doi:10.3772/j.issn.2095-915x.2018.03.009

基于密度分布单类支持向量机的科技文献分类研究

董微　赵捷

中国科学技术信息研究所　北京　100038

摘 要 在 OCSVM 单分类问题上，科技文献自动分类时交叉学科的分类并未得到良好的解决，且支持向量构造的超平面未考虑到非支持向量的影响，本文提出了一种基于密度分类的单类支持向量机的分类算法，将支持向量的密度分布引入目标函数。实验结果表明，该算法能够较好的将交叉学科的科技文献进行主题分类。

关键词：科技文献；文本分类，密度分布；单类支持向量机
中图分类号：G250.7
开放科学（资源服务）标识码（OSID）

Scientific Literature Classification Research Based on the Density Distribution of OCSVM

DONG Wei　ZHAO Jie

Institute of Scientific and Technical Information of China, Beijing 100038, China

Abstract The problem of the scientific literature automatic classification on cross subject in the One-Class SVMs is still not solved, and the small portion of samples called support vectors fully decide the hyperplane, whereas all the nonsupport vectors have no influence on the hyperplane. This study proposed a classification based on the density of one class support vector machines classification algorithm. Density distribution of support vectors is introduced into the objective function. The experimental results show that the proposed algorithm can classify the interdisciplinary scientific literature subject better.

Keywords: Scientific Literature; text classification; density distribution; OCSVM

基金项目：科技部科技创新战略研究专项“基础研究领域年度重点创新进展报告”（ZLY201636）；NSTL专项基金项目“开放学术资源建设”（2016XM16）。

作者简介：董微（1987–），博士，研究方向：数据挖掘、网络爬取，E-mail：dongw@istic.ac.cn；赵捷（1959–），高级工程师，研究方向：信息组织与构建、知识服务与知识链接、网络信息系统建设。

doi:10.3772/j.issn.2095-915x.2018.03.009

1 引言

随着科技的进步、信息量的增加，推动着图书馆数字化、网络化的全面发展，许多过去以印刷形式发行的报纸期刊也纷纷将自己的刊物搬到了因特网上。网络化、数字化时代的文献信息组织与揭示，为多元化的用户提供了直接和便捷的文献信息查找途径。目前，最常见的科技文献数据库检索方法是输入主题关键词作为检索条件，但检出的科技文献往往数以百计，用户需要在此基础上做进一步的筛选。因此快速的帮助用户获取相关的信息、缩小信息搜索范围、降低信息容量，使得研究如何实现电子科技文献面向主题的自动分类，为用户提供个性化服务具有重要的意义。

在对科技文献的主题进行自动分类过程中，往往只考虑到单分类，即一种文献往往只被归到一种主题。随着学科种类的增多以及交叉学科的出现，在分类的过程中，主题之间的界定往往存在交叉情况，因此本文针对科技文献主题的自动化分类进行研究，考虑了交叉学科分类困难的情况，提出了一种基于密度分布的单类支持向量机的分类算法，提高科技文献主题分类的准确性。

2 相关工作

为了有效利用数量日益增加的科技文献，迫切需要对它们的主题进行合理高效的分类。机器学习是解决这一问题的一条有效途径。常用的科技文献分类方法有：Naive Bayes分类法[1-2]，决策树分类法[3-4]，Rocchio分类器[5-7]，SVM[8-9]以及KNN[8,10-12]等分类法。例如，白小明等人[13]着重对SVM和KNN的科技文献自动分类算法进行分析。刘继才[14]设计实现了基于向量空间模型和KNN与SVM分类算法的科技文献自动分类系统。陈鑫卿等人[15]结合科技文献之间的互相引证关系、KNN分类法与贝叶斯分类法提出一种协调的科技文献分类方法。田萱等人[16]提出一种基于人工神经网络的文本表示模型并在此基础利用K-means算法对科技文献进行聚类分析。周丽红等人[17]通过关联规则的分类方法将科技文献划分为不同的类型。林堃等人[18]改进了基于关联规则的文本分类方法ARC-BC，利用ARC-BC分类器的封闭测试的结果对分类器进行调整规则置信度。姚力群等人[19]结合了局部线性和One-Class的思想对科技文本分类问题进行了研究，利用局部线性的思想寻找文本样本的内在支撑流形，利用One-Class的思想确定正负样本的分界面。王娟等人[20]提出了基于惩罚性矩阵分解(PMD)的文本软聚类算法。

上述的分类发现方法基本上是将科技文献分类到一种主题，鲜少涉及交叉学科的分类。本文针对科技文献自动分类时存在交叉学科的问题，提出了一种基于密度分类的单类支持向量机的分类算法，通过计算支持向量与非支持向量的概率密度来控制松驰变量，并将高斯函数引入目标函数。实验结果表明，该算法能够较好的将交叉学科的科技文献进行主题分类。

3 分类算法

在OCSVM单分类问题上，主要考虑了单学

科的分类问题，对交叉学科的分类并未得到良好的解决。利用OCSVM算法进行分类时，主要确定分类超平面，判断学科是否满足超平面，若满足则该文献被划分为指定的主题。然而，这不仅未考虑到交叉学科，且支持向量构造的超平面未考虑到非支持向量的影响，降低了分类的准确性。本文提出了一种基于密度分布的单类支持向量机的分类算法，将支持向量的密度分布引入目标函数。

3.1 样本库构建

首先，本文构建样本库，选取τ个主题的样本，其中M_i表示第$i(i=1,2\ldots,\tau)$个主题，如下图1所示。每个主题包含的样本是与本主题相关的。

构建训练集时，选取其中一个主题的样本做为正样本集，其它领域的样本做为负样本集。存在交叉学科的样本时，该样本需要作为共同的正样本集，并从负样本集的集合中剔除。

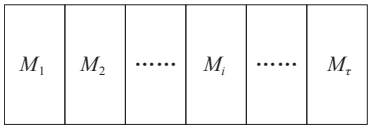

图1 样本库构建

3.2 支持向量的密度

Lee等人[21]提出了一种为空间每个点计算相关密度分布的方法：

$$\delta_i^{'}=\exp(\frac{M^K}{d(x_i,x_i^K)}) \quad \text{（公式1）}$$

其中，x_i^K是x_i的正样本集中第K个最近邻，$d(x_i,x_i^K)$表示为x_i和x_i^K的距离。M^K表示所有正样本集中的K最近邻的平均距离，即$M^K=\frac{1}{m}\sum_{i=1}^{m}d(x_i,x_i^K)$，$m$表示所有正样本集中的样本个数。

$\delta_i^{''}$表示正样本集中每个样本的外部概率密度。其中，x_i表示正样本集中的样例，x_i^L表示负样本集中的样例，x_i^L是x_i的第L个最近邻负样本集中的数据，并且N^L表示所有负样本集中的L最近邻的平均距离。

$$\delta_i^{''}=\exp(\frac{N^L}{d(x_i,x_i^L)}) \quad \text{（公式2）}$$

根据公式1与公式2，每个点δ_i的密度分布:

$$\delta_i=\exp(\frac{M^K}{d(x_i,x_i^K)}+\frac{N^L}{d(x_i,x_i^L)}) \quad \text{（公式3）}$$

3.3 基于密度分布的OCSVM算法

在OC-SVM算法中，超平面由少数的支持向量完全决定，而非支持向量对超平面完全没有影响，导致获取的超平面往往是次优解[22-23]。本文将上述的支持向量的密度分布概念引入本问题中。设有数据样本的样本数为n，通过核函数ϕ映射到高维特征空间，最优超平面与坐标原点最大距离为$\frac{\rho}{\|\omega\|}$，与分类超平面的距离为$\frac{\xi_i}{\|\omega\|}$，其中ρ为超平面的法向量，ω为超平面的截距。根据公式3，本文引入支持向量的密度分布δ_i，将上述问题转化为：

$$\min\frac{1}{2}\|\omega\|^2+\frac{1}{vn}\sum_i^n\delta_i\xi_i-\rho \quad \text{（公式4）}$$

约束条件为：

$$\omega\cdot\phi(x_i)\geq\rho-\xi_i$$
$$\xi_i\geq 0 \quad \text{（公式5）}$$

其中，δ_i引入公式3的正测试集的密度分布，$\gamma\in(0,1]$用来控制支持向量在训练样本中所占的比重，并且松弛变量$\xi=[\xi_1,\xi_2,\ldots,\xi_n]^T$被引入目标函数，将上述问题转化为对偶问题：

doi:10.3772/j.issn.2095-915x.2018.03.009

$$\min \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(x_i,x_j) \qquad (公式6)$$
$$s.t. \quad 0 \le \alpha_i \le \frac{\delta_i}{\nu n}, \sum_{i=1}^{n}\alpha_i = 1$$

在OC-SVM中，$\rho = \sum_{i=1}^{n}\alpha_i K(x,x_i)$为确定的阈值，与超平面的法向量$\omega$决定分离超平面。

引入高斯RBF核函数，OC-SVM分类函数转化为：

$$f(x) = sgn(\sum_{i=1}^{n}\alpha_i K(x,x_i) - \rho) \qquad (公式7)$$

根据公式6求出分类超平面，因此，当对测试集进行计算时，根据所求得每种主题的分类超平面，对测试集中的数据分别计算，若某测试数据分别满足多个分类超平面，则说明该测试数据为交叉学科的科技文献。

4 实验结果

本文从知网上根据主题词分别获取计算机、航天、医学、农业、生物学5个领域的文献共3000余篇，其中随机选取每个类别的70%作为训练集合，剩余的30%用于测试，并在训练集中随机选取λ%的正例文档和剩余的其它文档集合作为U集合，剩余的（1–λ%）的正例文档作为P集合，通过1-DNF算法获取初始反例集合，并通过迭代使用SVM算法构造文本分类器。本项目基于LibSvm[24]工具包，在数据预处理过程中，采用ICTCLAS分词工具[25]。

本文选取的λ范围为10~50，步长为10。根据公式8，计算1-DNF中反例比例λ对算法*ERR*%的影响。实验结果如图1所示。

$$ERR\% = \frac{\text{可信反例中包含正例的个数}}{\text{未标识集合中掺入的正例个数}} \qquad (公式7)$$

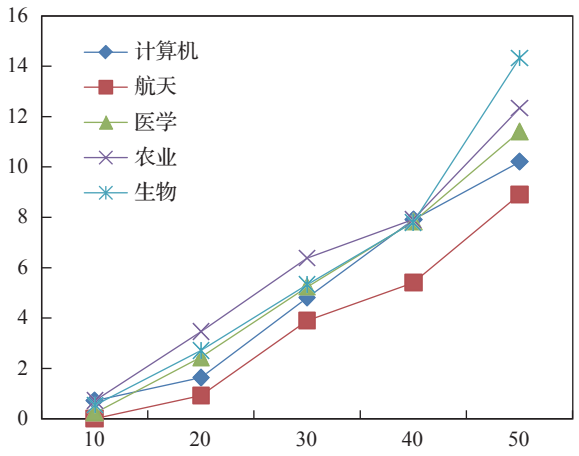

图1 可信反例随λ%的变化情况

由上图可以看出，随着λ%的增大，在五个领域*ERR*%均随之增大，因此，本文下述实验构建的文本分类器采用λ%为10%。

本文在计算概率密度时，K与L最近邻选取为10。根据在训练上计算的分类平面，在测试集上进行测试，将改进的OCSVM与传统的OCSVM算法进行比较，实验结果如图2和图3所示。

首先从图中观察改进的OCSVM算法的实验结果。从图2的精确率可以看出，当λ%取为10%时，计算机类与航天类主题的精确率较高，而医学、农业、生物学类相对较低；经过分析发现，医学类、农业类均与生物学类存在交叉学科，而计算机与航天学科的交叉主题的文章较少。从图3所示的召回率可以看出，5个领域的召回率均较高，其中计算机类与航天类分类较为全面；医学类、农业类与生物学类，生物学召回率较高，而医学类召回率较低，经过分析发现，部分医学类被判定为生物学。

通过分析与比较改进的OCSVM与传统的OCSVM算法，可以发现，改进的OCSVM的算

法相较于传统的OCSVM算法，大大的提高了精确率；并且在交叉学科方面，召回率也有很大改进。

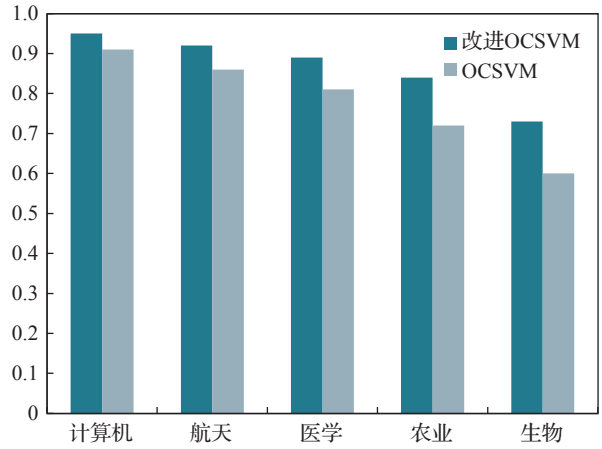

图2 5个领域精确率

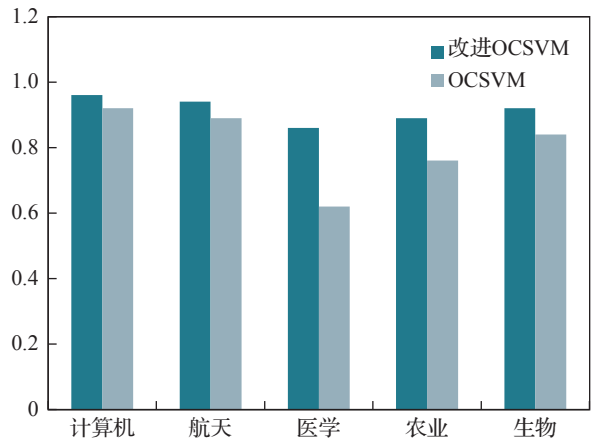

图3 5个领域召回率

5 结束语

随着图书馆数字化、网络化的推进，实现电子科技文献面向主题的自动分类，对Web资源开发与利用、为用户提供个性化服务有很重要的意义。本文针对交叉学科主题分类困难的问题，提出了一种基于密度分布的单类支持向量机的分类算法，通过实验证明，该算法能够较好的将交叉学科的科技文献进行主题分类。通过自动化对科技文献进行主题分类，可以大大提高用户检索的准确率，减少用户在繁杂的信息中检索有用信息的困难度，进一步推进图书馆的数字化发展。随着新兴事物的不断发展，越来越多新鲜术语的出现，进一步影响着主题的分类，作者将在下一步工作中致力于自适应的主题分类。

参考文献

[1] 马佳, 李小平. 基于贝叶斯网的数字图书馆个性化文献推荐系统的研究[J]. 科技视界, 2012(26):38-39.
[2] 王东, 熊世桓, 向程冠,等. 基于频繁2-项集的贝叶斯分类器[J]. 兰州理工大学学报, 2013, 39(4):99-104.
[3] 黄华. 基于DT和SVM算法的科技文献分类研究[C]. “决策论坛——区域发展与公共政策研究学术研讨会”论文集(上). 2016.
[4] 黄华. 基于决策树与SVM融合学习的科技文献分类方法研究[D]. 郑州: 河南工业大学, 2011.
[5] 邱定, 张激, 王金华,等. 基于Rocchio和KNN提出的新的文本分类技术[J]. 自动化与仪器仪表, 2017(8):107-110.
[6] 曾砺锋. 基于Rocchio方法和k均值聚类的支持向量机文本分类方法[J]. 软件导刊, 2008(6):37-39.
[7] 施聪莺, 徐朝军, 杨晓江. 基于规则和Rocchio分类器的学前综合教育资源分类[J]. 现代图书情报技术, 2009, 25(7):75-79.
[8] 陈玉芹. 多类别科技文献自动分类系统[D]. 武汉: 华中科技大学, 2008.
[9] 王方, 阮梅花, 朱海刚,等. 基于向量空间模型的科技文献自动分类研究[J]. 情报探索, 2013(12):1-3.
[10] 鲍文, 胡清华, 于达仁. 基于K-近邻方法的科技文献分类[J]. 情报学报, 2003, 22(4):451-456.
[11] 钱慎一, 朱艳玲, 朱颢东. 基于多层挖掘策略的特征选择及在科技文献分类中的应用[J]. 兰州理工大学学报, 2015, 41(6):109-113.

doi:10.3772/j.issn.2095-915x.2018.03.009

[12] Sricharan K, Hero A O. Efficient anomaly detection using bipartite k-NN graphs[J]. Advances in Neural Information Processing Systems, 2011:478-486.
[13] 白小明, 邱桃荣. 基于SVM和KNN算法的科技文献自动分类研究[J]. 微计算机信息, 2006, 22(36):275-276.
[14] 刘继才. 科技文献自动分类系统设计与实现[D]. 郑州: 河南工业大学, 2013.
[15] 陈鑫卿, 张永奎, 李荣陆. 一种协调的科技文献分类方法[J]. 计算机工程与应用, 2003, 39(26):91-93.
[16] 田萱, 刘希玉, 孟强. 基于BP神经网络的文档聚类研究[J]. 计算机科学, 2002, 29(8):93-95.
[17] 周丽红, 刘勘. 基于关联规则的科技文献分类研究[J]. 图书情报工作, 2012, 56(4):12-16.
[18] 林堃, 白清源, 谢丽聪,等. 基于规则置信度调整的关联文本分类[J]. 计算机科学, 2008, 35(3):173-176.
[19] 姚力群, 陶卿. 局部线性与One-Class结合的科技文本分类方法[J]. 计算机研究与发展, 2005, 42(11):1862-1869.
[20] 王娟, 范少萍, 郑春厚. 基于惩罚性矩阵分解的文本聚类分析[J]. 情报学报, 2012, 31(9):998-1008.
[21] Lee K Y, Kim D W, Lee K H, et al. Density-induced support vector data description[J]. Neural Networks, IEEE Transactions on, 2007, 18(1): 284-289.
[22] Tian J, Gu H, Gao C, et al. Local density one-class support vector machines for anomaly detection[J]. Nonlinear Dynamics, 2011, 64(1):127-130.
[23] Cao V L, Nicolau M, Mcdermott J. One-Class Classification for Anomaly Detection with Kernel Density Estimation and Genetic Programming[M]. Genetic Programming. Springer International Publishing, 2016:3-18.
[24] Chang C C, Lin C J. LIBSVM -- A Library for Support Vector Machines[EB/OL]. [2018-01-11]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[25] 张华平. NLPIL下载[EB/OL]. [2018-01-13]. http://ictclas.nlpir.org/newsdownloads?DocId=389.