一种新的兼类文本分类方法

秦玉平[1] 陈一荻[1] 王春立[2] 王秀坤[3]

(渤海大学工学院 锦州 121000)[1] (大连海事大学信息科学技术学院 大连 116026)[2]

(大连理工大学计算机科学与技术学院 大连 116024)[3]

摘 要 提出了一种基于超椭球的兼类文本分类算法。对每一类样本,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隔开。对待分类样本,通过判断其是否在超椭球内确定其类别。若没有超椭球包围待分类样本,则通过隶属度确定其所属类别。在标准数据集 Reuters 21578 上的实验结果表明,该方法较超球方法提高了分类精度和分类速度。

关键词 超椭球,兼类分类,缩放因子,隶属度

中图法分类号 TP181 **文献标识码** A

New Multi-label Text Classification Algorithm

QIN Yu-ping[1] CHEN Yi-di[1] WANG Chun-li[2] WANG Xiu-kun[3]

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

(College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China)[2]

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)[3]

Abstract A new multi-label text classification algorithm based on hyper ellipsoidal was proposed in this paper. For every class, the smallest hyper ellipsoidal that contains the samples of the class is structured, which can divide the class samples from others. For the sample to be classified, its class is confirmed by the hyper ellipsoidal that surrounds it. If the sample is not surrounded by any hyper ellipsoidal, the membership is used to confirmed its class. The experiments were done on Reuters 21578 and the experiment results show that the algorithm has a higher performance on classification speed and classification precision compare with hyper sphere algorithm.

Keywords Hyper ellipsoidal, Multi-label classification, Extension factor, Membership

1 引言

文本分类能大大缩短查询时间,保证搜索效果,改善文本信息杂乱的状况。因此,文本分类技术已成为机器学习领域的一个研究热点。目前,文本分类的主要研究成果有 k 最近邻算法[1]、朴素贝叶斯算法[2-4]、决策树算法[5]和支持向量机[6,7]等。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,对兼类分类问题尚未进行较深入的研究。文献[8]提出了一种基于 DAGSVM 的兼类文本分类算法,但该方法存在不可分区域,并且分类速度较慢。文献[9]提出了一种基于超球支持向量机的兼类文本分类算法,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,在剔除噪音点的同时实现兼类文本快速分类,但该算法只适合于每类样本程球形分布且聚类程度较高的情况。现实中的样本分布往往是带状的、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文本分类算法。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隔开。对待分类样本,通过判断其所属的超椭球确定其类别。

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,l 是样本数,$X_i=(x_{i1},x_{i2},\cdots x_{in})$。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:

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

这样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 $E_\alpha^2(\alpha=u_{i1},u_{i2},\cdots,u_{in})$,$E_\alpha$ 与超椭球体的 n 个半轴长度$(a_1,a_2,\cdots,a_n)$呈比例关系。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

到稿日期:2010-12-22 返修日期:2011-03-21 本文受国家自然科学基金项目(60603023),国家基础研究重大项目(973)研究专项(2001CCA00700),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;陈一荻(1985—),女,硕士生,主要研究领域为机器学习;王春立(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。

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重合，超椭球的 n 个轴与 n 个坐标轴重合。

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

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

式中，$S(S>0)$是缩放因子[10-12]。

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：

$$\min S$$

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

把式(2)代入式(4)中，得到：

$$\min S$$

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

求解优化问题(5)得到缩放因子 S。

3 算法描述

设给定兼类样本集 $A=\{x_i,E_i\}_{i=1}^{l}$，其中，l 是样本数，$x_i\in R^n$，$E_i=\{y_{ij}\}_{j=1}^{p}$，$y_{ij}\in\{1,2,\cdots,N\}$，$N$ 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^i 为 A 中属于第 i 类的样本子集，其中，$i=1,2,\cdots,N$。对于每一类样本 A^i，根据超椭球构造算法在特征空间寻找一个超椭球(m_i,a_i)，其中，$m_i=(m_{i1},m_{i2},\cdots,m_{in})$ 是该超椭球的球心，$a_i=(a_{i1},a_{i2},\cdots,a_{in})$为该超椭球的半轴长度。$n$ 个互相垂直方向分量为$(E_{i1},E_{i2},\cdots,E_{in})$，缩放因子为 S_i。

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

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

若对所有的超椭球$(m_i,a_i)(i=1,2,\cdots,N)$，都有 $D_i(x)>1$，则首先根据式(7)计算使 x 落在第 $i(i=1,2,\cdots,N)$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。

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

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

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

对待分类样本 x，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$(m_i,a_i)(i=1,2,\cdots,N)$球心的坐标 $U_i=(u_{i1},u_{i2},\cdots,u_{in})$；

步骤 2　根据式(2)计算 x 经坐标变换后的坐标 $Z_i=(z_{i1},z_{i2},\cdots,z_{in})(i=1,2,\cdots,N)$；

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

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

步骤 5　对每个类别 $i(i=1,2,\cdots,N)$，首先根据式(7)计算使 x 落在第 $i(i=1,2,\cdots,N)$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$\{i|r_i=r,i=1,2,\cdots,N\}$；

步骤 6　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
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，对超球和超椭球两种方法进行实验分析。超球方法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 $v=0.6$。超椭球方法的缩放因子见表 2。

表 2　超椭球方法的缩放因子

类别	oat	rice	corn	wheat	cotton	soybean
缩放因子	0.63	0.70	0.55	0.71	0.70	0.69

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

$$准确率(P)=N_c/N_a \tag{10}$$

$$召回率(R)=N_c/N_r \tag{11}$$

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

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$(AP)=(\sum P)/n$ (13)

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

定义 2　平均召回率$(AR)=(\sum R)/n$ (14)

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

定义 3　平均 F_1 值$(AF)=(\sum F_1)/n$ (15)

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

表 3 给出了超球算法和超椭球算法的宏平均准确率、宏平均召回率和宏平均 F_1 值的比较。表 4 给出了超球算法和超椭球算法的微平均准确率、微平均召回率和微平均 F_1 值

(下转第 224 页)

出版社，2008
[3] Malcolm B. Reducts within the variable precision rough sets model: A further investigation [J]. European Journal of Operational Research, 2001, 134: 592-605
[4] Greco S, Matarazzo B, Slowingski R. Rough approximation of a preference relation by dominance relation [J]. European Journal of Operation Research, 1991, 117: 63-83
[5] Greco S, Matarazzo B, Slowinski R. A new rough set approach to evaluation of bankruptcy risk[M]. Zopounidis C. Ed. Operational Tool in the Management of Financial Risk. 1998: 121-136
[6] Greco S, Matarazzo B, Slowinski R. Dominance-based rough set approach to rating analysis[M]. Fuzzy Economic Review, 1999
[7] Greco S, Matarazzo B, Slowinski R. A new rough set approach to multicriteria and multiattribute classication[M]// Polkowski L, Skowron A. eds. Rough Sets and Current Trends in Computing, 1998: 60-67
[8] Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis [J]. European Journal of Operational Research, 2001, 129: 1-47
[9] 张文修，仇国芳. 基于粗糙集的不确定决策[M]. 北京：清华大学出版社，2005
[10] 张文修，姚一豫，梁怡. 粗糙集与概念格[M]. 西安：西安交通大学出版社，2006
[11] 徐伟华，张文修. 基于优势关系下不协调目标信息系统的知识约简[J]. 计算机科学，2006，33(2)：182-184
[12] 袁修久，何华灿. 优势关系下的相容约简和下近似约简[J]. 西北工业大学学报，2006，24(5)：604-608
[13] 徐伟华，张晓燕，张文修. 优势关系下不协调目标信息系统的下近似约简[J]. 计算机工程与应用，2009，45(16)：66- 68
[14] 徐伟华，张晓燕，张文修. 优势关系下不协调目标信息系统的上近似约简[J]. 计算机工程，2009，35(18)：191-193
[15] 徐伟华，张文修. 基于优势关系下不协调目标信息系统的分布约简[J]. 模糊系统与数学，2007，21(4)：124-131
[16] 桂现才，彭宏. 优势关系下分布约简和最大分布约简问题研究[J]. 计算机工程与应用，2009，45(2)：150-153
[17] LI Yan, Sun Na-xin, Zhao Jin. Reductions in inconsistent decision systems based on dominance relations[C]// Proceedings of 2010 International Conference on Machine Learning and Cybernetics(ICMLC). 2010: 279-284
[18] LI Yan, Zhao Jin, Sun Na-xin, et al. Generalized Distribution Reduction in Inconsistent Decision Systems Based on Dominance Relations[C]//Proceedings of 2010 International Conference on Rough Sets and Knowledge Technology(RSKT). LNAI 6401, 2010: 151-158
[19] 陈娟，王国胤，胡军. 优势关系下不协调信息系统的正域约简[J]. 计算机科学，2008，35(3)：216-218

(上接第 205 页)

的比较。表 5 给出了超球算法和超椭球算法算法训练时间和分类时间的比较。

表 3 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(%)
超球算法	78.38	77.92	77.52
超椭球算法	80.88	78.82	79.62

表 4 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(%)
超球算法	1	71.34	73.91	72.14
	2	83.33	55.32	63.93
	3	100.00	50.00	65.00
超椭球算法	1	73.76	76.80	74.71
	2	85.19	59.26	71.67
	3	66.67	66.67	66.67

表 5 训练时间和测试时间比较

算法	训练时间(ms)	测试时间(ms)
超球算法	221	139
超椭球算法	276	101

从实验结果可以看出，超椭球方法的准确率和召回率都高于超球方法。这是因为样本在特征空间往往是带状的、凸的且各向异性的超椭球型分布，用超椭球包围的空间小于用超球包围的空间，从而提高了分类精度。超椭球方法较超球方法提高了分类速度，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)。超椭球方法的训练速度比超球方法略慢，这是因为构造超椭球时需要进行坐标变换，维数越高，计算量越大，同时还需优化缩放因子。

结束语 本文提出了一种基于超椭球的兼类文本类分类算法，描述了最小包围椭球的构造以及相应兼类文本分类算法，并与超球方法做了比较。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，通过样本映射增加样本之间的紧密度，缩小超椭球包围空间，进一步提高分类精度。

参 考 文 献

[1] Yang Yi-ming. An Evaluation of Statistical Approaches to Text Categorization[J]. Journal of Information Retrieval, 1999(1): 69-90
[2] Mccallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification[C]// AAAI Workshop on Learning for Text Categorization. Madison, 1998: 509-516
[3] Han J, Kamber M. Data Mining: Concepts and Techniques[M]. Beijing: Higher Education Press, 2001
[4] 林士敏，田凤占，陆玉昌. 贝叶斯学习、贝叶斯网络与数据采掘[J]. 计算机科学，2000，27(10)：69-72
[5] Takahashi F, Abe S. Decision-Tree-Based Multiclass Support Vector Machines[C]// International Conference on Neural Information Processing. Singapore, 2002: 1418-1422
[6] Platt J, Cristianini N, Shawe-Taylor J. Large Margin DAGs for Multiclass Classification[C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000: 547-553
[7] Tax D, Duin R. Uniform Object Generation for Optimizing One-class Classifiers[J]. Journal of Machine Learning Research, 2001(2): 155-173
[8] 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
[9] 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
[10] 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
[11] Shigeo A, Ruck T. A Fuzzy Classifier with Ellipsoidal Regions [J]. IEEE Transactions on Fuzzy Systems, 1997, 5(3): 358-368
[12] 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