

基于粗糙集粒子群支持向量机的特征选择方法

崔文岩¹, 孟相如¹, 李纪真¹, 王明鸣¹, 陈天平¹, 王 坤²

(1 空军工程大学 信息与导航学院, 陕西 西安 710077; 2 94872 部队司令部, 江西 樟树 331204)

摘 要: 将 Filter 型粗糙集属性约简方法与 PSO-SVM 方法相结合, 提出一种新的粗糙集粒子群支持向量机(RS-PSO-SVM)特征选择方法, 给出了该方法的特征选择具体步骤, 并对比分析了所提方法的性能。仿真实验表明: 提出的 RS-PSO-SVM 特征选择方法是有效的, 在保证所选特征集为最优情况下, 极大地缩短所用时间, 可以将其应用在多维数据的特征选择中。

关键词: 粗糙集; 属性约简; 粒子群支持向量机; 特征选择

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2015)01-0120-04

Feature Selection Based on RS-PSO-SVM

CUI Wen-yan¹, MENG Xiang-ru¹, LI Ji-zhen¹,
WANG Ming-ming¹, CHEN Tian-ping¹, WANG Kun²

(1 Institute of Information and Navigation, Air Force Engineering University, Xi'an 710077, China;

2 Command of Army 94872, Zhangshu 331204, China)

Abstract: A new RS-PSO-SVM feature selection combined Filter Rough Set attribute reduction with PSO-SVM was proposed. The specific steps of the new method were given, and the performance was analyzed contrasted with other two methods. Simulation experiments show that, the proposed feature selection method is effective which can reduce time consumed largely and ensure the selected feature set is the best. The proposed method can be used in the feature selection of multidimensional data.

Key words: rough set; attribute reduction; PSO-SVM; feature selection

1 引言

人们进行科研实验过程中往往需要处理高维数据, 这些特征并不是独立的, 彼此之间一般含有大量冗余, 这些冗余特征会增加数据分析的计算时间, 影响数据评估的精度。因此需要对数据进行特征选择。特征选择是指对数据从其所有特征之中挑选出一些最有效的特征以降低特征空间维数的过程^[1-2]。

依据是否依赖于后续学习算法, 可将特征选择方法分为过滤式(Filter)和封装式(Wrapper)两种。Filter 型特征选择算法独立于机器学习算法, 计算代价小, 效率高但效果一般, 选择出的特征集可能不

是规模最小的; Wrapper 型特征选择算法利用机器学习算法的训练准确率作为评估标准, 计算代价大, 效率低但选择效果好^[3-4]。

文献[5, 6]中以粒子群为特征集寻优方式, 以支持向量机为机器学习方法(PSO-SVM), 能取得良好的特征选择效果, 但是该特征选择方法会耗费大量的时间^[7]。基于粗糙集理论的特征选择称为属性约简, 经过属性约简后的特征集并不能完全去除特征之间的冗余性。本文将粗糙集与 PSO-SVM 理论相结合, 提出一种基于粗糙集的 PSO-SVM 特征选择方法(RS-PSO-SVM), 该方法兼顾 Filter 方法的高效率和 Wrapper 方法的高准确率。

收稿日期: 2014-03-09; 修回日期: 2014-05-03

基金项目: 国家自然科学基金项目(61201209); 陕西省自然科学基金项目(2013JQ8013)

2 相关理论

2.1 基于粗糙集理论的属性约简

信息系统可以利用其研究对象的集合来表示,这些对象的知识可通过指定对象的基本特征和它们的特征值来描述.形式上,信息系统可用四元组表示:

$$S = (U, A, V, f) \quad (1)$$

式中, U 为对象的非空有限集合; A 为属性的非空有限集合.

$$V = \bigcup_{a \in A} V_a \quad (2)$$

式中, V 为属性值域, V_a 是属性 a 的值域.

$$f: U \times A \rightarrow V \quad (3)$$

式中, f 是一个信息函数,它为每个属性赋予一个信息值,即 $\forall a \in A, x \in U, f(x, a) \in V_a$.

属性依赖度反映了属性间的一种映射关系,给定决策信息系统 $S = (U, C \cup D)$,那么条件属性集 C 对决策属性集 D 的依赖程度为

$$k = r(C, D) = \frac{|\bigcup_{X \in D^*} C_-(X)|}{|U|} \quad (4)$$

式中, $D^* = U/D$, $C_-(X)$ 为 X 的下近似.这里 $0 \leq k \leq 1$.因此,如果 $k = 0$, D 完全独立于 C ;如果 $0 < k < 1$, D 部分依赖于 C ;如果 $k = 1$, D 完全依赖于 C [8].

利用粗糙集的属性约简,从全部条件属性中发现部分必要的条件属性,以此来降低数据集的复杂性.

2.2 粒子群算法介绍

粒子群算法 (Particle Swarm Optimization, PSO) 是一种群体智能算法,模拟自然界鸟群觅食规律,通过不断迭代搜索获取最优解 [2,5]. 算法采用位置-速度搜索模型,假设粒子群由 n 个微粒组成,每个微粒有 m 维,粒子 i 的位置为 $X_i = (x_{i1}, x_{i2}, \dots, x_{im})$; 速度为 $V_i = (v_{i1}, v_{i2}, \dots, v_{im})$; 微粒个体极值为 $P_i = (p_{i1}, p_{i2}, \dots, p_{im})$; 全体极值为 $P_g = (p_{g1}, p_{g2}, \dots, p_{gm})$. 微粒通过自己的极值和群体极值共同决定下一步的位置和速度.在迭代过程中,微粒的每一步和上一步之间的迭代关系分别满足式(5)和式(6).

$$v_{id}^{k+1} = \omega \times v_{id}^k + c_1 \times r_1 \times (p_{id} - x_{id}^k) + c_2 \times r_2 \times (p_{gd} - x_{id}^k) \quad (5)$$

$$x_{id}^{k+1} = x_{id}^k + v_{id}^{k+1} \quad (6)$$

式中, $i = 1, 2, \dots, n$, n 是微粒个数; $d = 1, 2, \dots, m$,

m 是微粒的维数. c_1 和 c_2 为加速常数, r_1 和 r_2 为 $[0, 1]$ 内两个相互独立的均匀分布的随机函数; ω 为惯性权重,通常 ω 可由式(7)表示,

$$\omega = \omega_{\max} - \frac{\omega_{\max} - \omega_{\min}}{iter_{\max}} \times iter \quad (7)$$

式中, $\omega_{\max}$ 和 $\omega_{\min}$ 为 ω 的最大值和最小值, $iter$ 和 $iter_{\max}$ 分别为当前迭代次数和最大迭代次数.

2.3 支持向量机介绍

SVM 是从寻找线性可分情况下的最优分类超平面 (Optimal Hyperplane) 提出的 [9]. 分类原理: 对于二分类问题, 假定 l 个样本的训练集 $T = \{(x_i, y_i) | (x_i, y_i) \in R^n \times \{\pm 1\}, i = 1, 2, \dots, l\}$ 能被超平面 $H: w \cdot x + b = 0$ 准确无误地分开 (经验风险为 0), 并且使离超平面最近的样本到超平面的距离 (margin) 最大 (置信范围最小), 则此超平面即为最优超平面 (Optimal Hyperplane), 如图 1 所示.

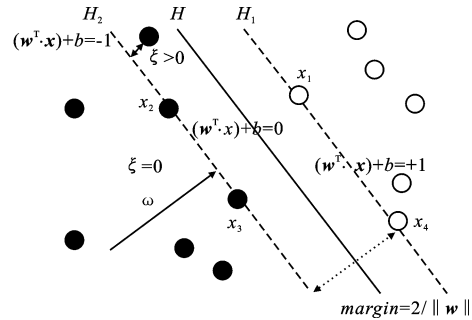


图1 支持向量机分类示意

图1中, 实心圆点和空心圆点分别表征两类样本; H 为最优分类超平面; H_1 和 H_2 是平行于 H 且经过两类中距 H 最近样本的超平面, 即类边界; margin 为分类边界间距离, 即分类间隔, 其值为 $2 / \|w\|$. SVM 在处理复杂非线性问题时采用的方法是: 首先, 将原始空间数据经非线性映射函数 φ 映射入高维特征空间, 而后再对特征空间中数据进行线性分类, 特征空间的最优分类面对应原始空间的最优分类曲面.

3 RS-PSO-SVM 特征选择算法

文献 [10] 中给出了利用粒子群算法对粗糙集属性约简过程进行迭代搜索的思路, 得益于粒子群算法的高效性, 仿真结果表明该方法能有效地降低最小约简寻优时间. 但是该算法在定义 PSO 的适应度时, 只考虑了依赖度和降维程度, 没有考虑分类精度, 使得约简结果含有冗余特征.

封装式算法 PSO-SVM [5] 的思路是利用粒子群

算法对特征集进行迭代搜索,以 SVM 对每次迭代得到的特征集进行评估,进而利用得到的评估准确率作为迭代搜索的判断准则,能得到无冗余的精简特征集.由于每次迭代都要利用 SVM 进行分类准确率评估,会导致算法计算量异常巨大,使得特征选择效率大大降低.本文综合考虑粗糙集属性约简和 PSO-SVM 势,将两种方法结合起来,提出一种新的基于粗糙集粒子群支持向量机(RS-PSO-SVM)特征选择方法,该方法在适应度函数中综合考虑属性依赖度、分类精度与降维能力三项内容来均衡选择效率与精度.

算法的具体步骤如下所示.

输入:系统 $S = (U, A)$, U 为论域, $A = C \cup D$, C 和 D 分别为条件属性和决策属性.

输出:最优特征集.

Step1:利用公式 $r_i' = \frac{r_i - \min\{r_i\}}{\max\{r_i\} - \min\{r_i\}}$ (其中 r_i 为第 i 个特征分量数值)对采集得到的样本进行归一化,使得数值映射到 $[0, 1]$ 区间.

Step2:利用粗糙集中的等间隔区间分割法^[10]对连续型数值 r_i' 进行离散化处理.

Step3:产生含有样本个数微粒的种群,每个微粒的初始速度随机给出,粒子的位置由经 Step2 离散化后的样本数据给出,其中取 1 代表包含该属性,取 0 表示不包含该属性.

Step4:根据适应度公式 $f = 1/3 * k + 1/3 * accuracy + 1/3 * \frac{D - Feature_dim}{D}$ (其中 k 为属性依赖度, $accuracy$ 为分类准确率, D 为原始特征维数, $Feature_dim$ 是所选特征维数.)对微粒适应度进行计算.

Step5:比较当前适应值与历史适应值,确定新的 $pBest_i$ 和 $gBest$.

Step6:判断终止条件,如果满足则停止,如果不满足则转 Step2,继续循环迭代.

Step7:当适应值最大时,将得到的粒子位置取出,属性为 1 即为对应特征存在,以此来确定最优特征子集.

4 实验分析

为了验证算法的有效性,采用 benchmarks 数据集集中的 german 数据集,选用 1 000 个 19 维的 german 数据集,其中 800 个用作训练集,200 个用作测试集.特征选择参数设置为 PSO 的种群规模为

28,最大迭代次数为 100,加速因子均取 2, SVM 训练工具选用高斯径向基核函数,核宽度 0.8,惩罚因子为 200.对基于 RS、PSO-SVM 以及 RS-PSO-SVM 三种方法特征选择的结果进行比较.结果分别如图 2 至图 5 和表 1 所示.

表 1 特征选择结果

	特征维数	分类精度/%	花费时间/s
原始数据	19	73.9	—
RS 约简	10	88.3	0.9
PSO-SVM	5	93.1	141
RS-PSO-SVM	5	93.1	10

三种方法在特征选择过程中,不同特征维数下的分类精度如图 2 所示.可以看出,随着特征集维数的下降,三种方法的分类精度均呈逐渐上升的趋势,RS 方法最后得到了 10 维特征,分类精度为 88.3%; PSO-SVM 和 RS-PSO-SVM 最后得到了 5 维特征,分类精度均为 93.1%,且两种方法特征选择过程类似.图 3 为 RS 方法特征选择过程中分类精度随时间的变化曲线,RS 方法在 0.9 s 时分类精度达到最大值,得到 10 维特征选择集;图 4 和图 5 分别为 PSO-SVM 和 RS-PSO-SVM 方法适应度随时间的变化曲线, PSO-SVM 方法在 141 s 时适应度达到最大值, RS-PSO-SVM 方法在 10 s 时适应度达到最大值.

从图 2 至图 5 可以看出,基于粗糙集的属性约简方法去掉了 9 维特征分量,相比较其他两种方法,其最终分类精度是最低的,但特征选择所用时间却是最短的,比本文所提方法低一个数量级,本文提出的方法得到了和 PSO-SVM 完全一样的 5 维特征集,因此分类精度也完全相同,说明融合粗糙集属性约简后并未改变特征选择的准确性,但所用时间较 PSO-SVM 算法下降了 10 多倍,使得算法能以极快的速度得到最优特征集.

通过以上分析结合仿真结果,可以看出 RS-PSO-SVM 特征选择方法是有效的,能在保证所选特征集为最优情况下,极大地缩短所用时间.

5 结束语

本文综合分析了典型的两种特征选择算法,在此基础上通过结合 RS 与 PSO-SVM 方法,在适应度函数中添加属性依赖度,加速迭代过程,缩短了特种选择时间.

给出了 RS-PSO-SVM 特征选择算法的具体步骤,利用 benchmarks 数据集对提出的算法进行了验证.实验结果表明:利用提出的 RS-PSO-SVM 特

征选择算法以较短的时间获得了具有高分类精度的特征集,可以用在高维数据的分析上。

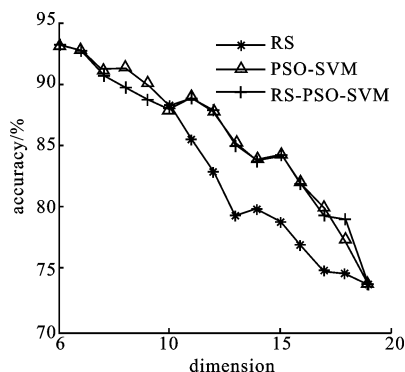


图2 三种方法在不同特征维数下的分类精度

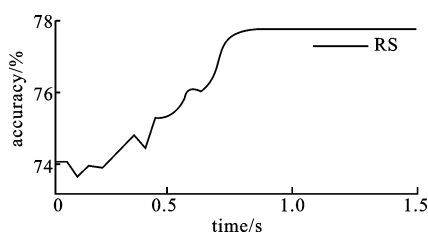


图3 RS方法下分类精度随时间的变化曲线

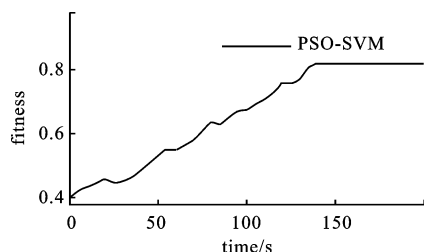


图4 PSO-SVM方法下适应度变化曲线

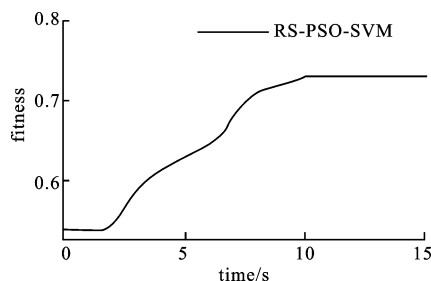


图5 RS-PSO-SVM方法下适应度变化曲线

参考文献:

- [1] 朱颖东,李红婵. 结合类内集中度和优化 RBF 神经网络的特征选择[J]. 微电子学与计算机, 2011, 28(2): 145-148.
- [2] 张俊才,张静. 使用粒子群算法进行特征选择及对支持向量机参数的优化[J]. 微电子学与计算机, 2012, 29(7): 138-141.
- [3] 史德琴,李俊山,耿志,等. 多模序列图像目标跟踪的特征融合粒子滤波实现[J]. 空军工程大学学报:自然科学版, 2011, 12(5): 61-66.
- [4] 姚旭,王晓丹,张玉玺,等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2): 161-166.
- [5] Li kun, Gao xianwen, Tian zhongda, et al. Using the curve moment and the PSO-SVM method to diagnose downhole conditions of a sucker rod pumping unit[J]. Petroleum Science, 2013, 13(10): 73-80.
- [6] Li yanbin, Zhang ning, Li cunbin. Support vector machine forecasting method improved by chaotic particle swarm optimization and its application[J]. Journal of Central South University of Technology, 2009, 16(9): 0478-0481.
- [7] Ding weiping, Wang Jiandong, Guan zhijin. Cooperative extended rough attribute reduction algorithm based on improved PSO[J]. Journal of Systems Engineering and Electronics, 2012, 23(1): 160-166.
- [8] 腾书华,周石琳,孙即祥,等. 基于条件熵的不完备信息系统属性约简算法[J]. 国防科技大学学报, 2010, 32(1): 90-94.
- [9] 温祥西,孟相如. 基于支持向量机和云模型的网络健康状态评估[J]. 北京邮电大学学报, 2012, 35(1): 10-14.
- [10] 武装. 基于粒子群的粗糙集属性约简与数控机床故障诊断研究[J]. 组合机床与自动化加工技术, 2012, 8(8): 51-54.

作者简介:

崔文岩 男, (1987-), 博士. 研究方向为网络可生存性分析.
E-mail: 690004398@qq.com.

孟相如 男, (1963-), 教授. 研究方向为宽带网.

李纪真 男, (1987-), 博士. 研究方向为网络安全预警.