

基于超椭球的多类文本分类算法研究

秦玉平¹ 陈一荻¹ 王春立² 王秀坤³

(渤海大学信息科学与工程学院 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 提出一种基于超椭球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围该类尽可能多样本的最小超椭球,使得各类样本之间通过超椭球隔开。对待分类样本,通过判断其是否被超椭球包围来确定类别。实验结果表明,与超球方法相比,该方法具有较高的分类精度和分类速度。

关键词 超椭球,多类分类,缩放因子

中图法分类号 TP181 **文献标识码** A

Study on Multiclass Text Classification Algorithm Based on Hyper Ellipsoidal

QIN Yu-ping¹ CHEN Yi-di¹ WANG Chun-li² WANG Xiu-kun³

(College of Information Science and Technology, 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 multiclass classification algorithm based on hyper ellipsoidal 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. For the sample to be classified, its class is confirmed by the hyper ellipsoidal that it belong to. The experiment results show that the algorithm has a higher performance on classification speed and classification precision, compared with hyper sphere algorithm.

Keywords Hyper ellipsoidal, Multiclass classification, Extension factor

1 引言

随着计算机技术和通讯技术的快速发展,文本自动分类技术越来越受到人们的关注。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]、支持向量机^[6,7]等。由于支持向量机具有出色的泛化性能,已成为解决文本分类问题的主要工具。

支持向量机是针对二分类问题提出的,在处理多类问题时,人们往往将其分解成一系列的二分类问题加以解决。常用的处理方法主要有1-a-1^[8]、1-a-r^[9]和DAGSVM^[6]。但这些算法存在不可分区域,且当类别较多时分类效率较低。文献[10]提出了一种超球支持向量机多类分类算法,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分离,在剔除噪音点的同时实现分类,但该算法只适用于每类样本成球形分布且聚类程度较高的情况。实际上,样本往往是带状的、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多类文本分类算法。对每一类训练样本,在特征空间求得一个包围该类尽可能多样本的最小超椭球,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来确定类别,有效地实现了多类文本分类。

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 $X_i=(x_{i1}, x_{i2}, \dots, x_{in})$ 。首先计算 l 个样本点的均值,得到超椭球球心坐标 $M=(m_1, m_2, \dots, m_n)$,然后根据式(1)计算样本点 $X_i(i=1, 2, \dots, l)$ 相对于超椭球球心的坐标 $U_i(i=1, 2, \dots, l)$ 。

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

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 $U_i(i=1, 2, \dots, l)$ 组成 $l \times n$ 阶矩阵 Y ,求矩阵 Y 的内积,令 $V=\frac{1}{l}(Y^T \cdot Y)$, 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 $E_a^2(a=u_{i1}, u_{i2}, \dots, u_{in})$, E_a 与超椭球体的 n 个半轴长度 $(a_1, a_2, \dots, a_n)$ 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 $Z_i=(z_{i1}, z_{i2}, \dots, z_{in})(i=1, 2, \dots,$

到稿日期:2010-09-09 返修日期:2010-11-12 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和决策支持系统,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。

n)。

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

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^[11-13]。

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

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

式中, $S(S>0)$ 是缩放因子。

为寻找包含尽可能多样本的最小超椭球,解决如下优化问题:

$$\begin{aligned} \min S \\ \text{s. t. } \left\| \left(\frac{z_{i1}}{a_1}, \frac{z_{i2}}{a_2}, \dots, \frac{z_{in}}{a_n} \right) \right\|^2 < 1 \end{aligned} \quad (4)$$

把式(2)代入式(4)中,得到:

$$\begin{aligned} \min S \\ \text{s. t. } \left\| \left(\frac{z_{i1}}{E_{i1}}, \frac{z_{i2}}{E_{i2}}, \dots, \frac{z_{in}}{E_{in}} \right) \right\| < S \end{aligned} \quad (5)$$

求解优化问题(5)得到缩放因子 S 。

3 分类算法描述

给定多类样本集 $A = \{x_i, E_i\}_{i=1}^N$, 其中 $x_i \in R^n$, $E_i \in \{1, 2, \dots, N\}$, N 是样本集 A 中的类别数。

设 A^i 为 A 中属于第 i 类的样本子集, 其中 $i=1, 2, \dots, N$ 。对于每一类样本 A^i , 在特征空间寻找一个最小超椭球 (m_i, a_i) , 其中 $m_i = (m_{i1}, m_{i2}, \dots, m_{in})$ 是该超椭球的球心, $a_i = (a_{i1}, a_{i2}, \dots, a_{in})$ 是该超椭球的半轴长度。通过变换使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, n 个互相垂直方向分量为 $(E_{i1}, E_{i2}, \dots, E_{in})$, 缩放因子为 S_i 。

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 相对于第 i ($i=1, 2, \dots, N$) 个超椭球球心的坐标 $U_i = (u_{i1}, u_{i2}, \dots, u_{in})$;

步骤 2 根据式(2)计算 x 经坐标系旋转后在第 i 个坐标系中的坐标 $Z_i = (z_{i1}, z_{i2}, \dots, z_{in})$;

步骤 3 根据式(6)计算判别式 D_i 的值;

$$D_i = \frac{z_{i1}^2}{a_{i1}^2} + \frac{z_{i2}^2}{a_{i2}^2} + \dots + \frac{z_{in}^2}{a_{in}^2} \quad (6)$$

步骤 4 如果满足 $D_i \leq 1$ ($i=1, 2, \dots, N$) 的 i 唯一, 则判定样本 x 为第 i 类, 转步骤 6; 如果满足 $D_i \leq 1$ ($i=1, 2, \dots, N$) 的 i 不存在, 转步骤 5; 否则先舍去 $D_i > 1$ 的 i , 然后转步骤 5;

步骤 5 对每个 i , 首先根据式(7)计算通过缩放使 x 落在第 i 个超椭球球面上的缩放因子 S_{i*} , 然后根据式(8)计算 x 所属类别 r , 转步骤 6;

$$S_{i*} = \left\| \left(\frac{z_{i1}}{E_{i1}}, \frac{z_{i2}}{E_{i2}}, \dots, \frac{z_{in}}{E_{in}} \right) \right\| \quad (7)$$

$$r = \arg \min_i \frac{S_{i*}}{S_i} \quad (8)$$

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 5 类共 896 篇文本, 用其中的 598 篇作为训练样本, 其余的 298 篇作为测试样本(见表 1)。将文本数据预处理后形成高维词空间向量, 采用信息增益的方法来进行特征降维, 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	wheat	corn	coffee	soybean	cocoa
训练集规模	204	168	97	79	50
测试集规模	101	84	48	40	25

实验中, 使用超球和超椭球两种方法进行实验分析。其中, 超球方法使用线性核函数, 系统参数 $C=10$, $v=0.01$; 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	wheat	corn	coffee	soybean	cocoa
缩放因子	0.78	0.56	0.80	0.65	0.73

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。

$$\text{准确率} = \frac{\text{分类正确的样本数}}{\text{实际分类的样本数}} \quad (9)$$

$$\text{召回率} = \frac{\text{分类正确的样本数}}{\text{应有的样本数}} \quad (10)$$

$$F_1 = \frac{\text{准确率} \times \text{召回率} \times 2}{\text{准确率} + \text{召回率}} \quad (11)$$

式中, 每一类的准确率、召回率和 F_1 称为微平均; 所有类的准确率、召回率和 F_1 称为宏平均。

表 3 给出了两种方法的微平均 F_1 值比较。表 4 给出了两种方法的宏平均准确率、宏平均召回率和宏平均 F_1 比较。表 5 给出了两种方法的训练时间和分类时间比较。

表 3 超椭球方法与超球方法的微平均 F_1 比较

类别	wheat	corn	coffee	soybean	cocoa
超球算法	73.25	82.38	76.45	72.34	75.21
超椭球算法	73.47	88.65	78.21	78.97	75.29

表 4 超椭球方法与超球方法的宏平均准确率、宏平均召回率和宏平均 F_1 比较

算法	宏平均准确率(%)	宏平均召回率(%)	宏平均 F_1 (%)
超球算法	76.64	73.47	75.02
超椭球算法	81.41	76.61	78.40

表 5 超椭球方法与超球方法的训练时间和分类时间比较

算法	训练时间(ms)	分类时间(ms)
超球算法	188	204
超椭球算法	243	185

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。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, 同时需优化缩放因子。

结束语 本文提出了一种基于超椭球的多类分类算法, 描述了最小包围椭球的构造和相应的分类算法, 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, 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, 陆玉吕. 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[M]. 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] Bennett K P. Combining Support Vector and Mathematical Programming Methods for Classification[M]. Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT Press, 1999: 307-326
- [9] Krebel U G. Pairwise Classification and Support Vector Machines [M]. Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT Press, 1999: 255-268
- [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(8): 116-118
- [11] 高俊祥, 杜海清, 刘勇. 采用光照不变特征的椭圆法运动阴影检测[J]. 北京邮电大学学报, 2009, 32(5): 109-113
- [12] Shigeo A, Ruck T. A Fuzzy Classifier with Ellipsoidal Regions [J]. IEEE Transactions on Fuzzy Systems, 1997, 5(3): 358-368.
- [13] 刘勇, 赵斌, 夏绍玮. 模糊超椭圆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

(上接第 207 页)

方面都还刚刚起步。在研究对象上,多集中在以影评为代表的体验商品。在研究方法上,多使用单一的定量研究(实证、建模、文本挖掘等)方法。本文提出的评论有用性影响因素模型还有待于进一步完善和检验,综合使用多种研究方法分析不同类型商品的评论有用性是本领域今后的研究方向。

参考文献

- [1] 郝媛媛,叶强,李一军. 基于影评数据的在线评论有用性影响因素研究[J]. 管理科学学报, 2010, 13(8): 78-96
- [2] Bei L T, Chen E Y I, Widdows R. Consumers' Online Information Search Behavior and the Phenomenon of Search and Experience Products[J]. Journal of Family and Economic Issues, 2004, 25(4): 449-467
- [3] Brown J J, Reingen P H. Socialties and word-of-mouth referral behavior[J]. Journal of Consumer Research, 1987, 14(3): 350-362
- [4] Chaiken S, Eagly A H. Communication modality as a determinant of message persuasiveness and message comprehensibility [J]. Journal of Personality and Social Psychology, 1976, 34(4): 606-614
- [5] Clemons E K, Gao G D, Hitt L M. When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry[J]. Journal of Management Information Systems, 2006, 23(2): 149-171
- [6] Cook D L, Coupey E. Consumer Behavior and Unresolved Regulatory Issues in Electronic Marketing[J]. Journal of Business Research, 1998, 41(3): 231-238
- [7] Eisend M. Two-Sided Advertising: A Meta-Analysis[J]. International Journal of Research in Marketing, 2006, 23(2): 187-198
- [8] Forman C, Ghose A, Wiesenfeld B. Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets [J]. Information Systems Research, 2008, 19(3): 291-313
- [9] Ghose A, Ipeirotis P. Designing Ranking Systems for Consumer Reviews: The Impact of Review Subjectivity on Product Sales and Review Quality[C]// Proceedings of the 16th Annual Workshop on Information Technology and Systems
- [10] Hennig-Thurau T, Gwinner K P, Walsh G, et al. Electronic Word-of-mouth Via Consumer-opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet? [J]. Journal of Interactive Marketing, 2004, 18(1): 38-52
- [11] Hovland C I. Social Communication[J]. Proceedings of the American Philosophical Society, 1948, 92(5): 371-375
- [12] Hu N, Pavlou P A, Zhang J. Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication[C]// Proceedings of the 7th ACM conference on Electronic commerce, 2006: 324-330
- [13] Huang P, Lurie N H, Mitra S. Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods[J]. Journal of Marketing, 2009, 73(2): 55-69
- [14] Jiang Z, Benbasat I. Investigating the Influence of the Functional Mechanisms of Online Product Presentations [J]. Information Systems Research, 2007, 18(4): 221-244
- [15] Kempf D S, Smith R E. Consumer Processing of Product Trial and the Influence of Prior Advertising: A structural Modeling Approach[J]. Journal of Marketing Research, 1998, 35(3): 325-337
- [16] Klein L. Evaluating the Potential of Interactive Media Through a New Lens: Search Versus Experience Goods[J]. Journal of Business Research, 1998, 41(3): 195-203
- [17] Kumar N, Benbasat I. The Influence of Recommendations on Consumer Reviews on Evaluations of Websites[J]. Information Systems Research, 2006, 17(4): 425-439
- [18] Lazonder A W, Harm J A B, Iwan G J H W. Differences Between Novice and Experienced Users in Searching Information on the World Wide Web[J]. Journal of the American Society for Information Science, 2000, 51(6): 576-581
- [19] Mudambi S M, Schuff D. What makes a helpful online review? A study of customer reviews on amazon. com[J]. MIS Quarterly, 2010, 34(1): 185-200
- [20] Nelson P. Information and Consumer Behavior [J]. Journal of Political Economy, 1970, 78(20): 311-329
- [21] Nelson P. Advertising as Information[J]. Journal of Political Economy, 1974, 81(4): 729-754
- [22] Park D H, Lee J, Han I. The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement [J]. International Journal of Electronic Commerce, 2007, 11(4): 125-148
- [23] Short J, Williams E, Christie B. The Social Psychology of Telecommunication [M]. New York: John Wiley & Sons, 1976
- [24] Weathers D, Sharma S, Wood S L. Effects of Online Communication Practices on Consumer Perceptions of Performance Uncertainty for Search and Experience Goods[J]. Journal of Retailing, 2007, 83(4): 393-401
- [24] 李文鸿, 胡喜卉, 孙劭文. 基于对跨国公司商品销售预测的数学模型[J]. 重庆理工大学学报: 自然科学版, 2011, 24(7): 117-119