

应用遗传算法优化子空间的SVM分类算法

蒋华荣 郁雪

(天津大学管理与经济学部 天津 300072)

摘要 提出了一种应用遗传算法优化子空间的SVM分类算法GS-SVM。该算法首先改进样本选择策略,采用基于置信度和凸包的样本选择方法,考虑类间距离和样本分布等因素,选择典型代表样本作为SVM的新训练集;然后采用矩阵式混合编码方式,利用遗传算法一并优化代表样本的特征子空间和SVM分类参数,并根据特征优化后的代表样本,构建SVM分类模型。在UCI的11个数据集上进行的仿真实验结果表明,该算法在大部分数据集上均可获得较小的样本规模和特征维数,以及较高的分类精度。

关键词 子空间分类,遗传算法,支持向量机,样本选择,凸包

中图分类号 TP181 文献标识码 A

GA-based Subspace Classification Algorithm for Support Vector Machines

JIANG Hua-rong YU Xue

(Department of Management and Economics, Tianjin University, Tianjin 300072, China)

Abstract This paper presented a new GA based Subspace classification algorithm for SVM(GS-SVM). A modified sample selection method is adopted to select a subset of training data based on both the confidence and the convex hull. Then the representative samples are selected to train the SVM models by considering the distances between classes and the sample distribution. The algorithm adopts the matrix-form mixed encoding. Genetic algorithm is used to optimize the feature subspace of representative samples and the classification parameters of SVM simultaneously. The SVM classification model is produced based on the representative samples with the optimized feature subspace. Experimental results on eleven UCI datasets illustrate that the proposed algorithm is able to select both smaller sample subset and feature size, and achieve higher classification accuracy than the traditional classification algorithms.

Keywords Subspace classification, Genetic algorithm, Support vector machine, Sample selection, Convex hull

1 概述

随着万维网技术和电子商务应用的迅速发展,在众多应用领域积累了大量高维复杂数据,有效分析挖掘高维复杂数据具有重大的现实意义,而高维复杂样本分类是其重要任务之一。传统的分类算法一般基于样本数据的全空间建立分类模型,然而随着数据维度的增加,数据点相对稀疏,全维空间上的数据间距离可能相差较小,因此传统的全维空间距离度量方式无法解决分类中样本相似性分析需要。此外,如果数据集中包含大量无关特征甚至噪声特征,则将直接导致分类模型识别率的降低。

特征选择采用某种评价标准来选取原始高维空间中最有效的特征或者最具有代表性的特征,以达到降低特征空间维数的目的的过程,通过特征选择滤掉不相关属性,减少维度,可以为后续计算节省大量的时间。因此处理高维分类问题的一般做法是使用特征选择选择其中一部分特征进行分类器设计,以提高分类精度或在保持分类精度的前提下简化分类模型。目前已有结合特征选择的复杂样本分类算法,例如

He等采用基于邻域的粗集模型将属性划分为强相关、弱相关不可或缺、弱相关冗余和不相关等属性子集,并在不可或缺的相关属性子集上训练分类模型^[1];Huang等用二进制编码的遗传算法(genetic algorithm, GA)同时进行属性选择和模式分类^[2];王世卿等用Relif算法得到遗传算法的初始种群,然后采用遗传算法一同优化特征子集和支持向量机参数^[3];Tian等提出一种双种群协同进化算法,将属性约减和神经网络模型优化过程协同进行,一并获得较优的网络结构和约减的属性维数^[4]。这些特征选择算法一般是对全部样本选择相同的特征属性子集,即将所有样本投影到一个全局子空间中,获得的属性子集维度较大,且没有考虑不同类样本的空间分布情况,因此分类效果受到影响。

子空间方法是近年来提出的一种统计模式识别方法^[5],其本质上是在原样本空间中寻找合适的子空间(特征子集),通过将高维样本投影到低维子空间上,在子空间上进行样本分类。一个理想的子空间分类算法应该具有以下特点:得到的分类规则简洁且约简后的维度较低,计算量较小,区分能力强,使各类样本均能很好地分离开来^[6]。子空间方法已被成

到稿日期:2013-01-29 返修日期:2013-05-10 本文受国家自然科学基金项目(61202030)资助。

蒋华荣(1989—),女,硕士生,主要研究方向为数据挖掘,E-mail: hrjiang@tju.edu.cn;郁雪(1977—),女,博士,讲师,主要研究方向为信息系统、Web智能推荐。

功用于脸部识别^[7,8]、目标识别^[9,10]和信用评估^[11]等各个方面。

子空间方法一般可分为线性子空间分类方法和非线性子空间分类方法。线性子空间分类利用一个线性子空间来表示每一个类,将高维数据中那些主要的、具有代表性的特征进行线性重组。例如 Watanabe 等提出了首个子空间分类算法 CLAFIC,其采用主成分分析,以相关矩阵的部分特征向量来构造子空间,实现了特征信息的压缩^[12];Wen 提出一种改进的判别常向量(discriminative common vectors)对判别准则的类间矩阵和类内矩阵进行微调,以提高子空间分类精度^[8]。非线性子空间分类假设每个类的样本分布在线性子空间中,处理非线性情况的常用方法是核方法。Zhang 等提出一种新的基于核子空间算法,该算法通过集成特征空间中的局部判别信息来降维,并使用核函数来解决非线性问题,以实现全局和局部最优分类^[13];Kitamura 等提出了基于子空间的最小二乘支持向量机和基于子空间的线性规划支持向量机,使用核主成分分析得到各个类的子空间,并用 SVM 来解决子空间之间存在覆盖的问题^[14];Wang 等提出一种类依赖的属性选择方法,即利用封装法分别对每个类进行属性选择,该方法能克服噪音的影响且分类精度高于不考虑分类类别的属性选择方法^[15]。Qiu 等提出了一种信息样本子空间方法(informative sample subspace, ISS),其利用信息论解决非线性子空间分类问题,将输入向量和目标类标号之间的互信息应用在子空间分类中^[6]。

在一些高维复杂数据集中,同一个类的样本空间中仍可能存在多个分布紧凑的子空间。本文提出一种应用遗传算法优化子空间的 SVM 分类算法(GS-SVM)。该算法首先采用改进的基于置信度和凸包的样本选择策略,在计算类间距离的同时考虑了样本分布的影响,以获得更准确的边界代表样本作为约简的 SVM 训练集合;然后采用矩阵式混合编码方式,利用遗传算法优化各代表样本的子空间特征和对应的 SVM 分类参数,使得每个代表样本可投影到不同的特征子空间,并根据特征优化后的代表样本构建 SVM 分类模型。本文所提算法的创新之处在于:

(1)同时考虑了类间样本距离和每个类的样本分布特点,改进样本选择算法,既能解决只考虑类间距离的样本选择算法对于那些类间无重叠或距离较远的数据集无法选出有效样本的问题,又能用复杂度较小的改进算法来近似每个类的样本分布情况,可得到更具有代表性的支持向量备选集,在一个规模较小的代表样本集上建立分类模型,可大大节约时间和空间成本。

(2)采用矩阵式混合编码方式,用遗传算法同时实现子空间特征选取和分类参数选择,可得到区分度更高和更具代表性的子空间,同时提高子空间分类精度。

(3)对于样本选择算法选出的每个代表样本都通过遗传算法选择和优化不同子空间特征,充分考虑那些同类但可能具有不同子空间特征的样本分布,可得到更接近样本原始分布的子空间。

本文第 2 节提出了基于置信度和凸包的支持向量混合选择算法;第 3 节针对已选择出的代表样本,利用遗传算法优化

其特征子空间,提出完整的 GS-SVM 算法,在代表样本及其子空间上训练 SVM 分类器;第 4 节介绍实验过程和实验结果分析;最后进行总结并指出进一步的研究方向。

2 支持向量选择

支持向量机(support vector machine, SVM)是 Vapnik 提出的一种基于结构化风险最小原则的机器学习方法^[16],其中思想是找到两类样本的最优分离超平面,使不同类的样本分别位于超平面的两侧且使两侧的空白区域最大。

2.1 支持向量选择研究现状

支持向量机需要求解一个受约束的二次规划问题,因此当训练集规模较大时,会出现训练速度慢、算法复杂、效率低下等问题。如果选择部分代表样本作为 SVM 支持向量来训练分类模型,则将极大地降低训练时间和存储开销。Cortes 和 Vapnik 提出 SVM 分类模型只由支持向量确定而与其他样本无关。研究学者已经提出了许多的样本选择方法。Shin 等提出一种基于邻域的模式选择方法,为了加速样本选择过程,该方法只考虑那些可能位于边界的样本。首先随机选出一部分样本作为初始样本,然后找出初始样本的邻居,只考虑具有异类邻居的样本,寻找这些样本的邻居的邻域,对那些处理过的样本进行标记,重复以上过程,直到所有位于分界边缘的样本都被选择和标记。该算法的模式选择结果受初始选择样本的影响很大,尤其是对分布复杂的样本(例如多峰的样本),可能由于初始样本选择不当而导致重要的边界样本无法被选中^[17];He 等提出一种基于邻域粗集模型的样本选择方法,该方法将样本划分为积极样本和边界样本,积极样本的邻域内样本的类标号相同,边界样本的邻域来自不同的类,使用这些边界本来训练 SVM 分类模型。但当样本维度足够大时,所有样本都将落入积极区域,使用该邻域粗集模型将无法准确得到边界样本^[1]。翟俊海等提出一种基于粗糙集技术的压缩近邻样本选择方法,该方法用粗糙集进行特征选择,然后选出特征空间中靠近边界域的样本^[18]。Wang 等提出一种基于置信度的样本选择方法,即以每一个样本为中心画一个不包含其他类的最大的球,将落在球内的样本数作为评价基准来确定支持向量备选集,落在某样本的球内的样本数越少该样本越有可能是支持向量,根据该评价基准,对每一个样本 x_i ,计算落在该样本的不含别类样本的最大球中的样本数 $N(x_i)$,根据 $N(x_i)$ 对所有样本进行排序,选择具有最小 $N(x_i)$ 的部分样本构成支持向量备选集。该算法考虑来自不同类的样本的信息,选择靠近其它类的样本作为支持向量,在一些数据集上能保持分类准确率的同时约减样本数目,但对于异类样本耦合较大的情况,类边界样本分布复杂,在一定程度上影响了该方法的分类性能^[19]。

SVM 的几何解释表明,对于可分的两类分类问题, SVM 的最优分离超平面与平分连接正负类样本凸包中两个最近点的线段的超平面相同;对于不可分的情况,将凸包换为约减的凸包,可以得到类似的结果。一般地,对于给定数据集 $D \subset \mathbb{R}^m$, $D = \{x_1, x_2, \dots, x_n\}$, D 的凸包可以表示为:

$$\text{cover}(x_1, x_2, \dots, x_n) = \left\{ \sum_{i=1}^n a_i x_i \mid x_i \in D, a_i \in [0, 1] \right\} \quad (1)$$

凸包可以看作样本的近似分布^[20]。如果能求出类的凸

包就能知道各个类的近似分布情况,但求解样本的凸包是一个 NP 难题^[21]。受标准 SVM 的几何解释的启发,Zhou 等提出一种基于凸包的样本选择方法,它将位于凸包附近的边界样本作为代表样本,代替整个凸包样本^[22]。Zhou 等定义了样本与凸包之间的距离。定义样本 y 与 D 的凸包之间的距离为 y 与 D 的所有凸包中最近的那个凸包之间的距离,该距离的求解是一个求最小值的二次规划问题。该算法在选择样本时只需考虑同类样本,选择部分凸包的边界样本来训练 SVM,但每次计算某个样本到样本子空间凸包的距离就必须求解一次二次规划,计算量较大。因此,该算法中求解样本到凸包的距离占了绝大部分开销,一定程度上制约了其在实际应用中的应用。

2.2 基于置信度和凸包的支持向量混合选择算法(CCBSS)

本文针对上述问题,提出了一种基于置信度和凸包的支持向量混合选择算法。该算法首先改进 Zhou 算法中的距离求解方法,将该样本到边缘样本集 CS 中各样本的距离之和作为该样本到 CS 的凸包的距离。给定一个数据集 $D \subset R^m$, $D = \{x_1, x_2, \dots, x_n\}$, 边缘样本集 $CS = \{x_1, x_2, \dots, x_k\} \subset D$, $k \leq n$, 样本 $x \in R^m$ 与 CS 之间的距离可以定义为:

$$\text{dist}(x, CS) = \sum_{i=1}^k d(x, x_i) = \sum_{i=1}^k \|x - x_i\| \quad (2)$$

基于改进的距离计算方法,确定每类凸包的边界样本。对于每类样本,先选择两个距离最远的样本构成初始样本子集,然后计算剩余的样本到该样本子集中两样本的距离和,选择距离和最大的样本添加到该样本子集中,所选的样本也被认为是该类凸包的边界样本,重复将距离和最远的样本添加到该样本子集中直到满足终止条件。利用改进后的距离计算方法求解凸包边缘样本的算法(Improved convex based sample selection method, ICSS)流程如表 1 所列。给定数据集 $DS = \{x_1, x_2, \dots, x_n | x_i \in R^m, i \in [1, n]\}$, m 和 n 分别表示样本的维度和数量。

表 1 改进的凸包边缘样本选择算法(ICSS)

输入: DS, num // DS 是样本集, num 是需要选择的边缘样本数
输出: Csample // 边缘样本集合
步骤 1 建立初始选择集: $Csample = \{Z_1, Z_2 [x_a, x_b] = \arg \max_{x_p, x_q \in DS} d(x_p, x_q)\}$
步骤 2 $L = 2$ // 已选样本数
步骤 3 while $L \leq num$
对 $\forall x \in DS \setminus Csample$ 计算 $\text{dist}(x, Csample)$
$ZL = \arg \max_{x \in DS \setminus Csample} \text{dist}(x, Csample)$
$Csample = Csample \cup ZL$
$L = L + 1$
end while
步骤 4 输出 Csample

采用该算法在寻找凸包的边缘样本时无需求解二次规划问题,在保证所选样本具有良好边缘性能的同时,使得求解样本到凸包距离所需的计算量大大减少。

最后,参考 Wang 的基于置信度的样本选择算法,在利用改进的凸包边缘样本选择算法得到凸包的边缘样本集后,进一步采用基于置信度的样本选择法来选择靠近其他类的样本。表 2 给出了完整的基于置信度与凸包边缘的样本选择算法(confidence and convex based sample selection method, CCBSS)。数据集 $D \subset R^m$, D 包含 C 个类, S_i 表示第 i 个类的样本集合, n_i 为第 i 个类的样本数, $x_j^i, i \in \{1, 2, \dots, C\}, j \leq n_i$

表示第 i 类的第 j 样本。

表 2 基于置信度与凸包边缘的样本选择算法(CCBSS)

输入: D, R, num // R 是选择样本的百分比, num 是要选择的凸包边缘样本数量
输出: Boundary // 选择样本的集合
步骤 1 Boundary = $\emptyset$ // 开始时集合为空
步骤 2 for $i = 1$ to C // C 是类的个数
输入第 i 个类的样本集 S_i
$CS = \text{ICSS}(S_i, num)$ // CS 为用 ICSS 算法得到的凸包边缘样本
Boundary = Boundary $\cup$ CS
$K = R \times n_i$ // K 是用置信度的方法从第 i 类样本中选择样本数
For 每一个 $x_j^i \in S_i$
$\text{sortx} = \text{sort}(\text{dist}(x_j^i, x_p^i)), x_p^i \in D, p \in [1, C]$
// 按距离从小到大升序排列
从 sortx 中找出第一个与 x_j^i 类标不同的 x_p^i 在 sortx 中的位置 V, x_j^i 的邻域中同类样本数为 $\text{nb}(x_j^i) = V - 2$
endfor
index = sort($\text{nb}(x_j^i)$)
$x_j^i \in S_i$
将 index 的前 K 项赋予 Cboundary
Boundary = Boundary $\cup$ Cboundary
endfor
步骤 3 输出 Boundary

为了验证 CCBSS 算法样本选择性能,随机生成一个包含 3 个类的两维数据集,3 个类分别满足均值为 1、3、9,方差均为 1 的正态分布,每个类包含 100 个样本,分别采用基于置信度的算法、KNN($K=5$)、ICSS 和 CCBSS 来进行样本选择,结果如图 1 所示。“x”、“+”和“ Δ ”分别表示 3 个类的样本,被圆圈住的样本是算法选择的样本,从图 1(a)和(b)中可以看出基于置信度的样本选择方法和 KNN 能有效选出距离较近的第 1 和第 2 类样本中位于分界面附近的代表样本,但对于像第 2 和第 3 类样本这种距离相对较远且类之间无重叠的情况,就无法选出那些位于分界面附近的样本。从图 1(c)中可以看出,ICSS 算法能有效选出每个类的边缘样本,并将第 2 和第 3 类样本较好地分离开来,但却无法有效选择第 1 类和第 2 类样本中重叠区域的代表样本。图 1(d)中 CCBSS 算法能选出各个类的最具代表性的样本,能有效地将 3 个类区分开来,这说明 CCBSS 既能选出有重叠的类的具有区分能力的代表样本,又能选出那些距离较远的类的代表样本。

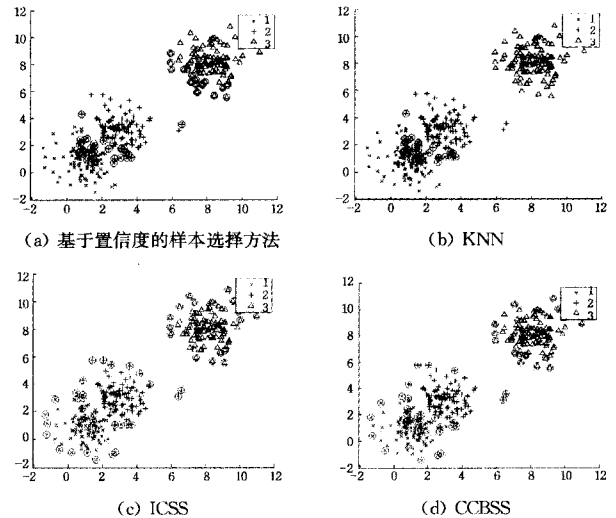


图 1

3 GS-SVM 算法

遗传算法是通过模拟生物系统中的自然选择和遗传机

制,以群体为进化基础,以适应度函数为评价依据,对群体中个体进行遗传操作的自适应优化搜索算法,具有较好的全局搜索能力。

针对高维复杂数据集中同类样本可能存在多个分布紧凑的子空间,本文采用遗传算法优化由基于置信度与凸包边缘的样本选择算法选出的代表样本的特征子空间,构建 SVM 分类模型。下面从编码、初始化、适应值评价及选择、交叉以及变异操作等方面逐一阐述,提出完整的 GS-SVM 算法。

3.1 编码

支持向量机一般通过引入核函数来处理线性不可分的情况,通过核函数将样本投影到高维特征空间,并在该特征空间中构造最优分类面。本文采用基于 RBF 核函数的 SVM 作为分类器,RBF 核参数的确定包括 RBF 核函数的半径 γ 和常数 C , C 是对最少错分样本和最大分类间隔的折中,控制对误分样本的惩罚程度。采用遗传算法一同优化代表样本的特征子空间和对应的核参数,染色体由 3 部分组成:特征子集选择部分、 γ 和 C 。每个个体表示的是一组代表样本的特征子集和核参数。采用矩阵式混合编码方式,特征子集选择部分采用二进制编码,对 γ 和 C 采用实数编码。种群大小用 L 表示,则种群中的个体表示为 $I_1, I_2, \dots, I_L$,个体 I_i 是一个 $n \times (m+2)$ 的矩阵:

$$I_i = [F^i \quad \gamma^i \quad c^i] = \begin{bmatrix} f_{11}^i & f_{12}^i & \dots & f_{1m}^i & \gamma^i & c^i \\ f_{21}^i & f_{22}^i & \dots & f_{2m}^i & \gamma^i & c^i \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ f_{n1}^i & f_{n2}^i & \dots & f_{nm}^i & \gamma^i & c^i \end{bmatrix} \quad (3)$$

式中, n 表示选择出的代表样本个数, m 表示样本特征维度, $i=1,2,\dots,L$ 。 $F^i = [f_s^i]_{n \times 1} = [f_s^i]_{n \times m}$,表示 n 个代表样本分别选择的 m 维特征子集, $f_s^i = 1$ 表示第 i 个个体的第 s 个样本的特征子空间中选择第 t 个属性, $f_s^i = 0$ 则表示没有选择该属性。例如, $f_1^i = (0100010)$ 表示第 i 个个体的第 1 个样本选择的特征子空间由第 2 个和第 6 个属性构成。 γ^i, c^i 分别表示第 i 个个体对应的核参数,对每个个体训练一个 SVM 分类器,因而 γ 和 C 的值是唯一的。

3.2 种群初始化

对原始数据进行规范化处理,使规范后数据在 $[0,1]$ 中取值。每个个体分两部分来初始化:随机将每个样本的属性位赋 0 或 1,得到初始子空间。为了避免得到的子空间过分稀疏,要求得到的所有样本子空间平均包含的属性要大于原来属性的 30%,至少有一个个体的每个样本子空间包含所有的属性; $\gamma \in [0.01, 50]$, $C \in [1, 100]$,在取值范围内为每个个体随机生成初始的核参数。

3.3 适应度函数及选择操作

分类准确率和子空间维度约减程度是设计适应值函数的两个标准,算法中由个体中的代表样本训练的 SVM 分类器的分类精度越高,代表样本的子空间维度约减程度越大,则该个体的适应值越大。采用加权求和方式用一个适应度函数将这两个目标结合在一起:

$$fitness_i = w_p \times precision_i + w_s \times n \times \left(\sum_{s=1}^n \sum_{t=1}^m f_s^i \right)^{-1} \quad (4)$$

式中, $precision_i$ 为该个体构成的基于子空间的 SVM 分类器的分类精度, n 和 m 分别表示样本的数量和维度。两个预先给定的权重 w_p 为分类准确率的权重; w_s 为子空间维度约减

的权重,且 $w_p + w_s = 1$ 。对精度要求越高, w_p 的取值越大,其取值范围为 $[0.5, 1]$ 。将所有样本分成 3 个子集:训练集、验证集和测试集,训练集用来训练 SVM 分类器,验证集用于评估个体的适应值,最后使用测试集来检验分类模型的准确率。

采用轮盘赌选择方法以及精英保留策略,个体的选择概率与其适应值成正比,个体 i 被选择的概率为:

$$p_i = \frac{fitness_i}{\sum_{i=1}^L fitness_i} \quad (5)$$

3.4 交叉操作

本文算法将个体的二进制部分(特征子空间选择)和实数部分(核参数)分别进行交叉操作。首先将需要进行交叉操作的个体两两分组。其次为每对个体随机生成一个长度为 m 的二进制串,用于二进制部分的交叉操作:从左到右扫描该二进制串,如果当前基因位值为 1,选择第一个父代与其对应的列,否则选择第二个父代对应的列,由此产生第一个子个体的子空间;重复以上操作,选择第一个父代中与 0 值对应的列,第二个父代中与 1 对应的列构成第二个子个体的子空间。最后为每对个体生成两个随机数 r_1 和 r_2 ($r_1 \in [0, 1]$, $r_2 \in [0, 1]$),对核参数部分进行交叉操作:

$$[\gamma_1', c_1'] = [r_1, r_2] \times [\gamma_1, c_1] + [(1-r_1), (1-r_2)] \times [\gamma_2, c_2] \quad (6)$$

$$[\gamma_2', c_2'] = [(1-r_1), (1-r_2)] \times [\gamma_1, c_1] + [r_1, r_2] \times [\gamma_2, c_2] \quad (7)$$

式中, γ_i 和 c_i ($i=1,2$) 表示交叉前的核参数, γ_i' 和 c_i' ($i=1,2$) 表示交叉后的核参数。

3.5 变异操作

个体的变异操作也分两部分来完成,变异概率均为 p_m 。二进制部分采用均匀变异,为每个个体随机生成一个 $n \times m$ 的矩阵 $M^i = [rd_s^i]_{n \times m}$,其中 $rd_s^i \in [0, 1]$, $i=1,2,\dots,L$, $s=1,2,\dots,n$, $t=1,2,\dots,m$ 。如果 $rd_s^i < p_m$,则个体 I^i 相应位置取反,即原来是 0 的变为 1,反之亦然。对个体中的实数编码部分生成一个 $[0, 1]$ 之间的随机数 r_m ,如果 $r_m < p_m$,则按下式对核参数 γ 进行变异操作:

$$\gamma'_i = \gamma_i + rand \times (\gamma^* - \gamma_i) \quad (8)$$

$$c'_i = c_i + rand \times (c^* - c_i) \quad (9)$$

式中, γ_i 和 c_i 表示原来的核参数, γ^* 和 c^* 表示精英个体的 γ 和 C 值, $rand$ 表示一个 $[0, 1]$ 之间的随机数。

3.6 GS-SVM 算法整体流程

GS-SVM 算法的流程如表 3 所列。

表 3 GS-SVM 算法的流程

步骤 1	对数据进行规范处理并分为训练集、验证集和测试集;
步骤 2	用基于置信度和改进凸包的混合样本选择方法(CCBSS)从训练集中选出代表样本集合 CS;
步骤 3	从代表样本集合 CS 得到样本的数量和维度,确定遗传算法中个体的矩阵规模,并初始化种群,预先设定最大迭代次数 G ,当前种群代数 $g=1$;
步骤 4	用评价集计算每个个体的适应值,按适应值大小进行排序,选出精英个体;
步骤 5	采用轮盘赌法生成交配池;
步骤 6	利用交叉概率对交配池中个体进行交叉操作;
步骤 7	利用变异算子,对交叉得到的个体进行变异操作得到新的种群;
步骤 8	如果 $g < G$ 且不满足收敛条件, $g=g+1$,转到步骤 2,否则转到步骤 9;
步骤 9	输出精英个体作为最终的分类模型和特征子空间。

整个 GS-SVM 算法的系统架构如图 2 所示。

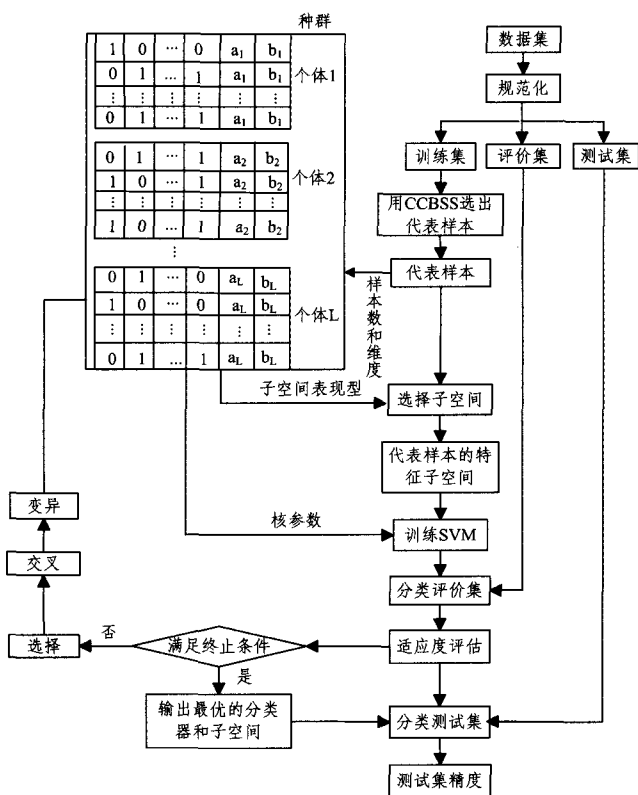


图2 GS-SVM算法的系统架构

4 实验

为了检验本文提出算法的性能,选取了11个来自UCI的真实数据集,各数据集简要介绍如表4所列。所有的数据集分成3部分:50%的数据为训练集,25%为评价集,25%为测试集,对于每个数据集,算法运行30次,取平均结果。

表4 实验数据集

数据集	训练集	评价集	测试集	类别	属性
diabetes	384	192	192	2	8
wdbc	285	142	142	2	31
german	500	250	250	2	24
breast-w	341	171	171	2	9
wpbc	99	44	45	2	33
spambase	2301	1150	1150	2	57
spectf heart	133	67	67	2	44
yeast	742	371	371	10	8
dermatology	183	91	92	6	33
segment	1155	577	578	7	19
vehicle	423	211	212	4	18

4.1 实验1

本实验不考虑特征子空间的选择,而是在全维空间中验证基于置信度和凸包的样本选择算法CCBSS的性能。对比算法分别为Wang等人提出的基于置信度的样本选择算法,K-最近邻样本选择法($K=5$)以及直接采用原始数据训练分类器,所有算法均采用SVM作为分类器,使用RBF核。对spambase和dermatology数据集,CCBSS算法选择样本百分比 $R=0.3$,其他数据集中 $R=0.5$,选择的凸包边缘样本数量 num 取样本维度与样本数中的较小值。基于置信度的算法中参数 R 取值与CCBSS通过凸包和置信度进行样本选择后的最终选择样本百分比相同。采用5倍交叉验证来确定核参数 $C \in \{0.5, 1, 5, 10, 30, 50, 100\}$, $\gamma \in \{0.01, 0.1, 0.5, 1, 5, 10, 15, 20, 30, 50\}$ 。

表5、表6分别给出了30次实验平均选择的代表样本数

和平均测试精度。从实验结果可以看出,虽然CCBSS和置信度方法比K-最近邻法所选代表样本略多,但能得到准确率更高的分类模型,尤其是对于多类且样本分布复杂的数据集,如dermatology, segment和vehicle,说明CCBSS和置信度方法更适于处理两类和复杂的多类分类问题,适用范围更广。CCBSS的分类精度在8个数据集上高于基于置信度的样本选择方法,并且在german, segment和vehicle等数据集上明显优于后者。此外,CCBSS算法训练样本仅为原始训练样本的48.3%,样本数大大减少,减少了数据的存储和处理开销,提高了算法效率。该实验说明CCBSS在置信度方法的基础上加入反映数据分布的样本是基本有效的,尤其是在复杂多类的数据集上。

表5 CCBSS和对比算法的平均代表样本数

数据集	原始数据	CCBSS	置信度	KNN
diabetes	384	199	199	250
wdbc	285	168	168	40
german	500	270	270	356
breast-w	343	177	177	32
wpbc	99	75	75	71
spambase	2301	774	774	751
spectf heart	133	106	106	83
yeast	345	337	337	541
dermatology	742	156	156	30
segment	183	608	608	165
vehicle	1155	249	249	300
Ave.	588	284	284	238

表6 CCBSS和对比算法的平均测试准确率

数据集	原始数据(%)	CCBSS(%)	置信度(%)	KNN(%)
diabetes	74.76	74.00	74.00	75.39
wdbc	97.45	97.52	97.17	95.34
german	72.00	70.84	67.41	73.40
breast-w	97.10	96.83	97.05	93.97
wpbc	76.29	76.29	76.29	76.29
spambase	93.26	91.83	91.13	90.47
spectf heart	79.70	79.70	79.37	79.52
yeast	62.64	57.53	57.51	61.04
dermatology	96.97	97.47	97.15	76.74
segment	91.57	93.35	92.29	45.45
vehicle	71.25	63.12	62.05	60.82
Ave.	83.00	81.68	81.04	75.31

4.2 实验2

本实验验证GS-SVM算法的性能,实验参数设置如下:种群规模 L 为60,最大迭代数 G 为200,交叉概率 P_c 为0.8,变异概率 P_m 为0.2。适应值函数中的准确率权重 w_p 设为0.8,子空间约减维度 w_s 设为0.2。终止条件为达到最大迭代次数或者是最优个体的适应值连续50代没有提高。

将本文所提的GS-SVM算法分别与GC-SVM, GKNN-SVM算法进行比较,其中GC-SVM, GKNN-SVM分别为采用置信度和采用K-最近邻选择样本,然后用GA优化子空间,用SVM进行子空间分类,这3种算法除了样本选择方法不同,其他进化操作和参数完全一样。表7列出GS-SVM和其他两种算法的平均测试精度和平均子空间维度约减率。比较表6和表7可以看出做子空间选择后测试精度都有所提高,3种算法分别提高了1.15%, 1.17%和3.46%。GS-SVM平均分类准确率最高,GC-SVM算法其次, GKNN-SVM相对较低。3种算法的维度约减率达到43%左右,说明GA能很好地选择出适应不同代表个体的特征子空间。

表7 GS-SVM、GC-SVM 和 GKNN-SVM 的试验结果

数据集	GS-SVM		GC-SVM		GKNN-SVM	
	维度约减(%)	准确率(%)	维度约减(%)	准确率(%)	维度约减(%)	准确率(%)
diabetes	73.63	49.86	72.62	51.56	74.44	53.88
wdbc	95.70	45.30	95.33	51.52	95.40	53.49
german	73.27	45.76	72.67	44.88	74.44	26.73
breast-w	95.81	70.49	95.83	71.26	95.69	70.61
wpbc	74.63	51.15	73.54	48.85	72.11	53.27
spambase	92.44	35.17	92.01	32.73	91.70	36.23
spectf heart	76.07	42.10	77.11	42.93	76.97	42.65
yeast	63.17	41.75	59.08	38.58	60.64	38.16
dermatolog	96.22	54.76	96.37	53.48	81.84	49.49
segment	92.31	24.29	92.07	24.68	69.52	38.83
vehicle	77.86	16.06	77.69	16.82	73.69	14.83
Ave.	82.83	43.34	82.21	43.39	78.77	43.47

表8 属性排序比较

数据集	GS-SVM	信息增益	SVM	RELIEF
diabetes	2,6	2,6	2,6	2,6
wdbc	23,22,12,26, 8,30,29,28, 15,31,3,27,9	24,25,22,29, 9,4,5,2,8, 15,28,12,14	22,29,24,23, 9,25,30,2, 26,5,12,3,4	22,29,24,23, 2,4,25,9,5, 8,3,28,26
german	15,1,2,3,18, 9,16,24,5, 14,22,23	1,3,2,5,4,9, 21,10,6,11, 17,16	2,3,1,19,18, 5,7,17,12, 15,16,10	1,5,3,14,17, 8,6,9,2,7, 16,13
breast-w	3,6,2,1,8	2,3,6,7,5	3,6,1,7,5	6,1,3,2,8
wpbc	25,15,12,20, 10,23,32,26, 31,21,24,1, 4,22,13,17	1,33,12,10, 11,15,16,13, 14,4,5,2,3, 8,9,6	22,13,1,33, 18,6,10,20, 11,21,26,3, 25,7,24,19	1,23,33,7, 10,3,9,8,26, 32,6,11,21, 28,18,29
spambase	7,52,24,16, 6,22,5,55, 19,27,49,13, 36,28,9,48, 53,45,10,46, 44,21,50,38, 26,34,11	52,53,56,7, 21,55,16,24, 57,25,23,27, 5,19,26,3, 11,17,2,8,6, 20,10,9,18, 12,54	53,7,23,16, 24,52,8,56, 25,17,20,57, 27,42,5,26, 46,45,21,18, 22,33,9,44, 15,6,49	27,25,12,23, 40,2,30,21, 32,34,9,26, 11,28,57,17, 5,7,37,43, 19,15,53,3, 1,36,45
spectf heart	28,6,39,41, 43,14,1,38, 8,9,10,20,7, 25,24,40,22, 31,26	41,42,27,17, 43,7,45,44, 26,31,40,16, 5,15,33,29, 8,37,35	41,33,27,34, 3,7,28,30, 12,44,11,15, 29,5,43,14, 16,36,26	45,43,44,27, 26,42,41,31, 7,16,17,30, 40,37,25,11, 15,36,9
yeast	3,4,8,1	3,4,8,2	4,8,3,2	3,4,8,1
dermatology	4,28,2,16, 21,5,23,15, 6,17,22,19, 18,12,27	20,21,22,33, 29,27,12,28, 25,6,8,9,16, 15,10	20,5,29,21, 15,31,22,28, 6,33,24,34, 26,25,7	21,33,22,28, 20,27,29,6, 12,16,25,8, 9,15,4
segment	2,1,15,12, 17,13,7	12,11,14,18, 20,13,17	20,3,12,7, 17,18,15	20,13,18,11, 12,14,3
vehicle	18,17,1,14, 10,4,3,2	12,7,8,11,9, 3,6,2	8,14,3,5,10, 17,4,7	8,18,7,12,9, 3,10,11

此外,为验证本文所提算法识别重要属性的能力,计算最终获得的子空间中第 t 个属性被选择的比重:

$$FW(t) = \frac{\sum_{s=1}^n \tilde{f}_s}{n} \quad (10)$$

式中, $\tilde{f}_s$ 为遗传算法优化最终获得的精英个体中代表样本属性选择部分, $\tilde{f}_s = 1$ 表示第 s 个代表个体优化的子空间中选择了第 t 个属性, $\tilde{f}_s = 0$ 则表示没有选择该属性, $t = 1, 2, \dots, m, n$ 为代表样本个数。根据 $FW(t)$ 由大到小对属性进行排序, $FW(t)$ 越大说明越多样本子空间包含第 t 个属性, 该属性越重要。为了验证该排序的有效性, 选择基于信息增益、SVM 和 RELIEF 的属性排序算法做对比实验, 实验结果如表 7 所列。表中 GS-SVM 选择 $FW(t) > \text{mean}(FW)$ ($t = 1, \dots, m$) 的属性 t , 为了方便对比, 所有算法选择的属性数量相同。

从表 8 可以看出在大部分数据集上, GS-SVM 与其他 3 个属性排序算法得到的排序结果相差不大, 特别是在数据集 diabetes, wdbc, german, breast-w, yeast, segment 和 vehicle 上, 说明 GS-SVM 算法得到的属性排序是有意义的, 可以用 GS-SVM 来进行属性排序。表 9 给出了在所有数据集上各算法选择的相同属性个数。从实验结果可以看出, 各个算法选择的相同属性个数相差不大, GS-SVM 的属性选择效果与基于 SVM 的属性选择方法最相似, 从而进一步验证了 GS-SVM 能够选择出重要属性来构造子空间。

表9 各个算法选择相同属性个数

算法	GS_SVM	信息增益	SVM	RELIEF
GS_SVM	128	62	69	67
信息增益	—	128	79	90
SVM	—	—	128	77
RELIEF	—	—	—	128

4.3 实验 3

此外, 我们还将所提算法与一些典型的传统算法进行比较, 例如随机子空间 (Rand Subspace)、KPCA 和广义判别分析 (generalized discriminant analysis, GDA), 实验结果如表 10 所列。

表10 GS-SVM 与其他算法的测试准确率

数据集	GS-SVM(%)	RandSubspace(%)	KPCA(%)	GDA(%)
diabetes	73.63	74.00	69.73	75.77
wdbc	95.70	92.95	85.56	96.77
german	73.27	73.19	70.53	70.00
breast-w	95.81	95.28	67.01	96.98
wpbc	74.63	74.67	73.13	70.72
spambase	92.44	93.16	82.76	91.82
spectf heart	76.07	79.22	70.95	79.62
yeast	63.17	58.82	54.90	62.10
dermatology	96.22	94.27	69.29	96.93
segment	92.31	94.41	94.18	95.03
vehicle	77.86	70.55	70.63	76.45
Ave.	82.83	81.87	73.56	82.93

GS-SVM 算法在 german, yeast 和 vehicle 等数据集上取得了较高的分类精度, GS-SVM 的平均分类精度明显高于 RandSubspace 和 KPCA, 与 GDA 分类精度相当, 在传统的子空间算方法中 GDA 的平均精度要远高于 KPCA。

结束语 本文提出一种改进的样本选择方法 CCBSS, 同时考虑类间信息和样本分布, 用 CCBSS 选出最有可能成为支持向量的代表样本, GS-SVM 通过 GA 来优化 CCBSS 选出的代表样本, 得到代表样本的特征子空间和 SVM 分类模型。在 11 个数据集上进行了实验, 实验结果表明 GS-SVM 能够获得更简单的 SVM 分类模型且分类精度高于大部分其他的算法, 此外, GS-SVM 还能用来鉴别重要属性。GS-SVM 不需要进行特征转换, 得到的分类结果更易于解释, 其提供了一种同时进行子空间选择和分类的分类模型, 能得到更高的分类精度。下一步的工作将采用协同进化算法来优化分类模型, 考虑使用多目标来进一步提高分类精度。

参考文献

- [1] He Qiang, Xie Zong-xia, Hu Qing-hua, et al. Neighborhood based sample and feature selection for SVM classification learning[J]. Neurocomputing, 2011, 74(10): 1585-1594

(下转第 275 页)

表明,该算法能够有效提高小类样本以及整体分类性能。

由于数据集本身的多样性和复杂性,样本的分布也呈现多样性,如果能够准确估计多数类样本的潜在分布,根据不同的分布采取不同的聚类方式以及智能采样技术,将会显著提高分类性能。此外,考虑对数据集用不同聚类方式进行多次聚类,然后通过某种策略对聚类结果进行融合,进一步提高少数类和数据集的整体分类性能是今后需要进一步研究的内容。

参考文献

- [1] Chawla N V, Bowyer K, Hall L, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357
- [2] Tomek I. Two modifications of CNN[J]. IEEE Transaction on Systems, Man and Communications, 1976, 26(1): 769-772
- [3] Kermanidis K, Maragoundakis K, Fakotakis N, et al. Learning greek verb complements; addressing the class imbalance[C]//Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, 2004; 1065-1071
- [4] Yen Show-jane, Lee Yue-shi. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset[C]//Proceedings of Intelligent Control and Automation, Series; Lecture Notes in Control and Information Sciences. Berlin/Heidelberg: Springer, 2006; 731-740
- [5] 蒋盛益, 苗邦, 余雯. 基于一趟聚类的不平衡数据下抽样算法

[J]. 小型微型计算机系统, 2012, 33(2): 232-236

- [6] 李雄飞, 李军, 屈成伟, 等. 数据挖掘中平衡偏斜训练集的方法研究[J]. 计算机研究与发展, 2012, 49(2): 346-353
- [7] 韩敏, 朱新荣. 不平衡数据分类的混合方法[J]. 控制理论与应用, 2011, 28(10): 1485-1489
- [8] 刘霄影, 吴建鑫, 周志华. 一种基于级联模型的类别不平衡数据分类方法[J]. 南京大学学报: 自然科学版, 2006, 42(2): 148-155
- [9] Tang Y, Zhang Y Q, Chawla N V, et al. SVMs modeling for highly imbalanced classifications[J]. IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(1): 281-288
- [10] 凌晓峰, Sheng V S. 代价敏感分类器的比较研究[J]. 计算机学报, 2007, 30(8): 1203-1212
- [11] 翟云, 杨炳儒, 曲武. 不平衡类数据挖掘研究综述[J]. 计算机科学, 2010, 37(10): 27-32
- [12] Ertekin S, Huang J, Bottou L, et al. Learning on the border: active learning in imbalanced data classification[C]//Proceedings of the ACM Conference on Information and Knowledge Management, Lisbon, Portugal, 2007; 127-136
- [13] 井小沛, 汪厚祥, 袁凯. 一基于修正核函数 SVM 的网络入侵检测[J]. 系统工程与电子技术, 2012, 34(5): 1036-1039
- [14] 李雄飞, 李军, 董元方, 等. 一种新的不平衡数据学习算法 PC-Boost[J]. 计算机学报, 2012, 35(2): 202-209
- [15] 林智勇, 郝志峰, 杨晓伟. 若干评价准则对不平衡数据学习的影响[J]. 华南理工大学学报: 自然科学版, 2010, 4(38): 126-135

(上接第 260 页)

- [2] Huang C L, Wang C J. A GA-based feature selection and parameters optimization for support vector machines[J]. Expert Systems with Applications, 2006, 35(4): 231-240
- [3] 王世卿, 曹彦. 基于遗传算法和支持向量机的特征选择研究[J]. 计算机工程与设计, 2010, 31(18): 4088-4092
- [4] Tian Jin, Li Min-qiang, Chen Fu-zan. Dual-population based co-evolutionary algorithm for designing RBFNN with feature selection[J]. Expert Systems with Applications, 2010, 37(10): 6904-6918
- [5] Oja E. Subspace methods of pattern recognition[M]. New York: Research Studies Press, 1983
- [6] Qiu Guo-ping, Fang Jian-zhong. Classification in an informative sample subspace[J]. Pattern Recognition, 2008, 41(3): 949-960
- [7] Cevikalp H, Neamtu M, Barkana A. Kernel common vector method: A Novel nonlinear subspace classifier for pattern recognition[J]. IEEE Transactions On Systems, Man and Cybernetics, 2007, 37(4): 937-951
- [8] Wen Ying. An improved discriminative common vectors and support vector machine based face recognition approach[J]. Expert Systems with Applications, 2012, 39(4): 4628-4632
- [9] Sakano H, Mukawa N, Nakamura T. Kernel mutual subspace method and its application for object recognition[J]. Electronics and Communications in Japan (Part II: Electronics), 2005, 88(6): 45-53
- [10] Kazuhiro F, Osamu Y. The kernel orthogonal mutual subspace method and its application to 3D object recognition[C]//8th Asian Conference on Computer Vision. Tokyo, 2007; 467-476
- [11] Zhu Mei-hong, Li Ai-hua. Random subspace method for improving performance of credit cardholder classification[C]//Modeling Risk Management for Resources and Environment in China.

Berlin, 2011; 257-264

- [12] Watanabe S, Lambert P F, Kulikowski C A, et al. Evaluation and selection of variables in pattern recognition[C]//Computer and Information Sciences II. New York, 1967
- [13] Zhang Peng, Peng Jing, Domeniconi C. Kernel pooled local subspaces for classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2005, 35(3): 489-502
- [14] Kitamura T, Takeuchi S, Abe S, et al. Subspace-based support vector machines for pattern classification[J]. Neural Networks, 2009, 22(5/6): 558-567
- [15] Wang Li-po, Zhou N, Chu Feng. A General Wrapper Approach to Selection of Class-Dependent Features[J]. IEEE Transactions on Neural Networks, 2008, 19(7): 1267-1278
- [16] Vapnik V N. The nature of statistical learning theory [M]. Springer, 1999
- [17] Shin H, Cho S. Invariance of neighborhood relation under input space to feature space mapping[J]. Pattern Recognition Letters, 2005, 26(6): 707-718
- [18] 翟俊海, 李胜杰, 王熙照. 基于粗糙集技术的压缩近邻规则[J]. 计算机科学, 2012, 39(2): 236-239
- [19] Wang Ji-gang, Neskovic P, Cooper L N. Training data selection for support vector machines[C]//Advances in Natural Computation. Changsha, 2005; 554-564
- [20] 周晓飞, 姜文瀚, 杨静宇. 基于子空间样本选择的最近凸包分类器[J]. 计算机工程, 2008, 34(12): 167-168
- [21] 姜文瀚, 周晓飞, 杨静宇. 核子类凸包样本选择方法及其 SVM 应用[J]. 计算机工程, 2008, 34(16): 212-214
- [22] Zhou Xiao-fei, Jiang Wen-han, Tian Ying-jie, et al. Kernel subclass convex hull sample selection method for svm on face recognition[J]. Neurocomputing, 2010, 10-12(73): 2234-2246