基于超椭球支持向量机的兼类文本分类算法

秦玉平[1] 王 祎[2] 伦淑娴[3] 王秀坤[4]

(渤海大学工学院 锦州 121013)[1] (渤海大学数理学院 锦州 121013)[2]

(渤海大学新能源学院 锦州 121013)[3] (大连理工大学计算机科学与技术学院 大连 116024)[4]

摘 要 提出一种基于超椭球支持向量机的多类文本分类算法。对每一类样本,利用超椭球支持向量机方法在特征空间求得一个超椭球,使其包含该类尽可能多的样本,同时将噪音点排除在外。分类时,利用待分类样本映射到每个超椭球球心的马氏距离确定其类别。在标准数据集 Reuters 21578 上的实验结果表明,该算法有效地提高了分类精度。

关键词 超椭球支持向量机,兼类分类,马氏距离

中图法分类号 TP181 **文献标识码** A

Multi-label Text Classification Algorithm Based on Hyper Ellipsoidal SVM

QIN Yu-ping[1] WANG Yi[2] LUN Shu-xian[3] WANG Xiu-kun[4]

(College of Engineering, Bohai University, Jinzhou 121013, China)[1]

(College of Mathematics and Physical, Bohai University, Jinzhou 121013, China)[2]

(New Energy College, Bohai University, Jinzhou 121013, China)[3]

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)[4]

Abstract A new multi-label text classification algorithm based on hyper ellipsoidal support vector machines was proposed. To each class sample, the hyper ellipsoidal that includes as much the class samples as possible and push the outlier samples away is trained in the featuer space. For the sample to be classified, the mahalanobis distance from the sample mapping to the center of each hyper ellipsoidal were used to decide the sample classs. The results of the experiment show that the proposed algorithm has a higher classification accuracy.

Keywords Hyper ellipsoidal SVM, Multi-label classification, Mahalanobis distance

1 引言

支持向量机(Support Vector Machines, SVM)是一种建立在统计学习理论基础上的新的分类技术[1]。它基于结构风险最小化原则,根据有限样本信息在模型的复杂度和学习能力之间寻求最佳的折衷,由于其出色的泛化性能,已成为目前解决文本分类问题的主要工具[2,3]。

支持向量机本质上是二类分类器,在处理多分类问题时,往往将其分解成一系列的二分类问题加以解决。常见的处理方法包括 1-a-r[4]、1-a-1[5]和 DAGSVM [6]。但这些方法都是针对一个样本属于一个类别的情况提出的。对兼类样本的分类问题尚未得到较深入的研究。文献[7]提出了一种基于 DAGSVM 的兼类文本分类算法,它但该方法存在不可分区域,并且分类速度较慢。文献[8]提出了一种基于超球支持向量机的兼类文本分类算法,它通过在特征空间求得最优超球面把每类样本最大限度地进行分离,但该算法只适合于每类样本呈超球形分布且聚类程度较高的情况。文献[9]提出了一种最小超椭球包含的兼类文本分类算法,但该算法得到的超椭球是包含所有样本的最小超椭球,没有考虑噪音点对超椭球构造的影响,从而影响分类精度。为此,本文提出了一种基于超椭球支持向量机的多类文本分类算法。对每一类样本,在特征空间求得一个包含该类尽可能多样本且将噪音点排除的最小超椭球,使各类样本之间通过超椭球球面隔开。对于待分类样本,通过判断其映射到每个超椭球球心的马氏距离确定其类别。

本文第 2 节介绍了超椭球支持向量机;第 3 节详细阐述了基于超椭球支持向量机的兼类文本分类算法;第 4 节给出了在 Reuters21578 标准语料库上的实验结果;最后得出结论。

2 超椭球支持向量机

设给定一类训练样本集$\{x_i\}_{i=1}^N$,其中,$x_i \in R^d$。设 X 是 $d \times N$ 的样本矩阵。在特征空间寻找一个超椭球 $E(a,R)$,其中 a 为球心,R 为半径。超椭球应包围尽可能多的样本,同时半径 R 应尽可能小。当不存在偏远点时,包围所有样本;当存在偏远点时,寻找一个能够包围大多数样本的最小超椭球,

本文受国家自然科学基金(60974071),辽宁省自然科学基金(201202003),辽宁省教育厅重点实验室项目(LS2010180)资助。

秦玉平(1965—),男,博士,教授,主要研究领域为机器学习,E-mail:jzqinyuping@gmail.com;王 祎(1989—),女,硕士生,主要研究领域为机器学习;伦淑娴(1972—),女,博士,教授,主要研究领域为模式识别;王秀坤(1945—),女,教授,博士生导师,主要研究领域为数据库系统。

允许一部分样本在超椭球的外面；当不知道是否含有偏远点时，则通过引入非负松弛变量 $\xi_i(i=1,2,\cdots,N)$，允许一部分样本位于超椭球的外面，采用与寻找最优分类面类似的方法，通过下面的目标函数得到最小超椭球[10-12]。

$$\min_{a,R,\xi_i} R^2+C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t. } (x_i-a)^{\mathrm{T}}\sum^{-1}(x_i-a)\leqslant R^2+\xi_i \quad (1)$$
$$\xi_i\geqslant 0, i=1,2,\cdots,N$$

式中，参数 C 用于折中位于超椭球外部的噪音点的个数和超椭球半径的长度，$\sum$是样本点的协方差矩阵。

为了求解上述优化问题，可以定义如下的 Lagrange 函数：

$$L(R,a,\beta,\gamma,\xi_i)=R^2+C\sum_{i=1}^{N}\xi_i-\sum_{i=1}^{N}\alpha_i\{R^2+\xi_i-(x_i-a)^{\mathrm{T}}\sum^{-1}(x_i-a)\}-\sum_{i=1}^{N}\beta_i\xi_i \quad (2)$$

其中，$\alpha_i\geqslant 0$ 和 $\beta_i\geqslant 0$ 为样本集的 Lagrange 系数。

求解式(2)的最小值，可令该泛函对 R、a 及 ξ_i 求偏导，并令导数等于 0。

$$\frac{\partial L}{\partial R}=2R(1-\sum_{i=1}^{N}\alpha_i)=0\Rightarrow\sum_{i=1}^{N}\alpha_i=1 \quad (3)$$

$$\frac{\partial L}{\partial \xi_i}=C-\alpha_i-\beta_i=0\Rightarrow 0\leqslant\alpha_i\leqslant C \quad (4)$$

$$\frac{\partial L}{\partial a}=-\sum_{i=1}^{N}2\alpha_i(x_i-a)=0\Rightarrow a=\sum_{i=1}^{N}\alpha_i x_i \quad (5)$$

将约束条件式(3)、式(4)及式(5)代入式(2)中，并进行合并整理，得到式(1)的 Lagrangian 对偶为：

$$\max_{\alpha_i\geqslant 0}\sum_{i=1}^{N}\alpha_i x_i^{\mathrm{T}}\sum^{-1}x_i-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j x_i^{\mathrm{T}}\sum^{-1}x_j$$
$$\text{s.t. }\sum_{i=1}^{N}\alpha_i=1 \quad (6)$$
$$0\leqslant\alpha_i\leqslant C, i=1,2,\cdots,N$$

若核函数为 $k(x_i,x_j)=g(x_i)^{\mathrm{T}}g(x_j)$，则式(6)的核形式为：

$$\max_{\alpha_i\geqslant 0}\sum_{i=1}^{N}\alpha_i k(x_i,X)Q^{\mathrm{T}}\Omega^{-2}Qk(x_i,X)^{\mathrm{T}}-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j k(x_i,X)Q^{\mathrm{T}}\Omega^{-2}Qk(x_j,X)^{\mathrm{T}}$$
$$\text{s.t. }\sum_{i=1}^{N}\alpha_i=1 \quad (7)$$
$$0\leqslant\alpha_i\leqslant C, i=1,2,\cdots,N$$

式(7)是一个二次凸优化问题，利用标准的二次优化解法，可以在多项式时间内获得最优解。

最小超椭球球心 a 为带权系数 α_i 的线性加权组合：

$$a=\sum_i\alpha_i g(x_i) \quad (8)$$

样本点 x 到超椭球球心的马氏距离为：

$$d^2(g(x),a)=(g(x)-a)^{\mathrm{T}}\sum^{-1}(g(x)-a)$$
$$=N(k(x,X)-\sum_{i=1}^{N}\alpha_i k(x_i,X))Q^{\mathrm{T}}\Omega^{-2}Q(k(x,X)-\sum_{i=1}^{N}\alpha_i k(x_i,X)) \quad (9)$$

最小超椭球半径 R 根据式(10)，即 KKT 条件求得。

$$\begin{cases} d^2(g(x),a)<R^2, & \alpha_i=0 \\ d^2(g(x),a)=R^2, & 0<\alpha_i<C \\ d^2(g(x),a)>R^2, & \alpha_i=C \end{cases} \quad (10)$$

根据 $d(g(x),a)$可确定样本 x 所属的类别。

3 算法描述

设给定兼类样本集 $A=\{x_i,E_i\}_{i=1}^{l}$ 和核函数 $K(x_i,x_j)$。其中，$x_i\in R^d$，$E_i=\{y_{ij}\}_{j=1}^{p}$，$y_{ij}\in\{1,2,\cdots,N\}$，$N$ 是样本集 A 中含有的总类别数，$p(p\leqslant N)$是样本 x_i 的兼类数，K 对应某特征空间中的内积，即 $K(x_i,x_j)=g(x_i)^{\mathrm{T}}g(x_j)$。

设 A^i 为 A 中属于第 $i(i=1,2,\cdots,N)$类的样本子集。对于每一类样本 A^i，根据超球支持向量机在特征空间寻找一个最小超椭球 $E(a_i,R_i)$。

对待分类样本 x，首先根据式(9)计算它映射到每个超椭球球心的马氏距离 $d_i(\varphi(x),a_i)$ $(i=1,2,\cdots,N)$，然后根据 $d_m(\varphi(x),a_m)$确定待分类样本 x 的类别。

若没有超椭球包围样本 x 的映射 $g(x)$，即对所有的 m $(m=1,2,\cdots,N)$，都有 $d_m(g(x),a_m)>R_m$，则先根据式(11)计算样本 x 属于第 m 类的隶属度，再根据式(12)确定样本 x 所属类别。

$$r_m=\frac{R_m}{d_m(g(x),a_m)} \quad (11)$$

$$r=\max\{r_1,r_2,\cdots,r_N\} \quad (12)$$

待分类样本 x 的分类过程具体描述如下：

步骤 1　根据式(9)计算 $d_m(g(x),a_m)$，$m=1,2,\cdots,N$；

步骤 2　若存在 $d_m(g(x),a_m)\leqslant R_m$，则 x 所属类别为$\{m\mid d_m(x)\leqslant R_m, m=1,2,\cdots,N\}$，转步骤 4，否则转步骤 3；

步骤 3　先根据式(11)计算 x 属于每一类的隶属度 r_m，然后根据式(12)计算隶属度的最大值 r，则 x 所属类别为$\{m\mid r_m=r, m=1,2,\cdots,N\}$，转步骤 4；

步骤 4　分类结束。

4 实验结果及分析

实验使用标准数据集 Reuters 21578，从中选取 6 类且一个文本所属类别最多为 3 类的 665 篇文本进行实验分析。用其中的 431 篇文本作为训练样本，其余的 234 篇文本作为测试样本(见表 1)。将文本数据经过预处理后形成高维词空间向量，采用信息增益的方法进行特征降维，向量中词的权重根据 tf-idf 公式计算。

表 1　训练语料和测试语料

类别	oat	rice	corn	wheat	cotton	soybean
类别标识	1	2	3	4	5	6
训练集	9	44	168	204	44	79
测试集	5	23	84	101	22	40

实验中，对文献[7-9]算法和本文算法分别进行实验分析。使用的核函数为径向基函数(Radial Basis Function, RBF)$K(x,y)=e^{-\gamma\|x-y\|^2}$，其中，参数 $\gamma=0.01$。惩罚参数 $C=100$。

实验环境为 CPU Pentium 1.6G，内存 512M，操作系统 Windows Xp。采用通用的准确率、召回率和 F_1 值作为评价指标。

$$\text{准确率}(P)=N_c/N_a \quad (13)$$

$$\text{召回率}(R)=N_c/N_r \quad (14)$$

$$F_1=(2*P*R)/(P+R) \quad (15)$$

其中，N_c 代表对某个测试样本测试后得到的正确类别数；N_a 代表对某个测试样本测试后得到的类别数；N_r 代表某个测试

样本实际的类别数。

定义 1 平均准确率(AP)=($\sum P$)/n (16)

n若为测试样本总数,则称为宏平均准确率(MAAP);n若为兼类数相同的样本数,则称为微平均准确率(MIAP)。

定义 2 平均召回率(AR)=($\sum R$)/n (17)

n若为测试样本总数,则称为宏平均召回率(MAAR);n若为兼类数相同的样本数,则称为微平均召回率(MIAR)。

定义 3 平均 F_1 值(AF)=($\sum F_1$)/n (18)

n若为测试样本总数,则称为宏平均 F_1 值(MAAF);n若为兼类数相同的样本数,则称为微平均 F_1 值(MIAF)。

表 2 给出了 4 种算法的宏平均准确率、宏平均召回率和宏平均 F_1 值比较。表 3 给出了 4 种算法的微平均准确率、微平均召回率和微平均 F_1 值比较。表 4 给出了 4 种算法训练时间和分类时间比较。

表 2 宏平均准确率、宏平均召回率和宏平均 F_1 值比较

算法	MAAP(%)	MAAR(%)	MAAF(%)
文献[7]算法	60.83	59.24	60.31
文献[8]算法	78.38	77.92	77.52
文献[9]算法	80.88	78.82	79.62
本文算法	82.34	80.96	81.84

表 3 微平均准确率、微平均召回率和微平均 F_1 值比较

算法	兼类数	MIAP(%)	MIAR(%)	MIAF(%)
文献[7]算法	1	65.32	63.76	64.21
	2	57.37	56.33	56.57
	3	55.66	54.57	55.38
文献[8]算法	1	71.34	73.91	72.14
	2	83.33	55.32	63.93
	3	100.00	50.00	65.00
文献[9]算法	1	73.76	76.80	74.71
	2	85.19	59.26	71.67
	3	66.67	66.67	66.67
本文算法	1	75.78	78.95	76.68
	2	87.12	62.33	73.93
	3	66.67	66.67	66.67

表 4 训练时间和测试时间比较

算法	训练时间(ms)	测试时间(ms)
文献[7]算法	393	297
文献[8]算法	221	139
文献[9]算法	276	101
本文算法	259	132

从实验结果可以看出,本文算法的准确率和召回率高于其他 3 种算法。其原因是文献[7]算法存在不可分区域,且没有考虑样本的分布;文献[8]算法仅适用于每类样本都呈超椭球形分布且紧密度较高的情况,当样本呈方向各异的超椭球形分布时,超球体较大,分类精度较低;文献[9]算法构造的超椭球的球心是同类样本的均值,且超椭球包含的是全部样本,当存在噪音点时,超椭球较大,影响分类进度;本文算法采用超椭球支持向量机方法得到超椭球的球心和半径,将噪音点隔离在外,确保超椭球的体积最小,并利用到超椭球球心的马氏距离确定样本的类别,考虑了样本的分布,从而提高了分类精度。本文算法的训练速度高于文献[7]算法和文献[9]算法,与文献[8]算法基本相当。这是因为文献[7]算法需要训练的分类器个数多于其他 3 个算法。文献[9]算法需要进行坐标变换,维数越高,计算量越大,同时还需优化缩放因子。本文算法的分类速度高于文献[7]算法,略低于文献[9]算法,与文献[8]相当。其原因是文献[7]算法的分类过程需要多个 SVM 二分类器,计算较复杂;文献[9]算法使用超椭球公式进行分类,分类时只涉及待分类样本;文献[8]算法使用欧式距离进行分类,本文算法使用马氏距离进行分类,分类时涉及所有支持向量。当样本分布呈超椭球形且存在多个方向各异的噪音点时,本文算法将有明显的优势。

结束语 提出了一种基于超椭球支持向量机的兼类文本类分类算法。利用超椭球对每类数据进行描述,同时将噪音点排除在该超椭球外,缩小了超椭球的体积。同时利用马氏距离进行分类,考虑了样本点的分布信息。因此该算法能有效地对新样本作出正确分类。在标准数据集 Reuters 21578 上的实验结果表明,该方法具有较高的分类精度。进一步的研究工作是利用增量学习的思想提高其在大样本情况下的训练速度。

参 考 文 献

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

[2] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Feature [A] // Proceedings of ECML-98, 10th European Conference on Machine Learning[C]. Berlin: Springer, 1998: 137-142

[3] 孙晋文,肖建国. 基于 SVM 的中文文本分类反馈学习技术的研究[J]. 控制与决策,2004,19(8):927-930

[4] Bennett K P. Combining Support Vector and Mathematical Programming Methods for Classification[A] // Advances in Kernel Methods: Support Vector Learning[C]. Cambridge, MA: MIT press, 1999: 307-326

[5] Krebel U G. Pairwise Classification and Support Vector Machines[A] // Advances in Kernel Methods: Support Vector Learning[C]. Cambridge, MA: MIT press, 1999: 255-268

[6] Platt J C, Cristianini N, Shawe-Taylor J. Large Margin DAGs for multiclass classification[A] // Advances in Neural Information Processing Systems[C]. Cambridge, MA: MIT Press, 2000: 547-553

[7] 王晔,黄上腾. 基于支持向量机的文本兼类标注[J]. 计算机工程与应用,2006,42(2):182-185

[8] 秦玉平,王秀坤,王春立. 基于超球支持向量机的兼类文本分类算法研究[J]. 计算机工程与应用,2008,44(19):166-168

[9] 秦玉平,陈一荻,王春立,等. 一种新的兼类文本分类方法[J]. 计算机科学,2011,38(11):204-205

[10] Wei X K, Huang G B. Mahalanobis Eillpsoidal Learning Machine for One Class Classification[C] // International Conference on Machine Learning and Cybernetics. 2007: 3528-3533

[11] 李永新,薛贞霞. 最大间隔椭球形多类分类算法[J]. 计算机工程,2010,36(7):185-189

[12] 李建民,李永新,薛贞霞. 基于马氏椭球学习机的监督野点探测[J]. 计算机工程与应用,2009,45(13):200-210