(60435010, 90604017, 60675010)、国家“八六三”高技术研究发展计划项目基金(2006AA01Z128)、国家“九七三”重点基础研究发展规划项目基金(2003CB311004)和北京市自然科学基金(4052025)资助. 何 清, 男, 1965年生, 研究员, 博士生导师, 主要研究领域为模糊数学、机器学习、人工智能. E-mail: heq@ics.ict.ac.cn. 史忠植, 男, 1941年生, 研究员, 博士生导师, 主要研究领域为人工智能、机器学习和分布式人工智能. E-mail: shizz@ics.ict.ac.cn.

Dimension). 关于统计学习理论最新综述包括在文献[7]中. 事实上, 对 PAC 的研究一直是理论性的、存在性的, Vapnik 的这个研究却是构造性的, 并将感知机的研究包括在其中, 他将这种模型称为支持向量机(Support Vector Machine, SVM). 在 SVM 的研究中我国学者做了大量的工作, 如文献[8-11]. 与此同时, 在基于覆盖思想的分类学习算法方面, 我国学者在最近十年相继提出了一些很有价值的分类学习方法. 本文着重分析基于覆盖的分类学习算法.

2 基于覆盖的分类学习算法概述

张铃、张钊教授给出了 M-P 神经元的几何意义, 通过球面投影变换将神经网络的最优设计问题转化为某种最优覆盖问题^[12]. 他们把神经元与几何上样本的球形邻域对应起来. 借助这种几何直观方法, 他们将多层前向神经网络作为“分类器”来进行设计, 并把这种设计看成用若干“球形邻域”, 将输入 $x^i (i=1, 2, 3, \dots, d)$ 按其所属的类别划分开来. 一种最简单的设计方法是用一组球形邻域将该类的 $x^i (i=1, 2, 3, \dots, d)$ 覆盖住, 而不覆盖不属于该类的 $x^i (i=1, 2, 3, \dots, d)$, 于是不同类的输入被不同组的球形邻域所覆盖. 然后再将属于同组的球形邻域对应的神经元输出用“或门”集中起来, 这样就给出一个分类器的设计. 一个覆盖中的球形邻域对应着一个隐层元. 最小覆盖对应于隐层元最少的神经网络结构. 他们还指出任何一种求(次优)最小覆盖的方法与球形邻域法相结合, 都能给出一个神经网络的学习算法, 称这种学习算法为邻域覆盖算法. 张铃、张钊教授还给出了邻域覆盖算法和交叉覆盖算法以及改进的函数覆盖算法和核覆盖算法. 继而将核函数法与构造性学习的覆盖算法相融合给出了一种新的核覆盖算法, 克服了以上两种模型的缺点, 其具有运算速度快、精度高、鲁棒性强的优点, 还能给出风险误差上界与覆盖个数的关系^[13]. 另外他们还提出了双交叉覆盖增量学习算法(BiCovering Algorithm, BiCA)^[14]以及多层前向网络的交叉覆盖设计算法^[15]. 该算法进一步通过构造多个正反覆盖簇, 使得网络在首次构造完成后还可以不断地修改与优化神经网络的参数与结构, 增加或删除网络中的节点, 进行增量学习. 基于神经元的几何理解非常直观地证明了 1989 年 Funabashi、Arai 和 Hecht-

Nielson 曾分别证明的三层前向神经网络可以任意逼近紧集上的连续函数和平方可积函数的定理. 邻域覆盖算法已用于手写汉字识别和金融股市预测. 张铃教授指出, SVM 算法与神经网络的基于规划的算法是等价的, 即在样本集是线性可分的情况下二者求到的均是最大边缘; 不同的是前者通常用拉格朗日乘子求解的复杂性将随规模呈指数增长而后的复杂性是规模的多项式函数. 文献[16]对上述工作作了详细介绍.

王守觉院士提出的仿生模式识别(拓扑模式识别)^[17-18], 是基于“认识”事物而不是基于“区分”事物. 与传统以“最佳划分”为目标的统计模式识别相比, 它更接近于人类“认识”事物的特性, 故称为“仿生模式识别”. 它的数学方法在于研究特征空间中样本集合的拓扑性质, 故亦称作“拓扑模式识别”. “拓扑模式识别”的理论基点在于它确认了特征空间中同类样本的连续性(不能分裂成两个彼此不邻接的部分)特性. 在仿生模式识别中引入了特征空间中同类样本的连续性规律后, 对一类事物的“认识”, 实质上就是对这类事物的全体在特征空间中形成的无穷点集合的“形状”的分析和“认识”. 仿生模式识别是以一类样本在特征空间的分布的最佳覆盖作为目标. 王守觉院士把一个神经元对应于多维空间中一个超平面或超曲面, 而神经元输出函数的基是输入空间的输入点偏离该超平面或超曲面的程度. 显然超平面或超曲面的几何概念对于帮助人们对神经网络行为的认识与分析是十分有效的. 用几何分析方法发展神经网络新模型、新算法的实例是把用于模式识别的神经网络中每个神经元看作是在多维空间中做的一个超平面或超曲面, 依此来划分输入空间的样本点. 而如果把先分离的某些样本点挖掉不再参加后面的划分, 这种思想方法可用于设计神经网络硬件^[19]. 他还研究了多维空间几何的基本分析方法与定理证明^[20].

徐宗本教授提出了一种基于视觉的分类方法 VCA(Visual Classification Algorithm)^[21]. 这种方法是针对以下情况的: 目前的很多分类方法主要是通过发现数据内在的结构来分类, 很少或没有注意到从模拟人的感觉和感知来进行分类. 该方法模拟人的视觉和感知原理, 其基本思想是把数据看作虚拟的图像, 通过挖掘与人的视觉和感知一致的数据知识来理解数据并给出有效的算法. 这种算法基于视觉中的尺度理论. 其几何意义表

明这是一种覆盖分类学习算法,其核心是选择合适的尺度使得同类样本区域融合在一起,这一点在思想上和下面将要详细阐述的基于分类超曲面的覆盖分类算法是一致的,所不同的是基于分类超曲面的覆盖分类算法是无交覆盖区域的融合且采用 Jordan 定理分类。

朱洪教授等给出了集合击中和弱集合覆盖的定义并且证明了解决这类问题是 NP 难的。他们设计了两个算法来解决这两个问题,证明了这两个算法的近似度,并且证明了这类问题的不可近似性^[22]。这项工作是关于集合覆盖的基础性工作。

何清等提出了基于超曲面的分类学习算法(Hyper Surface Classification, HSC)^[23]。这是一种覆盖分类算法,基本思想是从几何学和拓扑学角度出发的。该算法基于 Jordan 曲线定理,根据围绕数的奇偶进行分类判断,不需要考虑使用何种核函数,而直接地解决非线性分类问题。下面给出 Jordan 曲线定理和分类判别定理。

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 1(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

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

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。

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 , 如何判断它是在分类超曲面 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 的环绕数)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。

下面的算法给出了构造和使用分类超曲面的基本过程:

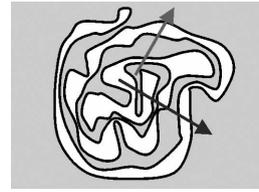


图 1 分类判别图示

假定 M 个训练样本: $K_{\text{train}} = \{x'(i) | i=1, 2, \dots, M\}$, 不失一般性, 假设样本点已预先标记好的类别 $t_i \in \{1, 2\}$; 训练空间 D 为一封闭区域, 且其边界满足 Jordan 曲线定理的基本条件。

1. 设前 m 个训练样本落在区域 D 内, $m \in \{1, 2, \dots, M\}$;
2. 将区域 D 划分成 n 个小区域, 设 $D = D_1 \cup D_2 \cup \dots \cup D_n$, $n \geq m$, 且满足以下条件:

(1) $D_j (j=1, 2, \dots, n)$ 的边界 H_j 满足 Jordan 曲线定理的基本条件;

(2) $x'(i) \in D_j, i=1, 2, \dots, m; j=1, 2, \dots, n$, 不妨设 $i=j$, 即每个小区域至多含一个样本点;

3. 设 $D_j (j=1, 2, \dots, m)$ 的边界 H_j 分别由 k_j 个超平面片组成, 可以将每片超平面表示为 $H_p^j (p=1, 2, \dots, k_j)$; 则可将 H_j 表示为含类别分量和超平面片的集合,

$$H_j = \{t_j, H_p^j | p=1, 2, \dots, k_j; t_j=1, 2\},$$

其中, t_j 为 D_j 所含样本点 $x'(j)$ 的类别;

4. 若 D_i 与 $D_j (i \neq j)$ 相邻, 且 $t_i = t_j$, 则合并边界 H_i 和 H_j (不妨设 H_p^i 与 H_q^j 重合, 即 $H_p^i = H_q^j$), 将重合超平面片消去; 则获得封闭区域 A_L , 其边界为

$$B_L = H_i \cup H_j = \{t_L, H_a^i, H_b^j | a=1, 2, \dots, k_i, \text{ 且 } a \neq p; b=1, 2, \dots, k_j, \text{ 且 } b \neq q; t_L = t_i\}.$$

继续合并相邻同类区域, 最终获得一组由若干超平面片组成的封闭超曲面——分类超曲面, 记为

$SHS_l^i = \{t_i, HP_l^i | i \in \{1, 2, \dots, r\}; l=1, 2, \dots, k_i; t_i=1, 2\}$, 其中 $r \leq m$; 以链的形式存储 SHS_l^i 的各超平面片 $HP_l^i (l=1, 2, \dots, k_i)$;

5. 输入新样本点 $X = x'(m+1)$, 按照以下方法进行判别(设 T 为样本点分类所得类别): 选择适当的由待定点 X 出发的射线 f_X , 设 f_X 与 SHS_l^i 的相交点数为 C_l^i , 分别计算 f_X 与 $\{SHS_a^1 | a=1, 2, \dots, r_1\}$ 和 $\{SHS_b^2 | b=1, 2, \dots, r_2\} (r_1 + r_2 = r)$ 的相交点数之和, 记为 $\sum_{a=1}^{r_1} C_a^1$ 和 $\sum_{b=1}^{r_2} C_b^2$, 则有

$$\sum_{a=1}^{r_1} C_a^1 \text{ 为奇数} \Leftrightarrow T=1; \text{ 或 } \sum_{b=1}^{r_2} C_b^2 \text{ 为奇数} \Leftrightarrow T=2.$$

实际上, 求 C_l^i 的过程就是求 f_X 与 HP_l^i 相交数量的过程;

6. 若不能判别 X 的类别, 就对 X 所在的小区域边界进行标定, 不妨设 $x'(m+1) \in D_{m+1}$, 则 D_{m+1} 的边界可表示为

$$H_{m+1} = \{t_{m+1}, H_p^{m+1} | p=1, 2, \dots, k_{m+1}\},$$

之后转入步 4, 继续合并相邻同类区域。

以上给出了基于分类超曲面的分类判别方法的基本算法, 即通过区域合并计算获得多个超平面组

成的双侧闭曲面,并作为分类超曲面对空间进行划分,也就是在样本点周围形成一个封闭区域,该区域由多个分类超平面片围成,并使得该区域覆盖某一类尽可能多的样本点,同时不覆盖异类样本点.

这种算法有两个关键步骤,一是局部化策略,另一个是用围绕数判断类别.该算法中,判别样本所属类别不需与所有分类边界链表做相交操作后再判断,而只需满足下面条件即可:由样本点所引射线与某完整分类边界链表相交点数为奇数.这样可提高判别速度.这种方法得到的分类超曲面是由若干个封闭闭曲面构成的,而曲面的局部是由低维平面片构成.每个闭曲面内部是一类样本,这样对闭曲面可以进行类别标记.样本类别可以是多个,所以这种方法对于多类问题的解决是很方便的,因为多个分类器可以在一次训练过程中产生,避免了两类分类器转化为多类分类器的技术处理过程.对二维和三维双螺旋及 UCI 中的数据分类实验结果说明:分类超曲面可以有效地解决在有限区域分布很复杂的海量(10^7)的非线性数据的多类分类问题,计算速度较高,同时对计算机资源要求很低,而传统的 SVM 不具备这种优点.另外小样本训练大样本测试结果表明:基于分类超曲面的分类法的泛化能力较好.该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,在现实中的数据分布大都满足这一条件^[25].

HSC 算法发展至今,已相继解决了二维两类分类^[23]、二维多类分类^[24]、二维一般连通区域分类^[25]、三维多类^[26]、高维换维分类问题^[27]、高维多类集成分类问题^[28].

3 基于覆盖数的机器学习 理论基础研究

著名的数学家 Smale 近年来也加入了机器学习理论基础的研究.他将逼近论用于建立机器学习的理论基础^[29-30].他的研究采用主流数学的方法和工具,所得到的结论既在理论上有重大学术价值,又对应用有具体的指导意义.学习理论研究从随机样本中学习对象,主要问题是样本容量,即对给定的置信度,为保证误差在给定的范围,需要学习多少样本^[29].覆盖数和重生核 Hilbert 空间的球填充数是这种样本容量的重要度量.对于度量空间中的紧集 S 和正数 $\eta > 0$,所谓覆盖数是指用半径为 $\eta > 0$ 的圆

盘覆盖紧集 S ,所用的圆盘个数 $m \in N$ 的最小值,记作 $N(S, \eta)$.重生核 Hilbert 空间的容量在学习理论分析中起着本质作用. Smale 对样本误差和逼近误差进行估计,在一些函数空间找到了样本个数与样本误差之间的关系,并指出了样本误差与覆盖数的关系,他还估计了一些具体空间的覆盖数. Zhou 给出了填充数的一个下界估计以及 Sobolev 光滑核的覆盖数的上界^[31].定义在紧集上的连续函数空间中的集合的覆盖数在学习理论中起着很重要的作用. Massimiliano 研究了覆盖数与其离散形式之间的关系,证明了当集合是光滑函数集合时,离散覆盖数逼近其连续覆盖数.他还指出了在重生核 Hilbert 空间中集合是一个球时的结果^[32]. Guo 等研究了 SVM 的覆盖数^[33]. Mendelson 对最近机器学习理论特别是样本复杂性问题的进展作了综述^[34].

4 基于极小样本集的机器学习 理论基础研究

所谓极小样本集是指能够反映训练得到的最终模型的全部性质的样本集,且其包含的样本数目应尽可能少.我们可以为 HSC 算法构造其极小样本集.

令 C 为有限样本集 S 的所有子集所构成的集合.定义 $C' \subseteq C$ 为 S 的一个覆盖,如果 S 中的每一个样本都属于 C' 中某个元素.定义相对于覆盖 C' 的极小样本集 (Minimum Sample Set) 为从 C' 每个元素中抽取一个且仅抽取一个样本构成的样本集,表示为 $S_{\min}|_{C'}$.

注意到可以利用 HSC 算法来构造覆盖集 C' ,即由 HSC 得到的分类超曲面 H 中的每个单位方体所包含的样本子集构成的集合,可以构成一个覆盖.我们定义基于超曲面 H 的极小样本集从每一个单位方体所包含的样本子集中仅选取一个代表样本所构成的集合,记为 $S_{\min}|_H$,且有如下的形式:

$$S_{\min}|_H = \bigcup_{u \subseteq H} \{\text{choosing one and only one } s \in u\}.$$

通常对于同一尺度的 HSC 模型来说极小样本集的组成样本是不唯一的,但是极小样本集中样本个数是唯一的.对于不同的分类算法极小样本集的构造方法可能不同,比如在 SVM 中极小样本集应该是支持向量,而 HSC 的极小样本集则是在每一个单位方体中取一个代表样本.

对近邻法也有相似的工作.为了减少近邻法的计算和存储量,很多研究针对如何减少近邻法中训

练集的样本数,这可以通过从原训练集中挑选一些有代表性的参考样本或者利用原样本集产生数目较少的新样本来实现,也相当于寻找极小样本集. 试图减少近邻法训练集样本数的研究开始于 Hart 的压缩近邻法(Condensed Nearest Neighbor rule, CNN)^[35]. 该算法得到的压缩集与原样本集是一致的,所谓一致是指原样本集中的样本用压缩集进行近邻法分类时可保证完全分类正确. CNN 算法的缺点是对原样本集样本的排列顺序敏感,而且压缩集中含有较多的冗余样本. 针对 CNN 算法中样本只能添入压缩集而不能删除的缺陷, Gates 提出了精简近邻法(Reduced Nearest Neighbor rule, RNN)^[37]. 他是将得到的压缩集进行精简,得到 CNN 压缩集的一个子集,而与原样本集保持一致性. 与 CNN 和 RNN 方法不同, Chang 试图从原样本集中产生一些新的参考样本,而不是从原来的样本中进行挑选^[38]. Devijver 和 Kittler 提出的多重剪辑算法 MULTIEDIT 的错误率是渐进贝叶斯最优的^[39]. 但 MULTIEDIT 算法一般只适用于训练样本数相当多的情形. 另外,它剪辑掉的多是类边界附近的点,仍有大量冗余的样本. Dasarathy 在总结以往工作的基础上,提出了极小一致子集(Minimal Consistent Subset, MCS)的概念和算法^[40]. 该算法是以最近异类子集(Nearest Unlike Neighbor Subset, NUNS)和最近异类样本(Nearest Unlike Neighbor, NUN)的概念为基础的. 虽然寻找的方法不同,在本质上各种不同的分类训练算法应该都存在一个极小样本集,它控制着最终模型的形式及性能.

5 基于超曲面的覆盖分类学习算法与理论的研究方向

基于超曲面的分类学习算法 HSC 作为一种新的算法,有很多问题亟待研究,这既包括算法优化,又包括理论分析,还包括应用中遇到的现实问题.

(1) 优化高维数据分类算法问题. 从理论上讲,这种方法可以推广到高维,因为 Jordan 定理在任意有限维空间都是成立的. 但是高维空间中的实现存在以下挑战性问题:一是高维空间的单位方体的合并计算复杂度随着维数增加而提高,另一方面高维超曲面的存储开销大. 但是这并不意味着 HSC 不能处理高维数据,借助数据预处理和集成学习技术,对于高维数据处理我们提出并实现了两种解决办法. 这两种高维处理方法与 HSC 算法特点紧密结

合,即形成基于样本数据重排的换维分类学习算法和基于集成学习思想^[42]的分维分类学习算法. 关于集成学习,周志华教授做了出色的工作^[43-45]. HSC 的集成特点是通过分维获得子分类器,不是通过划分样本集获得子分类器^[28]. 基于样本数据重排的换维分类学习^[27],将涉及到维排序和维组合问题,这些策略具有多样性,它们如何影响分类器性能,如何能找到最优的策略是有待研究的问题. 由于分类器集成方法是基于分维的,那么维的排序策略、划分策略、及权重策略就值得研究.

(2) 初始的划分尺度与泛化能力的关系问题. 基于超曲面的分类方法与传统方法相比,如 Parzen 窗分类器^[5],由于 HSC 采用了局部化策略,则克服了 Parzen 窗在样本分布不均衡情况下,由窗宽度较小所导致的分类区域过分零散,分类曲面复杂,推广性差,以及窗宽度较大时,分类区域融合过度所造成的分类误差大等问题. 在 HSC 中,由于采用局部化策略是对存在异类数据分布的同一单元区域进行,故这种方法是基于对数据分布的感知来工作的. Parzen 窗分类器的窗宽度是可以通过实验逐步择优的,但是一旦选定某个值,在整个分类过程中就不再变化,这对于分布不均衡的样本分类有明显的缺陷. 但是,初始的划分尺度对 HSC 的分类精度是有影响的,研究有关这种影响的估计以及提高精度的策略是重要的. 划分尺度与极小样本集之间的关系在一定意义上来说同仿生模式识别类似: HSC 将模式识别问题看成是模式的识别,而不是分类划分,不是模式分类. 因而,其数学模型与传统模式识别的“最优分类”界面的概念大不相同. 划分尺度越小, HSC 所得到的模型和训练样本的拓扑结构的匹配程度就越高,误识率就越低,但这也导致拒识率的提高,泛化能力的下降. 但是无论采用多大的划分尺度最终都会得到一个一致的分类模型,区别在于其细化程度的不同. 还要研究不同划分尺度产生的极小样本集间有何不同,并比较它们之间的性质.

(3) 最优的 HSC 覆盖问题. 张铃、张钹教授给出了 M-P 神经元的几何意义,通过球面投影变换将神经网络的最优设计问题转化为某种最优覆盖问题. 他们把神经元与几何上样本的球形邻域对应起来. 按照这种观点, HSC 可以看成神经元是由分类超曲面构成的神经网络,分类超曲面的个数就是神经元的个数. 所不同的是 HSC 的分类是靠样本围绕数来计算得出的,而神经网络是通过修正权值后加权计算获得. 如何找到超曲面个数少、分类性能好

的 HSC 分类器是一个重要的问题,这也是一个最优覆盖问题.

(4) HSC 的抽取规则问题. 以往人们觉得神经网络学习算法的分类过程不可理解,难以解释. 这样 Gallant 提出了一个简单的算法来解释连接主义专家系统所做的推理^[46]. 该算法通过产生规则来解释神经网络如何为某个给定案例得出结论. 其基本思想就是从当前已知的信息集中选择一个能有效地产生该结论的最小信息集合,也就是说,不管其他未知输入分量的取值为多少,只要满足该最小信息集合的取值要求就可以得出结论. Gallant 的这篇论文开创了神经网络规则抽取这一领域,成为该领域被引用最多的文献之一. 在 Gallant 之后,陆续有一些研究者对神经网络规则抽取进行了研究^[47-55]. 1995 年,Andrews 等人^[55]为从神经网络抽取的规则提出了一个评价体系,并提出了规则抽取算法的分类体系. 前者为不同规则抽取算法的比较提供了标准,也对新算法的设计具有指导作用,后者使得对规则抽取算法的系统化分析成为可能. 这两个体系为神经网络规则抽取这一领域的进一步发展奠定了基础,因此,Andrews 等人的这篇论文^[55]被认为是该领域的一个里程碑. HSC 分类超曲面以链表方式表达,我们已经实现了分类超曲面的可视化,但可视化并不意味着数据可理解,这些链表包含了分类信息,这些信息能否像神经网络规则抽取那样,从中抽出分类规则是值得研究的问题. 从分类超曲面中得到的分类规则就是可以学习,可以理解,可以传播的知识了.

(5) HSC 在数据理解中的作用问题. 理解数据,即获得数据集合的不同简洁程度表示,已成为机器学习研究的另一个重要研究方向. 解决这个问题的途径不能沿袭传统检验有效性的方法,即以检验个别事例为基础,而需要寻找必要的数学理论. 数据理解包括人对数据的理解和机器对数据的理解. 人对数据的理解,可以理解为借用符号机器学习的约简与可解释的特性,将一本使用数据语言书写的书翻译为人可理解的表示形式,从而丰富人的知识. 这就是数据挖掘的主要任务之一. 计算机的数据理解就是传统意义下的机器学习,分类超曲面可以看作数据分布的一个包络,这是数据理解的一个方面. 另一个方面的理解是数据分布的主曲线,主曲线相当于数据分布的骨架,这两者结合将会得到对数据更全面的理解.

(6) HSC 学习算法的计算复杂性. 其包括时间复杂性、空间复杂性、样本复杂性以及模型对新问题

的求解能力,或称为泛化能力. 从数学上来看,学习理论就是通过计算有限的随机样本获得数据中包含的知识^[56]. 函数的海量数据与学习的精确度(泛化能力)以及数据性质的多样性(领域依赖)等要求,需要人们考虑使用更多、更复杂的数学理论,如函数逼近论和宽度理论来揭示已有或正在发展的理论与方法所存在的问题及其对问题的适应性.

(7) 极小样本集的性质以及与 PAC 样本复杂度的关系. 从前面的论述我们可以看到,在 HSC 算法中我们可以构造出一个极小样本集,利用这个极小样本集进行训练会给出与用全部样本进行训练相同的模型. 基于覆盖类模式识别算法建立一个一般的极小样本集的概念使之不仅可以代表整个模型,而且可以反映出整个训练集的拓扑结构,并研究各种不同类型的算法使用不同的训练样本集时对最终分类器性能的影响. 此外 PAC 的样本复杂度理论给出了有多少随机抽取训练样例才足以可能近似正确(PAC)地学习到任意目标的概念. 通过 HSC 得到极小样本集的方法给出了选择的训练样本为了使在其上学习得到的模型有很好的性能应满足的性质(包含某一极小样本集). 我们将基于极小样本集这个概念,在一定的限制条件下,得到一个新的样本集所包含实例个数的边界,并与 PAC 的样本复杂度理论进行比较.

6 小 结

总之,在基于覆盖的分类学习算法方面,本文例举了最近十年提出的一些很有价值的分类学习方法和理论分析结果,但是,大量的理论问题尚未解决. HSC 学习算法作为一种覆盖分类算法,其性能提高以及计算复杂性、泛化能力等理论问题是值得深入研究的.

参 考 文 献

- [1] Vapnik V et al. The Nature of Statistical Learning Theory. New York; Springer-Verlag, 1995
- [2] Valiant L G. A theory of the learnable. Communications of the ACM, 1984, 27(11): 1134-1142
- [3] Blumer A, Ehrenfeucht A, Hassler D, Warmuth M K. Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension//Proceedings of the 19th Annual ACM Symposium on Theory of Computing. Berkeley, California, 1986; 273-282

- [4] Computational learning theory: Survey and selected bibliography. <http://delivery.acm.org/10.1145/130000/129746/p351-angluin.pdf>
- [5] Duda R O, Hart P E. Pattern Classification and Scene Analysis. New York: John Wiley & Sons, 1973
- [6] Duda R O, Hart P E, Stock D G. Pattern Classification. Beijing: China Machine Press, 2003(in Chinese)
(Duda R O, Hart P E, Stock D G. 模式分类. 北京:机械工业出版社, 2003)
- [7] Mendelson S. A few notes on statistical learning theory// Mendelson S, Smola A J eds. Advanced Lectures on Machine Learning, LNAI 2600. New York: Springer-Verlag, 2003; 1-40
- [8] Zhang Wen-Sheng, Ding Hui, Wang Jue. Study on computing the support vectors of massive data based on neighborhood principle. Journal of Software, 2001, 12(5): 711-720 (in Chinese)
(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): 711-720)
- [9] Tao Q, Wang J. Kernel projection algorithm for large-scale SVM problems. Journal of Computer Science Technology, 2002, 17(5): 556-564
- [10] Tao Qing, Wu Gao-Wei, Wang Jue. A generalized S-K algorithm for learning ν -SVM classifiers. Pattern Recognition Letters, 2004, 25(10): 1165-1171
- [11] Xu Jian-Hua, Zhang Xue-Gong, Li Yan-Da. A nonlinear perceptron algorithm based on kernel functions. Chinese Journal of Computers, 2002, 25(7): 689-695(in Chinese)
(许建华, 张学工, 李衍达. 一种基于核函数的非线性感知器算法. 计算机学报, 2002, 25(7): 689-695)
- [12] Zhang Ling, Zhang Bo. A geometrical representation of McCulloch-Pitts neural model and its applications. IEEE Transactions on Neural Networks, 1999, 10(4): 925-929
- [13] Wu Tao, Zhang Ling, Zhang Yan-Ping. Kernel covering algorithm for machine learning. Chinese Journal of Computers, 2005, 28(8): 1295-1301(in Chinese)
(吴涛, 张铃, 张燕平. 机器学习中的核覆盖算法. 计算机学报, 2005, 28(8): 1295-1301)
- [14] Tao Pin, Zhang Bo, Ye Zhen. An incremental bicovering learning algorithm for constructive neural network. Journal of Software, 2003, 14(2): 194-201(in Chinese)
(陶品, 张钺, 叶榛. 构造型神经网络双交叉覆盖增量学习算法. 软件学报, 2003, 14(2): 194-201)
- [15] Zhang Ling, Zhang Bo, Yin Hai-Feng. An alternative covering design algorithm of multi-layer neural networks. Journal of Software, 1999, 10(7): 737-742(in Chinese)
(张铃, 张钺, 殷海风. 多层前向网络的交叉覆盖设计算法. 软件学报, 1999, 10(7): 737-742)
- [16] Zhang Ling, Zhang Bo. Learning methods of neural networks//Zhou Zhi-Hua, Cao Cun-Gen eds. Proceedings of the Neural Networks and Its Applications. Beijing, 2004(in Chinese)
(张铃, 张钺. 神经网络的学习方法. 周志华, 曹存根主编. 神经网络及其应用. 北京, 2004)
- [17] Wang Shou-Jue. Bionic(topological) pattern recognition——A new model of pattern recognition theory and its applications. Acta Electronica Sinica, 2002, 30(10): 1417-1420(in Chinese)
(王守觉. 仿生模式识别(拓扑模式识别)——一种模式识别新模型的理论与应用. 电子学报, 2002, 30(10): 1417-1420)
- [18] Wang S J, Qu Y F, Li W J, Qin H. Face recognition: Biomimetic pattern recognition vs. traditional recognition. Acta Electronica Sinica, 2004, 32(7): 1057-1061
- [19] Cao W M, Hao F, Wang S J. The application of DBF neural networks for object recognition. Information Sciences — Informatics and Computer Science: An International Journal, 2004, 160(1-4): 153-160
- [20] Wang Shou-Jue, Wang Bai-Nan. Analysis and theory of high-dimensional space geometry for artificial neural networks. Acta Electronica Sinica, 2002, 30(1): 1-4(in Chinese)
(王守觉, 王柏南. 人工神经网络的多维空间几何分析及其理论. 电子学报, 2002, 30(1): 1-4)
- [21] Xu Zong-Ben, Meng De-Yu, Jing Wen-Feng. A new approach for classification: Visual simulation point of view// Proceedings of the ISNN 2005. LNCS 3497. Springer-Verlag, 2005; 1-7
- [22] Zhang Yong, Zhu Hong. Approximation algorithms for the problems of weak set cover. Chinese Journal of Computers, 2005, 28(9): 1497-1500(in Chinese)
(张涌, 朱洪. 一类弱集合覆盖问题的近似算法. 计算机学报, 2005, 28(9): 1497-1500)
- [23] He Qing, Shi Zhong-Zhi, Ren Lian. The classification method based on hyper surface//Proceedings of the IEEE International Joint Conference on Neural Networks. 2002; 1499-1503
- [24] He Qing, Shi Zhong-Zhi, Ren Lian. The multi-class classification method in large database based on hyper surface//Proceedings of the International Conference on Machine Learning and Application. Las Vegas; CSREA, 2002; 164-169
- [25] He Qing, Ren Li-An, Shi Zhong-Zhi. The large data direct classifying method based on hyper surface. Chinese Journal of Computers, 2003, 26(2): 206-211(in Chinese)
(何清, 任力安, 史忠植. 基于超曲面的海量数据直接分类法. 计算机学报, 2003, 26(2): 206-211)
- [26] He Qing, Shi Zhong-Zhi, Ren Li-An, Lee E S. A novel classification method based on HyperSurface. International Journal of Mathematical and Computer Modeling, 2003, 38: 395-407
- [27] He Qing, Zhao Xiu-Rong, Shi Zhong-Zhi. Classification based on dimension transposition for high dimension data. Soft Computing, 2007, 11(4): 329-334
- [28] Zhao Xiu-Rong, He Qing, Shi Zhong-Zhi. HyperSurface classifier ensemble for high dimensional data sets//Proceedings of the 3rd International Symposium on Neural Networks (ISNN 2006). LNCS 3971. Springer-Verlag, 2006; 1299-1304
- [29] Cucker F, Smale S. On the mathematical foundations of learning. Bulletin of the American Mathematical Society, 2002, 39(1): 1-49
- [30] Smale S, Zhou D X. Estimating the approximation error in learning theory. Anal. Appl., 2003, 1: 1-25

- [31] Zhou Ding-Xuan. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 2003, 49(7): 1743-1752
- [32] Massimiliano Pontila. A note on different covering numbers in learning theory. *Journal of Complexity*, 2003, 19(5): 665-671
- [33] Guo Ying, Bartlett Peter L, John Shawe-Taylor, Williamson R C. Covering numbers for support vector machine. *IEEE Transactions on Information Theory*, 2002, 48(1): 239-250
- [34] Mendelson S. Geometric parameters in learning theory//*Proceedings of the LNM 1850*, 2004: 193-235
- [35] Hart P E. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 1968, IT214(3): 515-516
- [36] Blumer A, Haussler D, Kearns M, Valiant L. A general lower bound on the number of examples needed for learning. *Information and Computation*, 1989, 82: 247-261
- [37] Gates G W. The reduced nearest neighbor rule. *IEEE Transactions on Information Theory*, 1972, IT218(3): 431-433
- [38] Chang C L. Finding prototypes for nearest neighbor classifiers. *IEEE Transactions on Computers*, 1974, C223(11): 1179-1184
- [39] Devijver P A, Kittler J. On the edited nearest neighbor rule//*Proceedings of the 5th ICPR*. Miami, Florida, 1980: 72-80
- [40] Dasarathy B V. Minimal consistent set (MCS) identification for optimal nearest neighbor decision systems design. *IEEE Transactions on Systems, Man and Cybernetics*, 1994, 24(3): 511-517
- [41] Simon H U. General lower bounds on the number of examples needed for learning probabilistic concepts. *Journal of Computer and System Sciences*, 1996, 52(2): 239-254
- [42] Hansen L K, Salamon P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990, 12(10): 993-1001
- [43] Zhou Zhi-Hua, Wu Jian-Xin, Jiang Yuan, Chen Shi-Fu. Genetic algorithm based selective neural network ensemble//*Proceedings of the 17th International Joint Conference on Artificial Intelligence*. Seattle, WA, 2001, 2: 797-802
- [44] Zhou Z-H, Wu J, Tang W. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 2002, 137(1/2): 239-263
- [45] Zhou Zhi-Hua, Chen Shi-Fu. Neural network ensemble. *Chinese Journal of Computers*, 2002, 25(1): 1-8 (in Chinese)
- (周志华,陈世福. 神经网络集成. *计算机学报*, 2002, 25(1): 1-8)
- [46] Gallant S I. Connectionist expert systems. *Communications of the ACM*, 1988, 31(2): 152-169
- [47] Setiono R, Liu H. Understanding neural networks via rule extraction//*Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Montreal, Canada, 1995: 480-485
- [48] Wu X. *Knowledge Acquisition from Databases*. Norwood NJ: Ablex, 1995
- [49] Alexander J A, Mozer M C. Template-based procedures for neural network interpretation. *Neural Networks*, 1999, 12(3): 479-498.
- [50] Zhou Zhi-Hua. Rule extraction from neural networks//Zhou Zhi-Hua, Cao Cun-Gen eds. *Proceedings of the Neural Networks and Its Applications*. Beijing, 2004 (in Chinese) (周志华. 神经网络规则抽取. 周志华, 曹存根主编. 神经网络及其应用. 北京, 2004)
- [51] Zhou Z-H. Rule extraction: Using neural networks or for neural networks? *Journal of Computer Science and Technology*, 2004, 19(2): 24-253
- [52] Zhou Z-H, Jiang Y, Chen S-F. A general neural framework for classification rule mining. *International Journal of Computers, Systems and Signals*, 2000, 1(2): 154-168
- [53] Zhou Zhi-Hua, Chen Shi-Fu. Rule extraction from neural networks. *Journal of Computer Research and Development*, 2002, 39(4): 398-405 (in Chinese) (周志华, 陈世福. 神经网络规则抽取. *计算机研究与发展*, 2002, 39(4): 398-405)
- [54] Zhou Z-H, Jiang Y, Chen S-F. Extracting symbolic rules from trained neural networks. *AI Communications*, 2003, 16(1): 3-15
- [55] Andrews R, Diederich J, Tickle A B. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 1995, 8(6): 373-389
- [56] John Shawe-Taylor, Williamson Robert C P. Generalization performance of classifiers in terms of observed covering numbers fischer//Simon H U ed. *Proceedings of the EuroCOLT'99*. LNAI 1572. Springer-Verlag, 1999: 274-285
- [57] Wang Jue, Shi Chun-Yi. Investigations on machine learning. *Journal of Guangxi Normal University (Natural Science Edition)*, 2003, 21(2): 1-15 (in Chinese) (王珏, 石纯一. 机器学习研究. *广西师范大学学报(自然科学版)*, 2003, 21(2): 1-15)



HE Qing, born in 1965, professor, Ph.D. supervisor. His research interests include fuzzy mathematics, machine learning and artificial intelligence.

SHI Zhong-Zhi, born in 1941, professor, Ph.D. supervisor. His research interests include artificial intelligence, machine learning, neural computing and cognitive science.

Background

This paper studies classification problem that belongs to the machine learning category. The advances in the area of classification based on covering algorithm are summarized. Specially, the Hyper Surface Classification is introduced and analyzed in detail. Moreover, the future research directions are pointed out. The understanding of data is highly relevant to how one senses and perceives them. The covering learning algorithms can always simulate human visual cognition to represent the data distribution. Some covering learning algorithms were proposed in the decade. Zhang Ling and Zhang Bo proposed a geometry classification method, where the original input space is transferred into a quadratic space by the use of a global project function. Then, the well-known point set covering method was used to perform the partition of the data in the transformed space. Biomimetic Pattern Recognition (BPR) theory is firstly proposed by Wang Shou-Jue as a new model for pattern recognition. Xu Zong-Ben proposed an

approach called Visual Classification Algorithm (VCA) for classification, with an expectation of resolving some of the problems mentioned above. Lee Daewon and Lee Jaewook proposed a learning algorithm for semi-supervised classification. For Hyper Surface Classification (HSC), Hyper Surface Classification (HSC) is a novel classification method based on hyper surface is put forward by He & Shi & Ren (2002). However, what we really need is an algorithm that can deal with data not only of massive size but also of high dimensionality. Thus He, Zhao & Shi proposed a simple and effective kind of dimension reduction method without losing any essential information in 2006. Another solution to the problem of HSC on high dimensional data sets is proposed. A judgment sampling method based on Minimal Consistent Subset (MCS) is proposed to select of a representative subset of the original training data.