基于非线性流形学习和支持向量机的文本分类算法

任剑锋 梁 雪 李淑红
(河南财经政法大学计算机与信息工程学院 郑州 450002)

摘 要 为解决文本自动分类问题,提出一种流形学习和支持向量机相结合的文本分类算法(LLE-LSSVM)。LLE-LSSVM算法利用非线性流形学习算法LEE对高维文本特征进行非线性降维,挖掘出特征内在规律与本征信息,从而得到低维特征空间,然后将其输入到LSSVM中进行学习,同时利用混沌粒子群算法对LSSVM参数进行优化,建立文本分类模型。仿真实验结果表明,LLE-LSSVM算法提高了文本分类准确率,减少了分类运行时间,是一种有效的文本分类算法。

关键词 文本分类,支持向量机,流形学习,遗传算法
中图法分类号 TP311,TP301 **文献标识码** A

Text Categorization Algorithm Based on Manifold Learning and Support Vector Machines

REN Jian-feng LIANG Xue LI Shu-hong
(School of Computer and Information Engineering, Henan University Economics and Law School, Zhengzhou 450002, China)

Abstract In order to solve the text classification problem, this paper put forward a text classification algorithm based on manifold learning and support vector machine (LLE-LSSVM). Firstly, high dimension text characteristics are reduced by LEE algorithm, and the inner rule and characteristics of the information are mined to obtain meaningful low-dimensional feature space. Secondly the features are input into the LSSVM to be learnt while using chaotic particle swarm algorithm to optimize LSSVM parameters. Lastly establishes the text classification model. The simulation results show that the proposed algorithm improves text classification accuracy and reduces the classification time, and it is an effective text classification algorithm.

Keywords Text categorization, Support vector machines, Manifold learning, Genetic algorithm

1 引言

随着信息处理和通信技术的飞速发展,网络资源日益丰富,其大部分以文本形式储存,如何从海量文本中找到用户所需要的信息,越来越重要。文本页分类(Text Categorization, TC)是解决该问题的重要工具,因此文本自动分类技术成为一个富有挑战性且十分迫切的研究课题[1]。

在文本自动分类过程中,有两个关键问题需要解决:文本特征提取、选择和文本分类器的设计[2]。网络文本是一种特殊结构的文本,具有样本稀疏、维数高和特征不太明显等特点,特征间非线性关联性较大,信息冗余严重[3]。传统特征提取方法主要有主成分分析(PCA)、投影寻踪等,这些方法都属于线性流形学习方法,无法选择最佳文本分类特征,影响后继文本分类器的分类效果[4,5]。非线性流形学习能够找出原始高维特征空间中嵌入的低维流形,挖掘出特征内在规律与本征信息,从而实现数据约简,为文本特征提取开拓了新的空间[6]。当前文本分类器主要采用人工智能技术中的神经网络和支持向量机[7,8]。神经网络具有自身难以克服的缺陷,如局部极优、网络结构复杂等,分类结果常出现过学习问题,文本分类效果与实际需求有一定的差距[9]。支持向量机(Support Vector Machines, SVM)是一种基于统计学习理论和结构风险最小原则的机器学习方法,较好地解决了非线性、维数灾等难题,不易产生类似神经网络的局部最优及过学习等缺陷,在分类领域性能一般优于神经网络[10]。然而SVM学习性能和泛化能力取决于其参数的合理选择,目前SVM参数选择方法主要有梯度下降算法、遗传算法(Genetic Algorithm, GA)和粒子群优化算法(Particle Swarm Optimization, PSO)等,这些方法均有自己的不足,难以获得优SVM参数,影响SVM在文本自动分类中的性能[11]。

为了提高文本自动分类精度,首先采用非线性流形学习中的局部线性嵌入算法(Local Linear Embedding, LLE)对文本特征进行选择,然后采用混沌粒子群算法对SVM分类器参数进行优化,最后采用仿真实验对文本分类模型的性能进行测试,验证其有效性和可行性。

2 文本自动分类的工作原理

文本自动分类根据给定文本内容,将其判别为事先确定

到稿日期:2011-02-25 返修日期:2011-05-11 本文受河南省科学技术厅科技攻关科学项目(112102210199),河南省科学技术厅基础与前沿研究项目(112300410201)资助。

任剑锋(1979—),男,硕士,讲师,主要研究方向为软件工程;梁 雪(1982—),女,讲师,主要研究方向为计算机网络;李淑红(1972—),女,博士,副教授,主要研究方向为计算机应用。

的若干个文本类别某一类的过程，文本内容可以是媒体新闻、电子邮件、科技报告、网页和技术专利等。文本分类算法的框架图如图 1 所示。

图 1　文本分类算法的框架

3　基于流形学习和支持向量机的分类算法

3.1　局部线性嵌入算法

流形学习将高维空间数据通过一定策略映射到低维数据空间，并使数据之间的几何关系和距离测度保持不变，较好地解决了数据处理中的“维数灾难”问题。尤其是近几年提出的非线性流形学习算法 LLE，其能够较好地发现文本空间中的局部几何结构，且具有计算简单的优点，因此本研究采用 LLE 算法对文本进行降维处理。

给定文本数据集 $X=\{x_1,x_2,\cdots,x_n\}$，LLE 算法在保持文本数据映射前后、局部几何结构不变的同时，获得低维数数据集：$Y=\{y_1,y_2,\cdots,y_n\}$。LLE 算法的具体运行步骤如下：

1)各个点的邻域的选取。首先确定邻域点个数 k，然后采用欧式距离法计算出 x_i 最近的 k 个近邻点。

2)重构权值矩阵的计算。设 w_{ij} 表示数据点 x_j 对 x 的重构贡献，那么权重矩阵 W 通过最小化式(1)求得：

$$\varepsilon_1(W)=\sum_i \| x_i-\sum_j w_{ij}x_j \|^2 \tag{1}$$

3)低维嵌入计算。通过最小化式(2)求得低维向量 y_i：

$$\varepsilon_2(W)=\sum_i \| y_i-\sum_j w_{ij}y_j \|^2 \tag{2}$$

式中，$\sum_i y_i=0$ 且 $\frac{1}{N}\sum_i y_i y_i^{\mathrm{T}}=I$。

LLE 算法的具体流程如图 2 所示。

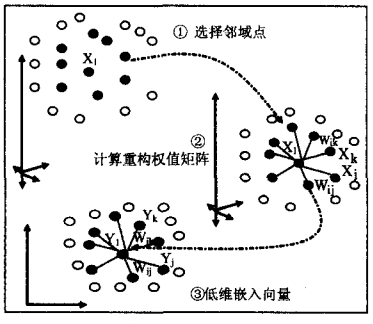

图 2　LLE 算法的工作流程

3.2　支持向量机及参数优化

3.2.1　支持向量机分类算法

由于文本样本的数量比较大，支持向量机对于大样本，学习速度比较慢，因此本文采用最小二乘支持向量机(Least Square Support Vector Machines，LSSVM)算法。LSSVM 是标准 SVM 的一个变种，加快了求解的速度，降低了计算复杂度，因此本文采用 LSSVM 对文本进行分类。

设共有 n 个文本训练样本$\{(x_1,y_1),\cdots,(x_n,y_n)\}$，$x_i$ 和 y_i 分别表示输入和输出向量，$y_i\in\{1,-1\}$，LSSVM 分类实质上是最小化下列函数。

$$J(\omega,\xi)=\frac{1}{2}\| \omega \|^2+\frac{1}{2}\gamma\sum_{i=1}^{n}\xi_i^2 \tag{3}$$

式中，ξ 为松驰变量，ω 为权向量。

引入拉格朗日乘子对式(3)进行优化求解，即

$$L(\omega,b,\xi,\alpha)=J(\omega,\xi)-\sum_{i=1}^{n}\alpha_i(y_i(\omega'\phi(x_i)+b)-1+\xi_i) \tag{4}$$

式中，α_i 表示拉格朗日乘子。

对式(4)进行优化，然后对其进行求偏导得到：

$$\begin{cases}\dfrac{\mathrm{d}L}{\mathrm{d}\omega}=0\rightarrow\omega=\sum_{i=}^{n}\alpha_i y_i\phi(x_i)\\ \dfrac{\mathrm{d}L}{\mathrm{d}\xi}=0\rightarrow\xi_i=\alpha_i/\gamma\\ \dfrac{\mathrm{d}L}{\mathrm{d}b}=0\rightarrow\sum_{i=1}^{n}\alpha_i=0\\ \dfrac{\mathrm{d}L}{\mathrm{d}\alpha_i}=0\rightarrow y_i(\omega'\phi(x_i)+b)+\xi_i-1=0\end{cases} \tag{5}$$

消除 ω 和 ξ_i，并转换为求解分块矩阵。

$$\begin{bmatrix}0 & Y'\\ Y & ZZ'+\gamma^{-1}\end{bmatrix}\begin{bmatrix}b\\ a\end{bmatrix}=\begin{bmatrix}0\\ I\end{bmatrix} \tag{6}$$

这样，最终得到 LSSVM 分类函数。

$$y(x)=\mathrm{sign}(\sum_{i=1}^{n}\alpha_i y_i k(x,x_i)+b) \tag{7}$$

式中，$k(x,x_i)$表示核函数。

3.2.2　支持向量机参数优化

混沌粒子群算法(CPSO)是在基本粒子群优化(Particle Swarm Optimization，PSO)算法的基础引入混沌思想形成的，其利用混沌运动具有的对初始条件的敏感性、随机性和遍历性等优点，提高了种群的多样性和粒子搜索的遍历性，从而改善了基本 PSO 局部寻优能力弱，后期收敛速度慢和易陷入局部最优等缺陷，加快了收敛速度和提高精度。CPSO 是一种基于群体智能算法，它通过粒子间的协作与竞争，在多维空间中搜索最优解，最后找到全局最优解。

引入 Logistic 方程构造混沌系统：

$$z_{m+1}=\mu z_m(1-z_m),m=0,1,2,\cdots \tag{8}$$

式中，μ 表示控制参量，当值为 4 时，系统完全处于混沌状态，可以迭代出一系列特定时间序列。

本研究 LSSVM 采用高斯核函数，其需要优化的参数为 γ 和 σ，将 γ 和 σ 组合在一起作为 CPSO 的粒子，采用文本分类误差作为 CPSO 的适应度函数，通过粒子间相互协作找到最优粒子，即 LSSVM 模型的最优参数。CPSO 对 LSSVM 参数的优化流程如图 3 所示。

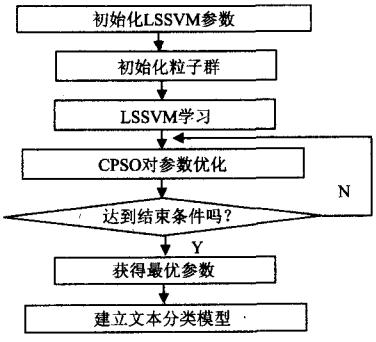

图 3　LSSVM 参数优化过程

3.3 文本自动分类器的设计

因为网络中的信息有多种类型，所以文本自动分类实质上是一个多分类问题，而支持向量机(SVM)是针对两分类问题进行设计的，因此必须针对 SVM 两分类器采用一定的策略构造成一个文本多分类器，从而实现文本自动分类。当前多分类器构造有两种方式："一对多"和"一对一"，"一对多"十分复杂，效率低，不适合实时性要求高的在线文本分类，因此本研究采用"一对一"的方式构造文本多分类器，具体构造如图 4 所示。

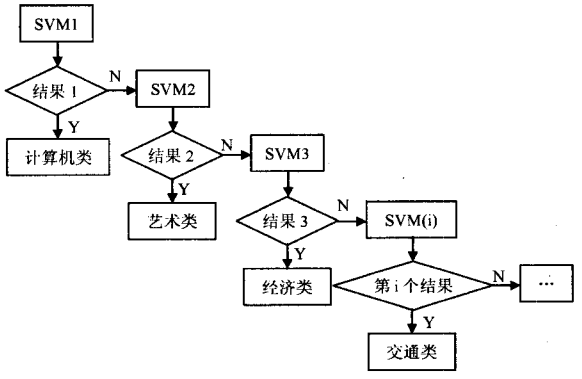

图 4 文本自动分类器模型

4 仿真实验

4.1 数据集来源

采用中国科学院计算技术研究所的中文语料库作为仿真数据，对文本分类器进行分类测试。其中训练语料集包含 1500 篇文本文件，测试语料包含 377 篇文本文件，两者无重叠，其有 10 个文本类别，见表 1。

表 1 文本分类数据集

类别	训练集	测试集	合计
计算机	120	30	150
艺术	130	33	163
教育	150	38	188
交通	100	25	125
环境	80	20	100
经济	100	25	125
医药	120	30	150
军事	150	38	188
政治	300	75	375
体育	250	63	313

4.2 仿真结果与分析

为了使结果具有可比性，采用主成分分析(PCA)的最小二乘支持向量机作为对比算法(PCA-LSSVM)，主要从分类精度对算法进行评价。

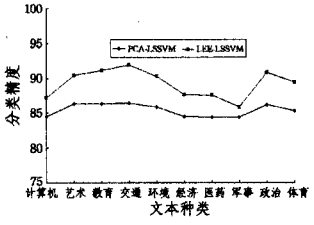

图 5 PCA-LSSVM 和 LLE-LSSVM 分类性能

首先对文本数据集进行中文分词和词性选择以及文本向量化，然后分别采用 PCA 和 LLE 算法对高维特征向量进行降维，最后采用 LSSVM 对文本分类模型进行建模，同时采用 CPSO 对 LSSVM 参数进行优化，从而建立 PCA-LSSVM 和 LLE-LSSVM 文本分类模型。采用建立的文本分类模型对文本测试样本进行分类得到的文本分类结果如图 5 所示。

从图 5 可知，对于每一种文本类别，LLE-LSSVM 分类精度都要比 PCA-LLSVM 精度高，而且分类速度也明显加快，这说明通过 LLE 对文本特征进行降维后，提取的特征可以很好地反映文本原始特征的信息，计算量大大减小，时间效率得到显著提高；LLE 算法在复杂度和抗噪性等方面比 PCA 算法有着更大的优势，同时采用 CPSO 对 LSVM 进行参数优化，进一步提高了文本分类精度，仿真结果表明，LLE-LSSVM 算法应用于文本自动分类是切实可行的、有效的。

结束语 文本分类是一种高维、非线性分类问题，传统特征降维方法速度慢，分类精度不高，本研究将非线性流学习中的 LLE 算法引入到文本分类特征降维中，采用分类能力强的 LSSVM 作为分类器，并采用 CPSO 对 LSSVM 的参数进行优化。仿真对比实验表明，LLE-LSSVM 提高了文本分精度和分类效率，在文本分类中有着很好的应用前景。

参 考 文 献

[1] 袁鼎荣，钟宁，张师超. 文本信息处理研究述评[J]. 计算机科学，2011，38(2)：9-14

[2] 黄豫清，戚广志，张福炎. 从 Web 文档中构造半结构化信息的抽取器[J]. 软件学报，2000，11(1)：73-78

[3] 冯书晓，徐新，杨春梅. 国内中文分词技术研究新进展[J]. 情报杂志，2002，11：29-30

[4] 胡佳妮，等. 中文文本分类中的特征选择算法研究[J]. 光通信研究，200，5(3)：44-46

[5] 朱明，王俊普. 一种最优特征集的选择算法[J]. 计算机研究与发展，1998，35(9)：803-805

[6] 王自强，钱旭. 基于流形学习和 SVM 的 Web 文档分类算法[J]. 计算机工程，2009，35(15)：38-40

[7] 冯永，李华，钟将，等. 基于自适应中文分词和近似 SVM 的文本分类算法[J]. 计算机科学，2010，37(1)：251-255

[8] Cai Deng，He Xiao-fei，Han Jia-wei. Document Clustering Using Locality Preserving Indexing [J]. IEEE Transactions on Knowledge and Data Engineering，2005，17(12)：1624-1637

[9] 邓乃扬，田英杰. 数据挖掘中的新方法：支持向量机[M]. 北京：科学出版社，2004

[10] 巩知乐，张德贤，胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真，2009，26(7)：165-168

[11] 向昌盛，周子英，张林峰. 相空间重构和支持向量机参数联合优化研究[J]. 湖南科技大学学报：自然科学版，2010，25(4)：81-85

[12] 曾立梅. 基于文本数据挖掘的硕士论文分类技术[J]. 重庆邮电大学学报：自然科学版，2010，22(5)：669-672