一种新的兼类样本类增量学习算法

秦玉平[1] 伦淑娴[1] 王秀坤[2]

(渤海大学工学院 锦州 121000)[1] (大连理工大学计算机科学与技术学院 大连 116024)[2]

摘 要 提出了一种基于超椭球的兼类样本类增量学习算法。对兼有同一类别的样本,在特征空间构建一个能包围该类尽可能多样本的最小超椭球,使各类样本之间通过超椭球球面分开。增量学习过程中,对新增样本中的每一新类别构建超椭球,对新增样本中的各历史类别重新构建超椭球,使得算法在很小的空间代价下实现了兼类样本类增量学习,同时保留了与新增样本类别无关的历史类训练结果。分类过程中,根据待分类样本是否在超椭球内或隶属度来确定其所属类别。实验结果表明,该算法较超球方法具有较快的分类速度和较高的分类精度。

关键词 超椭球,兼类,增量学习,隶属度

中图法分类号 TP181 **文献标识码** A

New Multi-label Sample Class Incremental Learning Algorithm

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

(College of Engineering, Bohai University, Jinzhou 121000, China)[1]

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)[2]

Abstract To multi-label sample, a class incremental learning algorithm based on hyper ellipsoidals was proposed. For every class, the smallest hyper ellipsoidal that contains most samples of the class was structured, which can divide the class samples from others. In the process of class incremental learning, the hyper ellipsoidals of new class were structured, and the historical hyper ellipsoidal that its class exists in the incremental samples was structured again. The multi-label class incremental learning is realized in a small memory space, and the history results that has nothing to do with the new sample classes are saved at the same time. For the sample to be classified, its class is confirmed by the hyper ellipsoidal that it belongs to or its membership. The experimental results show that the algorithm has a higher performance on classification speed and classification precision compared with hyper sphere algorithm.

Keywords Hyper ellipsoidals, Multi-label, Incremental learning, Membership

1 引言

增量学习是一种智能化数据挖掘与知识发现技术,主要研究成果有主成分分析[1]、最近邻方法[2]、Boosting算法[3]和支持向量机[4-8]等。但这些研究成果大都是针对历史类别的样本增量学习。文献[9,10]分别提出了一种类增量学习算法,但这两种方法都是针对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,对其类增量学习问题尚未得到较深入的研究。文献[11]提出了一种兼类样本类增量学习算法,其通过在特征空间求得最优超球面把各类样本最大限度地进行分离,在剔除噪音点的同时实现分类。但该算法只适合于每类样本呈球形分布且聚类程度较高的情况。在实际中,样本往往呈带状的、凸的且各向异性的超椭球型分布。为此,本文提出了一种基于超椭球的兼类样本类增量学习算法。对每一类训练样本,在特征空间求得一个包围该类样本的最小超椭球,使得各类样本之间通过超椭球面隔开。与超球方法相比,该方法缩小了包围的空间,同时减少了分类器中参加计算的样本数量,提高了分类精度和分类速度。

本文第2节介绍了超椭球模型;第3节详细阐述了基于超椭球的兼类样本类增量学习算法;第4节给出了在Reuters 21578标准语料库上的实验结果;最后得出结论。

2 超椭球模型构建

设给定一类训练样本集$\{X_i\}_{i=1}^{l}$,其中,$X_i=(x_{i1},x_{i2},\cdots,x_{in})$,$l$是样本数量。首先计算$l$个样本点的均值,得到超椭球球心坐标$M=(m_1,m_2,\cdots,m_n)$,然后根据式(1)计算样本点$X_i(i=1,2,\cdots,l)$相对于超椭球球心的坐标$U_i$。

$$U_i=(u_{i1},u_{i2},\cdots,u_{in})=X_i-M=(x_{i1},x_{i2},\cdots,x_{in})-(m_1,m_2,\cdots,m_n) \quad (1)$$

如此变换,就把坐标系原点移到了超椭球的球心。为了变换各坐标轴方向,用坐标值$U_i(i=1,2,\cdots,l)$组成$l\times n$阶

到稿日期:2011-10-10 返修日期:2012-03-18 本文受国家自然科学基金项目(60974071),辽宁省教育厅重点实验室项目(LS2010180),辽宁省教育厅优秀人才项目(201102005)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;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。

矩阵 Y，求矩阵 Y 的内积，令 $V=\frac{1}{l}(Y^{\mathrm{T}}\cdot Y)$，$V$ 为 $n\times n$ 阶矩阵，其特征值给出了 n 个互相垂直方向分量的平方 $E_\alpha^2(\alpha=u_{i1},u_{i2},\cdots,u_{in})$，它与超椭球体的 n 个半轴长度 $(a_1,a_2,\cdots,a_n)$ 成比例关系。与特征值对应的 n 个特征向量构成旋转矩阵 R，根据式(2)计算旋转后的坐标值 $Z_i=(z_{i1},z_{i2},\cdots,z_{in})$。

$$Z_i=(z_{i1},z_{i2},\cdots,z_{in})=X_i\cdot R \tag{2}$$

对每个样本点进行上述操作后，实现了超椭球球心与坐标原点的重合，超椭球的 n 个轴与 n 个坐标轴的重合。

根据式(3)计算超椭球的 n 个半轴长度 $(a_1,a_2,\cdots,a_n)$ 为

$$(a_1,a_2,\cdots,a_n)=S(E_{i1},E_{i2},\cdots,E_{in}) \tag{3}$$

式中，$S(S>0)$ 是缩放因子[12-14]。

为寻找包含所有样本的最小超椭球，需要解决如下优化问题

$$\min S$$

$$\text{s. t.}\left\|\left(\frac{z_{i1}}{a_1},\frac{z_{i2}}{a_2},\cdots,\frac{z_{in}}{a_n}\right)\right\|^2<1,i=1,2,\cdots,l \tag{4}$$

把式(2)代入式(4)中，得到

$$\min S$$

$$\text{s. t.}\left\|\left(\frac{z_{i1}}{E_{i1}},\frac{z_{i2}}{E_{i2}},\cdots,\frac{z_{in}}{E_{in}}\right)\right\|<S,i=1,2,\cdots,l \tag{5}$$

求解优化问题式(5)，得到缩放因子 S。

3 兼类样本类增量学习算法

给定初始兼类样本集 $A=\{x_i,E_i\}_{i=1}^l$，其中，$x_i\in R^n$，$E_i=\{y_{ij}\}_{j=1}^p$，$y_{ij}\in\{1,2,\cdots,N\}$，$N$ 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。

设 A^m 为 A 中兼有类别 $m(m=1,2,\cdots,N)$ 的样本子集。对于每一个样本集 A^m，构建一个超椭球 $E(a_m,r_m)$，其中，$a_m=(a_{m1},a_{m2},\cdots,a_{mn})$ 是该类超椭球的球心，$r_m=(r_{m1},r_{m2},\cdots,r_{mn})$ 为该超椭球的半轴。该超椭球的缩放因子为 s_m。

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

步骤 1　对每个样本子集 $B^q(q=N+1,\cdots,M)$，构建超椭球 $E(a_q,r_q)$，其缩放因子为 s_q；

步骤 2　对样本子集 $B^q(q=1,2,\cdots,N)$，若 $B^q\neq\Phi$，则 $B^q=B^q+A^q$，重新构建超球 $E(a_q,r_q)$，更新其缩放因子 s_q。

对待分类样本 x，首先根据式(1)计算 x 相对于第 $i(i=1,2,\cdots,M)$ 个超椭球球心的坐标 $U_i(x)$，然后根据式(2)计算 x 在第 $i(i=1,2,\cdots,M)$ 个超椭球所在坐标系中的坐标 $Z_i(x)=(z_{i1},z_{i2},\cdots,z_{in})$，再根据式(6)计算判别式 $D_i(x)$ 的值，最后根据 $D_i(x)$ 判定 x 所属的类别。

$$D_i(x)=\frac{z_{i1}^2}{a_{i1}^2}+\frac{z_{i2}^2}{a_{i2}^2}+\cdots+\frac{z_{in}^2}{a_{in}^2}(i=1,2,\cdots,M) \tag{6}$$

若对所有的超椭球 $E(a_i,r_i)(i=1,2,\cdots,M)$，都有 $D_i(x)>1$，首先根据式(7)计算使 x 落在第 $i(i=1,2,\cdots,M)$ 个超椭球球面上的缩放因子 S_i^*，然后根据式(8)计算待分类样本 x 属于第 i 类的隶属度，最后根据式(9)确定待分类样本 x 所属类别。

$$S_i^*=\left\|\left(\frac{z_{i1}}{E_{i1}},\frac{z_{i2}}{E_{i2}},\cdots,\frac{z_{in}}{E_{in}}\right)\right\|(i=1,2,\cdots,M) \tag{7}$$

$$r_i=\frac{S_i}{S_i^*} \tag{8}$$

$$r=\max r_i \tag{9}$$

分类过程具体描述如下：

步骤 1　根据式(1)计算 x 相对于超椭球 $E(a_i,r_i)(i=1,2,\cdots,M)$ 球心的坐标 $U_i(x)$；

步骤 2　根据式(2)计算 x 在第 i 个超椭球所在坐标系中的坐标 $Z_i(x)(i=1,2,\cdots,M)$；

步骤 3　根据式(6)计算判别式 $D_i(x)(i=1,2,\cdots,M)$；

步骤 4　若存在超椭球 $E(a_i,r_i)$，使得 $D_i(x)\leqslant 1$，则 x 所属类别为 $\{i|D_i(x)\leqslant 1,i=1,2,\cdots,M\}$，转步骤 6；否则转步骤 5；

步骤 5　对每个类别 $i(i=1,2,\cdots,M)$，首先根据式(7)计算使 x 落在第 i$(i=1,2,\cdots,M)$ 个超椭球球面上的缩放因子 S_i^*，然后根据式(8)计算待分类样本 x 属于第 i 类的隶属度 r_i，再根据式(9)计算最大隶属度 r。待分类样本 x 所属类别为 $\{i|r_i=r,i=1,2,\cdots,M\}$；

步骤 6　分类结束。

4 实验结果及分析

本文使用标准数据集 Reuters 21578，从中选取 6 类($N=6$)且一个文本兼类数最多为 3($p\leqslant 3$)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)。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，向量中每个词的权重根据 tf-idf 公式计算。

表 1　训练集和测试集

类别	oat	rice	corn	wheat	cotton	soybean
类别标识	1	2	3	4	5	6
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，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。进行三增量学习，每次新增加的兼类样本都兼有同一个新类别。第一次增加兼有第 4 类(wheat)的兼类样本，第二次增加兼有第 5 类(cotton)的兼类样本，第三次增加兼有第 6 类(soybean)的兼类样本。

采用通用的准确率、召回率和 F_1 值作为评价指标。

$$\text{准确率}(P)=N_c/N_a \tag{10}$$

$$\text{召回率}(R)=N_c/N_r \tag{11}$$

$$F_1=(2*P*R)/(P+R) \tag{12}$$

式中，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

$$(AP)=(\sum P)/n \tag{13}$$

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

定义 2　平均召回率

$$(AR)=(\sum R)/n \tag{14}$$

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

定义 3　平均 F_1 值

$$(AF)=(\sum F_1)/n \tag{15}$$

(下转第 224 页)

结束语 从实际的道路运输安全管理决策支持需求出发，在对现有的基础和应用条件分析的基础上，提出了道路运输安全管理 MDAS 的总体解决方案，包括系统的全局网络拓扑结构和系统体系结构，并深入地研究了其中涉及的关键技术。以此技术为基础，开发完成了重庆市道路运输安全管理 MDAS，并将其应用于重庆市的道路运输安全管理决策支持工作中。实际运行表明：设计的系统能够实现时间、空间、车辆、人员、企业多种维度的营运车辆超速报警综合分析，能够满足实际决策支持工作的需要。

参 考 文 献

[1] Recker W W. The California Advanced Transportation Management Systems Testbed [C]// WESCON'93. Conference Record, 1993: 273-278

[2] Ossowski S, Fernandez A, Serrano J M. Designing Multi-agent Decision Support System-The Case of Transportation Management [C]// Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-agent Systems. 2004: 1470-1471

[3] 孙棣华，张星霞，张志良. 地图匹配算法及其在智能交通系统中的应用[J]. 计算机工程与应用，2005，41(20)：225-228

[4] Liu Wen-yuan, Fang Shu-fen. OLAP Realization technology Research Based on MDX [C]// Proceedings International Conference on. Machine Learning and Cybernetics. 2002, 4: 2205-2209

(上接第 207 页)

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

实验中，对超球和超椭球两种方法进行实验分析。实验环境为 Pentium 1.6G，内存 512M，操作系统 Windows XP。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 $\gamma=0.01$，系统参数 $v=0.6$。

表 2 给出了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 3 给出了在初始数据集和每次增量后的训练时间以及分类时间。

表 2 两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

学习过程	算法	MAAP	MAAR	MAAF
初始样本集	超球算法	91.51	91.04	90.57
	超椭球算法	92.16	91.78	91.93
第 1 次增量	超球算法	81.32	78.68	79.02
	超椭球算法	83.25	80.06	81.33
第 2 次增量	超球算法	79.21	78.24	78.77
	超椭球算法	82.75	79.48	80.99
第 3 次增量	超球算法	78.38	77.92	77.52
	超椭球算法	80.88	78.82	79.62

表 3 两种算法的训练时间和分类时间比较

学习过程	算法	训练时间(ms)	分类时间(ms)
初始样本集	超球算法	110	58
	超椭球算法	127	46
第 1 次增量	超球算法	94	114
	超椭球算法	112	77
第 2 次增量	超球算法	16	122
	超椭球算法	24	81
第 3 次增量	超球算法	15	139
	超椭球算法	21	101

从实验结果可以看出，超椭球方法的准确率和召回率高于超球方法。其主要原因是样本在特征空间不是规整地呈超球型分布，而是呈带状的、凸的且各向异性的超立方体或超椭球型分布，用超椭球包围的空间小于用超球包围的空间，从而提高了其分类精度。超椭球方法较超球方法提高了分类速度，其主要原因是每次分类时，超椭球方法的分类器中只涉及一个样本(待分类样本)，而超球方法的分类器涉及多个样本(所有支持向量)。但超椭球方法的训练速度比超球方法略慢，这是因为增量学习时，新增样本以及新增样本中兼有的历史类样本都参加训练，同时需要进行坐标变换，其维数越高，计算量越大。另外，优化缩放因子也要花费一定的时间。

结束语 本文提出了一种基于超椭球的兼类样本类增量学习算法，描述了最小包围椭球的构造和相应的类增量学习算法，并与超球方法作了比较。在标准数据集 Reuters 21578 上的实验结果表明，该方法在分类精度和分类速度上都优于超球方法。进一步的研究工作是引入核函数理论，并在规模较大、兼类数较多的数据集上进行试验。

参 考 文 献

[1] Li Y. On incremental and robust subspace learning[J]. Pattern Recognition, 2004, 37(7): 1509-1518

[2] Lam W, Keung C K, Liu D. Discovering useful concept prototypes for classification based on filtering and abstraction[J]. IEEE Transtrations on Pattern Analysis and Maehine Intelligence, 2002, 24(8): 1075-1090

[3] Gangardiwala A, Polikar R. Dynanncally weighted majority voting for incremental learning and comparison of three boosting based approache[C]// Proeeedings of the IEEE International Joint Conference on Neural Networks. Montreal, Canada, 2005: 1131-1136

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

[5] Zhang J P, Li Z W, Yang J. A divisional incremental training algorithm of support vector machine[C]// Proceeding of the IEEE International Conference on Mechatronics and Automation. Niagara Falls, Canada, 2005: 853-855

[6] Diehl C, Cauwenberghs G. SVM incremental learning, adaptation and optimization[C]// Proceedings of the International Joint Conference on Neural Networks. 2003: 2685-2690

[7] 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

[8] 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

[9] Zhang B F, Su J S, Xu X. A class-incremental learning method for multi-class support vector machines in text classification[C]// Proceedings of the Fifth International Conference on Machine Learning and Cybernetics. Dalian, China, 2006: 13-16

[10] 秦玉平，李祥纳，王秀坤，等. 基于超球支持向量机的类增量学习算法研究[J]. 计算机科学，2008，35(18)：116-118

[11] 秦玉平，王秀坤，王春立. 实现兼类样本类增量学习的一种算法[J]. 控制与决策，2009，24(1)：137-140

[12] 高俊祥，杜海清，刘勇. 采用光照不变特征的椭球法运动阴影检测[J]. 北京邮电大学学报，2009，32(5)：109-113

[13] Shigeo A, Ruck T. A fuzzy classifier with ellipsoidal regions[J]. IEEE Transactions on Fuzzy Systems, 1997, 5(3): 358-368

[14] 刘勇，赵斌，夏绍玮. 模糊超椭球分类算法及其在无约束手写体数字识别中的应用[J]. 清华大学学报，2000，40(9)：120-124