文章编号: 1001-0920(2004)05-0573-04

粗 SVM 分类方法及其在污水处理过程中的应用

范昕炜, 杜树新, 吴铁军
(浙江大学 工业控制技术国家重点实验室 智能系统与决策研究所, 浙江 杭州 310027)

摘 要: 提出一种基于粗糙集理论和支持向量机理论的粗 SVM 分类方法. 该方法采用粗糙集属性约简方法以减少属性个数, 且在属性约简过程中选出几组合适的属性集组成新的属性集, 使模型具有一定的抗信息丢失能力, 同时充分利用 SVM 的良好推广性能, 提高了预测分类精度. 对城市污水处理厂运行状态的实验结果表明了该方法的优越性.

关键词: 支持向量机; 粗糙集; 分类精度; 污水处理过程
中图分类号: TP391.4 **文献标识码**: A

Rough support vector machine and its application to wastewater treatment processes

FAN Xin-wei, DU Shu-xin, WU Tie-jun
(National Key Laboratory of Industrial Control Technology, Institute of Intelligent Systems and Decision Making, Zhejiang University, Hangzhou 310027, China. Correspondent: DU Shu-xin, E-mail: shxdu@iipc.zju.edu.cn)

Abstract: A new classification algorithm named rough support vector machine (RSVM) is presented based on support vector machine (SVM) and rough set theory. RSVM has high predictive classification accuracy with much less attributes, which means less sensors and less cost. And it keeps certain redundant attributes to have high predictive accuracy in the case of lost information caused by sensor fault. RSVM increases classification accuracy with good generalization performance. The numerical experiments for a wastewater treatment process show the effectiveness of the proposed algorithm.

Key words: support vector machine; rough set; classification accuracy; wastewater treatment process

1 引 言

实际生产过程中, 需要大量传感器监测过程的运行状态, 以确保生产正常运行. 运行状态的监测/监控本质上是一个模式分类问题, 因此需要相应的模式分类方法以实现生产过程的运行状态监测. 传感器的数量直接影响系统的成本, 如何减少一些测量不重要变量(属性)的传感器, 并保证具有较高的预测分类精度, 这是一个值得研究的实际课题. 粗糙集是减少属性数目的有效方法. 20 世纪 90 年代中期提出的支持向量机(SVM)[1~4]是基于结构风险最小化准则, 使期望泛化误差的上界最小, 且具有好的推广能力的一种方法. 本文提出一种基于粗糙集理论和支持向量机理论的分类方法, 称之为粗 SVM 分类方法(RSVM). 利用粗糙集理论[5]中属性约简方法, 降低样本维数, 并保留一定的冗余属性, 这样可以减少测量属性的传感器数目, 降低生产成本, 同时提高训练模型的抗信息丢失能力. 另外, 利用 SVM 良好的推广性能, 在小样本训练的情况下得到较高的预测分类精度.

收稿日期: 2003-01-09; **修回日期**: 2003-06-27.
作者简介: 范昕炜(1973—), 男, 江西广丰人, 博士, 从事数据挖掘、模式分类等研究; 吴铁军(1950—), 男, 江苏南京人, 教授, 博士生导师, 从事智能系统控制与决策等研究.

2 粗糙集和 SVM 理论的回顾

2.1 粗糙集理论的属性约简概念[5]

设信息系统为 $IS = (U, A)$. 其中: U 是全域(对象的有限集, $U = \{x_1, x_2, \cdots, x_m\}$), A 是属性集(特征, 变量). 每个属性 $a \in A$. 定义信息函数为 $f_a: U \to V_a$, 其中 V_a 是 a 值的集, 称为属性 a 的域. 检验属性集是否独立可以看属性一个个被去掉后, 是否会增加信息系统中基本集的数目. 如果 $\mathrm{Ind}(A) = \mathrm{Ind}(A - a_i)$, 那么属性 a_i 称为冗余的; 否则, 属性 a_i 对 A 来说是必不可少的. 若属性集不是独立的, 则能找到所有可能的最小属性子集, 这样就得到了相同数目的整个属性集的基本集(约简), 并找到所有不可缺少的属性集(核). 约简的方法有很多, 如基因算法和 Johnson 算法[6].

2.2 v-SVM 算法[1,7]

支持向量机方法是在线性样本空间中构造基于结构风险最小化准则的最优分类超平面. 对于非线性样本数据, 则通过非线性函数将输入空间映射到高维(可能是无限维)线性特征空间, 从而在高维空间中构造最优分类器. 给出训练向量 $x_i \in R^n$, $i = 1, 2, \cdots, l$, 属于两类, 即 $y_i \in \{1, -1\}$. v-SVM 算法最优化问题的对偶最优化问题为[7]

$$\min_{\alpha} \frac{1}{2}\alpha^{\mathrm{T}} Q \alpha, \tag{1}$$

$$\text{s.t.}\quad e^{\mathrm{T}}\alpha \geq v,\ y^{\mathrm{T}}\alpha = 0,$$
$$0 \leq \alpha_i \leq 1/l;\ i = 1, 2, \cdots, l. \tag{2}$$

其中: e 是单位向量, Q 是 $l \times l$ 正半定矩阵, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, $K(x_i, x_j) \equiv \varphi(x_i)^{\mathrm{T}}\varphi(x_j)$ 是核函数. 决策函数 $f^*(x)$ 和 b 分别为

$$f^*(x) = \mathrm{sign}\left(\sum_{i=1}^{l} y_i \alpha_i (K(x_i, x) + b)\right), \tag{3}$$

$$b = -\frac{1}{2s}\sum_{x \in S_+ \cup S_-}\sum_{j} \alpha_j y_j K(x, x_j). \tag{4}$$

其中 $S_{\pm}$ 为大小相同($s > 0$)的两个集, 包含支持向量 x_i, 且分别满足 $0 \leq \alpha_i \leq 1/l$ 和 $y_i = \pm 1$.

虽然上述 v-SVM 分类器是二类别分类器, 但很容易将其组合起来处理多类别的情况[8,9].

3 粗 SVM 分类方法(RSVM)

3.1 基本分类过程

在许多实际模式分类应用(如生产过程的运行状态监控)中, 通常会遇到下述 3 个问题:

1) 样本的属性太多会导致相应的传感器数目增加, 从而使生产成本上升;

2) 要求训练模型具有一定的抗信息丢失能力, 不容易受传感器故障等影响;

3) 由于技术或经济条件的限制, 无法很快获得建模对象的大量数据, 只能在小数据集的基础上进行模式分类, 并要求有较高的预测分类精度.

对于具有以上特点的模式分类问题, 本文提出了粗 SVM 分类方法. 通过把粗糙集约简方法引入 SVM 分类器, 并保留一定的冗余属性, 提高了分类精度. 分类方法的基本步骤如下:

1) 样本预处理: 去掉不完全数据, 使剩下的样本都是全部属性具有属性值的样本;

2) 属性约简: 利用粗糙集理论的属性约简概念和方法对样本集进行处理, 得到一系列的约简属性集;

3) 产生符合要求的新属性集: 结合工艺和生产实际等方面的专家经验, 挑选出一组符合要求的由几组约简属性集组成的新属性集(下节将作详细讨论);

4) 建立训练模型: 对新属性集的样本进行处理, 用"一对一"方法[8]解决多类别问题, 根据式(1)和式(2)解二次优化问题, 得到 $k(k-1)/2$ 个决策函数 $f^*(x)$, 最后得到训练模型;

5) 分类预测: 用得到的训练模型对测试数据进行分类, 并输出结果和分类精度.

3.2 新属性集的产生方法

该分类方法把粗糙集理论中约简的方法应用于 SVM 分类, 但不是简单地把求出的一组最优约简属性集送给 SVM 处理, 而是结合专家经验来保留几组最优或较优的约简属性集, 再将其合并成一组新的属性集送给 SVM 处理, 即有一定的冗余量. 这样做的主要优点在于: 减少一定属性的同时提高了训练模型的可靠性, 使其具有一定的抗信息缺失能力. 如果只是简单地将粗糙集和 SVM 结合起来, 就会在某一测量最优属性集的传感器出现故障或受到干扰情况下, 导致关键信息丢失或失真, 从而严重影响分类精度. 本文给出的粗 SVM 分类方法由于保留了几组约简属性集, 其中每组约简属性集的分类能力都与全体属性集的分类能力相同[5], 即在测量某一属性的传感器出现问题或其他原因数据缺失的情况下, 其余几组属性集仍保留有正确的分类信息, 使后续处理训练出较好的模型, 而不会导致分类效果大幅度变差.

产生新的符合要求的属性集的具体做法是:

1) 根据粗糙集约简方法得到的一系列约简属性集, 依次搜索出几组约简属性集, 要求每一组的属

性不能相同(核属性除外);

2) 将这几组属性集组合成一组新的属性集, 如果这个新属性集的属性个数超过了属性总数目的 50% 或搜索到最后一个, 就停止本次搜索返回上一步, 否则进入下一步;

3) 通过 1) 和 2) 的循环执行, 可找出所有符合要求的属性集, 若得到了几组新的属性集, 则需根据专家经验选出一组最符合实际要求的属性集

4　污水处理过程运行状态监控实验

污水处理过程是一个复杂的生化过程, 如何通过各处理阶段的可测变量进行处理过程运行状态的分类(正常情况/异常情况), 从而判断某一环节是否出现故障, 是污水处理过程运行与管理的关键 这里以某城市污水处理厂为例, 讨论粗 SVM 分类方法的具体应用 实验数据来自 UCI 机器学习数据库

整个数据集有 527 个样本, 包括 38 个属性和 13 种分类 为简便起见, 将其合并成 6 类, 如表 1 所示 其中: 1 类为正常情况, 2 类为性能超过平均值的正常情况, 3 类为低输入的正常情况, 4 类为二沉池故障, 5 类为暴雨, 6 类为固体溶度过负荷

表 1　6 类共 527 个样本的组成

类　别	1	2	3	4	5	6
样本个数	332	116	65	7	3	4

全部样本分成两部分, 60% 用于训练, 40% 用于测试 由基因算法得到的约简属性组如表 2 所示, 新属性集包括属性 1, 3, 4, 6, 8, 9, 12, 13, 14, 16, 20, 21, 22, 24, 25, 27, 30, 31, 34, 35, 37. 按粗糙集理论[5], 每一组属性集都和 38 个属性集的分类能力相同 将这 3 组属性组合成一个新的属性集, 便有了冗余, 即具有一定的抗信息丢失能力.

表 2　3 组约简属性集

	属　性						
第 1 组	4	8	12	16	20	22	30
第 2 组	3	6	21	25	27	35	37
第 3 组	1	9	13	14	24	31	34

为了进行比较, 采用 RSVM, 粗糙集分类方法(RSC) 和 v-SVM 这 3 种算法来训练分类器 其中支持向量机的核函数采用多项式核函数, 即

$$K(x_i, x_j) = (x_i x_j + 1)^d,$$

式中 $d = 3$. 选取参数 $v = 0.000\,001$, 结果如表 3 所示

表 3　3 种算法的分类精度

分类精度/%	RSVM	RSC	v-SVM
测试数据	78.21	73.72	75.66
训练数据	87.55	100	100
全体数据	83.80	89.46	90.26
属性个数	21	21	38
总的 SVs 数目	82	—	103

从表 3 可知, RSVM 算法的预测分类精度最高, 比 RSC 算法高出 4.49, 比 v-SVM 算法高出 2.55, 使用的属性为 v-SVM 算法的 55.26%, 支持向量数目为 v-SVM 算法的 79.61%.

下面研究 RSVM 分类方法关于抗信息缺失性能的问题, 核函数仍为多项式核函数, 参数 v 不变 20RSVM 表示 21 个属性的测量中有一个属性值由于意外原因丢失, 本实验假设属性 37 缺失 7RSVM 表示使用 7 个属性进行训练, 即采用没有信息冗余的属性集进行训练, 属性集包括属性 3, 6, 21, 25, 27, 35, 37. 6RSVM 表示 7 个属性的测量中有一个属性值由于意外原因丢失, 也假设属性 37 缺失 实验结果如表 4 所示

表 4　RSVM 分类方法抗信息缺失的效果

分类精度/%	RSVM	20RSVM	7RSVM	6RSVM
测试数据	75.00	75.00	64.10	52.56
训练数据	99.57	99.57	64.38	46.78
全体数据	89.71	89.71	64.26	49.10
属性个数	21	20	7	6
总的 SVs 数目	104	108	49	45

由表 4 可知, 20RSVM 和 RSVM 分类精度差不多, 而 7RSVM 和 6RSVM 分类精度则相差很多. 这表明 RSVM 具有一定的抗信息缺失能力, 同时说明保留冗余属性对抗信息丢失具有重要意义 6RSVM 的训练和测试数据的分类精度约为 50%, 说明由于信息缺乏, 该方法已经不起作用了, 得到的结果没有什么意义

5　结　论

本文提出的粗 SVM 分类方法通过利用粗糙集理论中属性约简的想法, 降低了样本属性并保留一定的冗余属性, 同时利用 SVM 良好的推广性能, 在小样本训练的情况下得到了较高的预测分类精度 该方法的主要优点在于:

1) 减少属性, 即减少传感器数目, 降低生产成本;

2) 保留一定的冗余属性, 使训练模型可靠, 具有一定的抗信息缺失能力;

3) 根据小样本得到的训练模型具有较高的预测精度

对污水处理过程运行状态监测的实验结果表明了该方法的有效性

参考文献(References):

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

[2] Kreβel U. Pairwise classification and support vector machines [A]. *Advances in Kernel Methods Support Vector Learning* [C], Cambridge: MIT Press, 1999. 255-268.

[3] Joachims T. Text categorization with support vector machines [R]. Dortmund: University of Dortmund, 1997.

[4] Cai Y D, Liu X J, Xu X B, et al. Prediction of protein structural classes by support vector machines [J]. *Computers and Chemistry*, 2002, 26(3). 293-296.

[5] Pawlak Z. *Rough Sets-theoretical Aspects of Reasoning about Data* [M]. Boston, London: Kluwer Academic Publishers, 1992. 1-53.

[6] Aleksander Ohrn. Discernibility and rough sets in medicine: Tools and applications [D]. Trondheim: Norwegian University of Science and Technology, 1999. 53, 63-65.

[7] Scholkopf B, Smola A, Williamson R C, et al. New support vector algorithms [J]. *Neural Computation*, 2000, 12(5): 1207-1245.

[8] Hsu C W, Lin C J. A comparison of methods for multiclass support vector machines[J]. *IEEE Trans on Neural Networks*, 2002, 13(2): 415-425.

[9] Chang C C, Lin C J. Training *v*-support vector classifiers: Theory and algorithms [J]. *Neural Computation*, 2001, 13(9): 2119-2147.

(上接第 572 页)

5 结 语

根据Lyapunov稳定性理论, 针对一类线性不确定性系统, 给出了正常跟踪控制器和可靠跟踪控制器存在的充分条件; 通过求解LMI给出了设计两种控制器的方法. 仿真数例表明, 文中所给的控制器设计方法是可行的. 通过对正常跟踪控制系统与可靠跟踪控制效果的比较, 进一步看出对系统进行可靠控制设计的必要性

参考文献(References):

[1] 张华春, 谭民. 状态反馈控制系统的容错控制器设计[J]. 控制与决策, 2000, 15(6): 724-726.
(Zhang H C, Tan M. Design of fault-tolerant controller to state feedback control systems [J]. *Control and Decision*, 2000, 15(6): 724-726.)

[2] Veillette R J, Medanic J V, Perkins W R. Design of reliable control system [J]. *IEEE Trans on Automatic Control*, 1992, 37(3): 770-784.

[3] Liu J, Wang J L. Reliable robust minimum variance filtering with sensor failures[A]. *Proc* 2001 *ACC*[C]. Arlington, 2001. 1041-1046.

[4] Yang G H, Wang J L, Soh Y C. Reliable H_∞ design for linear system [J]. *Automatica*, 2001, 37(5): 717-725.

[5] Zhao Q, Jiang J. Reliable tracking control system design against actuator failures [A]. *Proc* 1997 *SICE* [C]. Tokushima, 1997. 1019-1024.

[6] Liao F, Wang J L, Yang G H. MI-based reliable robust tracking control against actuator faults with application to flight control[A]. *Proc* 39*th IEEE Conf on Decision and Control*[C]. Sydney, 2000. 3914-3919.

[7] Liao F, Wang J L, Yang G H. LMI-based reliable robust preview tracking control against actuator faults [A]. *Proc* 2001 *ACC*[C]. Arlington, 2001. 1047-1052.

[8] Liao F, Wang J L, Yang G H. Reliable robust flight tracking control: An LMI approach[J]. *IEEE Trans on Control Systems Technology*, 2002, 10(1): 76-89.