文章编号: 1001-0920(2009)01-0137-04

实现兼类样本类增量学习的一种算法

秦玉平[1,2], 王秀坤[1], 王春立[1]
(1. 大连理工大学 电子与信息工程学院, 辽宁 大连 116024; 2. 渤海大学 信息科学与工程学院, 辽宁 锦州 121000)

摘 要: 针对兼类样本,提出一种类增量学习算法. 利用超球支持向量机,对每类样本求得一个能包围该类尽可能多样本的最小超球,使各类样本之间通过超球隔开. 增量学习时,对新增样本以及旧样本集中的支持向量和超球附近的非支持向量进行训练,使得算法在很小的空间代价下实现兼类样本类增量学习. 分类过程中,根据待分类样本到各超球球心的距离判定其所属类别. 实验结果表明,该算法具有较快的训练、分类速度和较高的分类精度.

关键词: 支持向量机; 超球; 兼类; 类增量学习
中图分类号: TP181 **文献标识码**: A

An incremental learning algorithm for multi-class sample

QIN Yu-ping[1,2], *WANG Xiu-kun*[1], *WANG Chun-li*[1]
(1. School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China; 2. College of Information Science and Engineering, Bohai University, Jinzhou 121000, China. Correspondent: QIN Yu-ping, E-mail: jzqinyuping@gmail.com)

Abstract: To multi-class sample, an incremental learning algorithm is proposed in this paper. Hyper-sphere support vector machine is used to get the smallest hyper-sphere that contains most samples of a class, which can divide the class samples from others. In the process of class incremental learning, the new samples, the history support vectors and the history samples that near the hyper-sphere are trained. Therefore, the multi-class incremental learning can be realized in a small memory space. For the sample to be classified, the distances from it to the centre of every hyper-sphere are used to confirm the classes that the sample belongs to. The experimental results show that the algorithm has a higher performance on training speed, classification speed, and classification precision.

Key words: Support vector machines; Hyper-sphere; Multi-class; Class incremental learning

1 引 言

支持向量机(SVM)[1]是一种基于统计学习理论的机器学习方法,具有很好的泛化能力. 支持向量机的泛化性能并不依赖于全部训练数据,而是该全部数据的一个子集,即支持向量,并且支持向量数与整个训练数据集数相比是很小的,因此支持向量机对于增量学习是一种强有力的工具.

支持向量机增量学习的主要研究成果有 Batch SVM 增量学习算法[2,3]、训练精确解算法[4]、分块增量算法[5]、快速增量学习算法[6]、α ISVM 算法[7]和多分类增量学习算法[8]等,这些方法都不涉及类别的增加. 文献[9]给出了一种类增量学习算法,将新增类样本作为正类,原有类样本作为负类,训练得到的二值分类器作为二叉树的根结点. 该算法节省了训练时间,取得了一定的效果,但只适用于单类别样本的类增量学习.

兼类是样本的一个属性,即一个样本可能属于几个类别,其类增量学习问题尚未得到研究. 本文提出一种基于超球支持向量机的兼类样本类增量学习算法 MCIL (Multi-class incremental learning),快速有效地实现了兼类样本的类增量学习.

本文首先介绍了超球支持向量机理论;然后阐述了基于超球的兼类样本类增量学习算法,并给出了在 Reuters 21578 标准语料库上的实验结果;最后得出结论.

2 超球支持向量机

给定一类训练样本集$\{x_i\}_{i=1}^{l}$和核函数 $K(x_i, x_j)$,寻找特征空间的一个超球(a, R). 其中: $x_i \in$

收稿日期: 2007-10-08; **修回日期**: 2008-02-01.
基金项目: 国家自然科学基金项目(60603023); 国家 973 计划项目(2001CCA00700).
作者简介: 秦玉平(1965—),男,辽宁建平人,教授,博士生,从事机器学习和决策支持系统的研究; 王秀坤(1945—),女,辽宁辽阳人,教授,博士生导师,从事数据库系统和决策支持系统的研究.

R^n, K 对应某特征空间 Z 中的内积,即 $K(x_i, x_j) = \langle g(x_i), g(x_j)\rangle$,变换 $g: X \to Z$ 是将样本从输入空间映射到特征空间; a 为球心, R 为球半径. 超球应尽量包围样本的大部分映射,同时半径 R 应尽可能地小. 当不存在偏远的点时,则寻找一个能包围所有样本映射的最小超球;当存在偏远的点,允许一部分样本映射在超球的外面时,则寻找一个能包围大多数样本映射的最小超球;当不知道是否含有偏远的点时,则通过引入一个非负松弛变量 ξ_i, $i = 1, 2, \cdots, l$, 允许一部分样本映射位于超球的外面. 采用与寻找最优分类面类似的方法,通过对下面的目标函数的最小化得到最小超球[10,11]:

$$\min F(R, a, \xi_i) = R^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i. \tag{1}$$

$$\text{s.t.} \quad \| g(x_i) - a \|^2 \le R^2 + \xi_i, \quad i = 1, 2, \cdots, l; \tag{2}$$

$$\xi_i \ge 0, \quad i = 1, 2, \cdots, l. \tag{3}$$

其中 $0 < v \le 1$ 用来控制超球的半径与它所能包围的样本数之间的折衷. v 越小,惩罚越大,对于允许在超球外面存在样本的约束程度也就越大.

为了求解上述优化问题,可以定义如下的 Lagrange 函数:

$$L(R, a, \beta, \gamma, \xi_i) = R^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\beta_i\{R^2 + \xi_i - \| g(x_i) - a \|^2\} - \sum_{i=1}^{l}\gamma_i\xi_i, \tag{4}$$

其中 $\beta_i \ge 0$ 为该类样本集的 Lagrange 系数.

求解式(4)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,得

$$\frac{\partial L}{\partial R} = 2R\left(1 - \sum_{i=1}^{l}\beta_i\right) = 0 \Rightarrow \sum_{i=1}^{l}\beta_i = 1; \tag{5}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{vl} - \beta_i - \gamma_i = 0 \Rightarrow 0 \le \beta_i \le \frac{1}{vl}, \quad i = 1, 2, \cdots, l; \tag{6}$$

$$\frac{\partial L}{\partial a} = -\sum_{i=1}^{l}2\beta_i(g(x_i) - a) = 0 \Rightarrow a = \sum_{i=1}^{l}\beta_i g(x_i), \quad i = 1, 2, \cdots, l. \tag{7}$$

将约束条件(5) ~ (7)代入(4),并进行合并整理,得

$$\max L(\beta) = L(R, a, \beta, \gamma, \xi_i) = \sum_{i=1}^{l}\beta_i K(x_i, x_i) - \sum_{i,j=1}^{l}\beta_i\beta_j K(x_i, x_j). \tag{8}$$

$$\text{s.t.} \quad \sum_{i=1}^{l}\beta_i = 1; \tag{9}$$

$$0 \le \beta_i \le \frac{1}{vl}, \quad i = 1, 2, \cdots, l. \tag{10}$$

由式(7)可知,最小超球中心为带权系数 β_i 的线性加权组合,即

$$a = \sum_{i=1}^{l}\beta_i g(x_i). \tag{11}$$

当 $\beta_i > 0$ 时,对应的样本称为支持向量;当 $0 < \beta_i < \frac{1}{vl}$ 时,对应的样本位于超球附近,任选其中一个该类样本 x 与超球球心之间的距离可确定超球半径

$$R^2 = \| g(x) - a \|^2 = K(x, x) - 2\sum_{i=1}^{l}\beta_i K(x, x_i) + \sum_{i,j=1}^{l}\beta_i\beta_j K(x_i, x_j); \tag{12}$$

当 $\beta_i = \frac{1}{vl}$ 时,对应的样本位于超球外面,称为野值或含噪声的样本.

3 兼类样本类增量学习算法

给定初始兼类样本集 $A = \{x_i, E_i\}_{i=1}^{l}$ 和核函数 $K(x_i, x_j)$. 其中: $x_i \in R^n$, $E_i = \{y_{ij}\}_{j=1}^{p}$, $y_{ij} \in \{1, 2, \cdots, N\}$, N 为样本集 A 中含有的总类别数, $p(p \le N)$ 为样本 x_i 的兼类数; K 对应某特征空间 Z 中的内积,即 $K(x_i, x_j) = \langle g(x_i), g(x_j)\rangle$,变换 $g: X \mid\to Z$ 是将样本从输入空间映射到特征空间.

设 A^m 为 A 中兼有类别 m 的样本子集, $m = 1, 2, \cdots, N$. 对于每一类样本集 A^m,利用超球支持向量机在特征空间确定一个超球 (a_m, R_m). 其中: a_m 为该类超球的球心, R_m 为该超球的半径. SV_m 为该超球的支持向量, NSV_m 为该超球附近的非支持向量.

设新增兼类样本集为 B, B^q 为 B 中兼有类别 q 的样本子集, $q = 1, 2, \cdots, N, N+1, \cdots, M$. 类增量学习算法描述如下:

Step1: 对于每个样本子集 B^q, $q = N+1, \cdots, M$, 在特征空间训练一个超球 (a_q, R_q),保存其支持向量 SV_q 和超球附近的非支持向量 NSV_q;

Step2: 对于每个样本子集 B^q, $q = 1, 2, \cdots, N$, 若 $B^q \ne \Phi$,则 $B^q = B^q + \mathrm{SV}_q + \mathrm{NSV}_q$,在特征空间重新训练超球 (a_q, R_q),更新 SV_q 和 NSV_q.

对于待分类样本 x,根据

$$[d_m(x)]^2 = \| g(x) - a_m \|^2 = \| g(x) - \sum_i \beta_i^m g(x_i^m) \|^2 = K(x, x) + \sum_{i,j}^{l}\beta_i^m\beta_j^m K(x_i^m, x_j^m) - 2\sum_{i=1}^{l}\beta_i^m K(x, x_i^m) \tag{13}$$

计算它到各超球球心 a_m 的距离 $d_m(x)$, $m = 1,2,\cdots,N$. 根据 $d_m(x)$ 的值判断 x 所属的类别. 若 $d_m(x) > R_m, m = 1,2,\cdots,N$,则根据

$$r_m = R_m / d_m(x) \tag{14}$$

计算样本 x 属于第 m 类的隶属度,根据

$$r = \max_m r_m \tag{15}$$

确定 x 所属的类别.

待分类样本 x 的分类过程描述如下:

Step1: 根据式(13) 计算 $d_m(x)$, $m = 1,2,\cdots,N$;

Step2: 若存在 $d_m(x) \leq R_m$,则 x 所属类别为 $\{m \mid d_m(x) \leq R_m, m = 1,2,\cdots,N\}$,转 Step4;否则转 Step3;

Step3: 先根据式(14) 计算 r_m,然后根据式(15) 计算 r, x 所属类别为 $\{m \mid r_m \geq r, m = 1,2,\cdots,N\}$,转 Step4;

Step4: 分类结束.

4　算法性能分析

由算法的学习过程可知,每次训练只针对一类样本,且历史数据中只有支持向量和超球附近的非支持向量参加,因此该算法适用于处理规模较大的兼类数据集. 另外,每次增量学习只对新增样本集含有的类重新训练,因此每次增量学习需要训练的超球的个数最多为 N(样本集的类别数),并且约束条件简单,同时有效地保留了与新增样本无关类的历史训练结果,易于推广和改进. 分类过程中,通过 N 次简单的距离计算,便可确定样本所属类别,有效地实现了兼类样本的分类,而且分类速度较快,准确率较高. 该算法对类别较多、兼类数较少的大规模数据集更有效.

5　实验结果及分析

本文使用标准数据集 Reuters 21578,从中选取 6 类且一个文本所属类别最多为 3 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). 将文本数据进行预处理,形成高维词空间向量,采用信息增益的方法进行特征降维,向量中每个词的权重根据 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
训练集规模	9	44	168	204	44	79
测试集规模	5	23	84	101	22	40

实验中采用通用的平均准确率(AP),平均召回率(AR) 和平均 F_1 值(AF) 作为评价指标.

准确率

$$(P) = N_c / N_a; \tag{16}$$

召回率

$$(R) = M_c / N_r, \tag{17}$$

$$F_1 = \frac{2 * P * R}{P + R}. \tag{18}$$

其中: N_c 代表对每个测试样本测试后得到的正确兼类数; N_a 代表对每个测试样本测试后得到的所有兼类数; N_r 代表每个测试样本的实际兼类数.

定义 1　平均准确率

$$(\mathrm{AP}) = \frac{\sum P}{n}. \tag{19}$$

若 n 为测试样本总数,则称为宏平均准确率(MAAP);若 n 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

$$(\mathrm{AR}) = \frac{\sum R}{n}. \tag{20}$$

若 n 为测试样本总数,则称为宏平均召回率(MAAR);若 n 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

$$(\mathrm{AF}) = \frac{\sum F_1}{n}. \tag{21}$$

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

实验环境为 CPU Pentium 1.6 G,内存 512 M,操作系统 Windows Xp. 使用的核函数为径向基函数(RBF) $K(x,y) = \mathrm{e}^{-\gamma \|x-y\|^2}$,其中 $\gamma = 0.01$,系统参数 $v = 0.6$. 算法实现参考了 Chang 和 Lin 所开发的 libsvm[12],并在此基础上进行了相应的修改.

实验中,初始样本集中含有 3 类(第 1 类为 oat,第 2 类为 rice,第 3 类为 corn) 兼类样本. 进行 3 次增量学习,每次新增加的兼类样本都兼有同一个新类别. 第 1 次增加兼有第 4 类(wheat) 的兼类样本,第 2 次增加兼有第 5 类(cotton) 的兼类样本,第 3 次增加兼有第 6 类(soybean) 的兼类样本. 表 2 为 MCIL 算法在初始数据集和每次增量学习后的微平均准确率、微平均召回率和微平均 F_1 值;表 3 为 MCIL 算法在初始数据集和每次增量学习后的宏平均准确率、宏平均召回率和宏平均 F_1 值;表 4 为 MCIL 算法在初始数据集和每次增量后的训练时间以及分类时间.

从实验结果可以看出,MCIL 算法有效地实现了兼类样本的类增量学习,在保证单类样本分类精度的同时,实现了对兼类样本的分类,并具有较好的

表 2 MCIL 算法的微平均准确率、微平均召回率和微平均 F_1 值

学习过程(样本数)	兼类数为 1			兼类数为 2			兼类数为 3		
	MIAP	MIAR	MIAF	MIAP	MIAR	MIAF	MIAP	MIAR	MIAF
初始样本集(209)	91.00	93.00	91.67	100	58.33	72.22	…	…	…
第 1 次增量(165)	80.47	82.25	81.07	80.84	50.00	62.28	100	50.00	65.00
第 2 次增量(26)	77.32	78.69	77.78	83.33	50.00	60.72	100	50.00	65.00
第 3 次增量(31)	71.34	73.91	72.14	83.33	55.32	63.93	100	50.00	65.50

表 3 MCIL 的宏平均准确率、宏平均召回率和宏平均 F_1

学习过程	MAAP	MAAR	MAAF
初始样本集	91.51	91.04	90.57
第 1 次增量	81.32	78.68	79.02
第 2 次增量	79.21	78.24	78.77
第 3 次增量	78.38	77.92	77.52

表 4 MCIL 的训练时间和分类时间 ms

学习过程	训练时间	分类时间
初始样本集	110	58
第 1 次增量	95	114
第 2 次增量	18	122
第 3 次增量	17	139

准确率、召回率和 F_1 值. 另外,随着增量学习的不断进行,该算法能保持较快的训练、分类速度和较高的分类进度,扩展能力强,更适合于单类样本的类增量学习.

6 结 论

本文提出了一种基于超球支持向量机的兼类样本类增量学习算法,有效地解决了兼类样本类增量学习和兼类样本分类问题,训练、分类速度快,分类精度高. 对于类别数较多、样本兼类数较少的大规模数据集,效果更加明显. 实验结果表明,本文算法是一种较为实用的兼类样本类增量学习方法.

参考文献(References)

[1] Vapnik V. The nature of statistical learning theory[M]. New York: Springer, 1995.

[2] Syed N, Liu H, Sung K. Incremental learning with support vector machines[C]. Proc of the Workshop on Support Vector Machines at the Int J Conf on Artifical Intelligence. Stockholm, 1999: 352-356.

[3] Domeniconi C, Gunopulos D. Incremental support vector machine construction[C]. Proc of IEEE Int Conf on Data Mining. San Jose, 2001: 589-592.

[4] Cauwenberghs G, Poggio T. Incremental and decremental support vector machine [J]. Machine Learning, 2001, 44(13): 409-415.

[5] Zhang Jinpei, Li Zhongwei, Yang Jing. A divisional incremental training algorithm of support vector machine [C]. Proc of the IEEE Int Conf on Mechatronics and Automation. Niagara Falls, 2005, 8: 853-855.

[6] 孔锐, 张冰. 一种快速支持向量机增量学习算法[J]. 控制与决策, 2005, 20(10): 1129-1132.
(Kong R, Zhang B. A fast incremental learning algorithm for support vector machine[J]. Control and Decision, 2005, 20(10): 1129-1132.)

[7] 萧嵘, 王继成, 孙正兴, 等. 一种 SVM 增量学习算法 α ISVM[J]. 软件学报, 2001, 12(12): 1818-1824.
(Xiao R, Wang J C, Sun Z X, et al. An incremental SVM learning algorithm α ISVM[J]. J of Software, 2001, 12(12): 1818-1824.)

[8] 朱美琳, 杨佩. 基于支持向量机的多分类增量学习算法[J]. 计算机工程, 2006, 32(17): 77-79.
(Zhu M L, Yang P. Multi-class incremental learning based on support vector machines [J]. Computer Engineering, 2006, 32(17): 77-79.)

[9] Zhang Bofeng, Su Jinshu, Xu Xin. A class-incremental learning method for multi-class support vector machines in text classification[C]. Proc of the 5th Int Conf on Machine Learning and Cybernetics. Dalian, 2006: 13-16.

[10] 张翔, 肖小玲, 徐光祐. 基于样本之间紧密度的模糊支持向量机方法[J]. 软件学报, 2006, 17(5): 951-958.
(Zhang X, Xiao X L, Xu G Y. Fuzzy support vector machine based on affinity among samples [J]. J of Software, 2006, 17(5): 951-958.)

[11] 唐发明, 王仲东, 陈绵云. 支持向量机多类分类算法研究[J]. 控制与决策, 2005, 20(7): 746-749.
(Tang F M, Wang Z D, Chen M Y. On multiclass classification methods for support vector machines[J]. Control and Decision, 2005, 20(7): 746-749.)

[12] Chang Chinchang, Lin Chihjen. LIBSVM: A library for support vector machines [J/OL]. http://www.csie.ntu.tw/cjlin/libsvm, 2005.