基于流形正则化的文档分类算法研究

徐海瑞 张文生 吴 双
(中国科学院自动化研究所 北京 100190)

摘 要 基于流形正则化框架提出一种分类算法(MLD-RLSC),以解决高维文档分类问题。该算法通过构建训练样本的最近邻图来估计数据空间的几何结构并将其作为流形正则化项,结合多变量线性回归获得高维文档的低维流形结构,并采用k近邻分类器对低维流形进行分类,得到针对多类问题的分类器。该算法能够充分利用训练样本的类别信息来帮助学习以提取有效特征。通过在Reuters-21578数据集上的实验,证明该算法的分类性能和运行速度比传统分类器有较大的提高。

关键词 局部鉴别嵌入,流形学习,文档分类,k近邻,流形正则化
中图法分类号 TP181 **文献标识码** A

Document Classification Algorithm Based on Manifold Regularization

XU Hai-rui ZHANG Wen-sheng WU Shuang
(Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract A novel document classification algorithm based on manifold regularization framework, which is called MLD-RLSC, is presented to resolve high dimensional document classification. In the proposed MLD-RLSC, a nearest neighbor graph was constructed and the intrinsic geometrical structure of the sample space was taken as a manifold regularization term, then it was incorporated into the objective function of the multivariate linear regression to extract lower dimensional space. The classification and predication in the lower dimensional feature space are implemented with kNN. Aiming to extract effective features for the multi-class problem, MLD-RLSC can make use of all labeled samples. Experimental results on Reuters-21578 dataset demonstrate that the proposed algorithm is of higher classification accuracy and faster running speed.

Keywords LDE, Manifold learning, Text categorization, kNN, Manifold regularization

1 引言

文档分类技术是信息检索和搜索引擎的关键技术基础,其主要任务是在给定的分类体系下,根据文档的内容自动地确定与文档关联的类别。近年来,随着信息技术的飞速发展,网络中的文档资源呈现爆炸式增长。文档分类通过对文档自动进行类别标注,能有效地辅助人们组织和管理信息,目前其已成为信息检索领域的研究热点。对于文档分类而言,其难点之一是提取文本特征。原始文档数据的维数很高,通常需要降维处理,而降维的特征空间维度对文档分类效果影响甚大。其维度过高势必消耗大量的计算资源,给文档的快速分类造成极大困难。而维度太低又无法正确表示目标文本,从而有可能丢弃大量具有较高相关度的文本资源。另一个难点是文档分类对效率的要求,如何提高分类效率,从而迅速判定目标文档与主题是否相关,也是整个研究的关键。

传统的文档分类算法在处理高维大规模数据集时,存在着计算复杂、分类效率和精度不高的弱点。其主要原因之一在于高维文档空间特征表示不够理想。常用的如PCA[1],LDA[2,3]等算法,主要通过保持样本空间的全局几何结构来提取有效的低维特征,而往往简单的线性映射无法合理地表示文档复杂的空间结构。近些年提出的流形学习[4]作为一种数据降维技术,如LLE[5],LPP[6],NPE[7]算法等,能够有效地发现隐藏在高维数据中的非线性流形结构,避免"维数灾难",提高分类器的性能和计算效率。

在文档分类应用中,样本虽然维数很高,但对分类任务而言,可认为其本质上分布于一个低维的流形子空间,因此可以利用训练样本帮助学习平滑的流形子空间。本文采用基于线性回归的流形正则化框架,充分利用训练样本的类别信息,提出了一种新的分类算法MLD-RLSC,用于高维文档分类问题。本算法基于局部鉴别嵌入(LDE)算法[8],通过构建所有训练样本的最近邻关联图来估计数据空间的边缘几何结构,然后将该结构作为正则化项引入到分类问题中,通过对多变量的线性回归增加此正则化项与合适的指示矩阵,并结合kNN分类算法[9],得到针对多类问题的分类器。实验结果表明,在相同测试数据条件下,本算法的分类性能和运行速度相对于传统分类器有较大的提高。

到稿日期:2011-04-24 返修日期:2011-07-18 本文受国家自然科学基金(90924026),国家(863)项目(2008AA01Z121)资助。
徐海瑞(1984—),男,硕士生,主要研究方向为机器学习与数据挖掘,E-mail:hairui.xu@ia.ac.cn;张文生(1966—),博士,研究员,博士生导师,主要研究方向为机器学习理论与算法、人机交互技术;吴 双(1985—),硕士生,主要研究方向为机器学习与数据挖掘。

本文主要解决高维文档数据分类问题，第 2 节介绍本文的研究基础：LDE 和线性回归；第 3 节论述基于流形正则化的文档分类算法；第 4 节通过实验证明本文算法的优势；最后是全文的总结。

2 研究基础

2.1 局部鉴别嵌入算法

局部鉴别嵌入 LDE(Local Discriminant Embedding)是最近提出的一种用于数据降维的流形学习算法。由于它能够在计算数据映射的过程中同时考虑数据局部几何结构和类别标记信息，因此在经过 LDE 映射后，能够保存相同类别内的紧凑性和不同类别间的分离性，有利于提高后继分类器的性能和计算效率。其算法描述如下：

Step1 构建邻接图。设 G 和 G' 均表示一个具有 n 个顶点的无向图，且每个顶点 x_i 的类别标记为 c_i。构造图 G 的方法如下：如果顶点 x_i 和 x_j 具有相同的类标签(即 $c_i=c_j$)，且 x_j 是 x_i 的 k 个最近邻居之一，则在顶点 x_i 和 x_j 之间构建一条边。同理，构造图 G' 的方法如下：如果顶点 x_i 和 x_j 具有不同的类标签(即 $c_i\neq c_j$)，且 x_j 是 x_i 的 k 个最近邻居之一，则在顶点 x_i 和 x_j 之间构建一条边。

Step2 计算图 G 和 G' 中边的权值。为简化计算，在图 G 中，如果顶点 x_i 和 x_j 之间有一条边，则赋权值 $w_{ij}=1$，否则 $w_{ij}=0$；同理，如果图 G' 中，顶点 x_i 和 x_j 之间有一条边，则赋权值为 $w'_{ij}=1$，否则 $w'_{ij}=0$。

Step3 计算局部鉴别嵌入。优化目标函数为 $\max\sum_{i,j}\|V^Tx_i-V^Tx_j\|^2w'_{ij}$，约束条件为：$\sum_{i,j}\|V^Tx_i-V^Tx_j\|^2w_{ij}=1$，问题可转换为求解下式的特征值和特征向量：

$$X(D-W)X^Tv=\lambda X(D'-W')X^Tv$$

式中，$X=(x_1,\cdots,x_m)$，W 和 W' 分别表示图 G 和 G' 中的边权值矩阵，D 和 D' 表示对角矩阵，其定义为：$d_{ij}=\sum_j w_{ij}$ 和 $d'_{ij}=\sum_j w'_{ij}$。设 $v_0,\cdots,v_{d-1}$ 是其特征向量，按照其对应的特征值由小到大排列，即 $\lambda_0\leqslant\cdots\leqslant\lambda_{d-1}$，则由高维数据得到的低维嵌入可表示为 $x_i\rightarrow y_i=V^Tx_i$，其中 $V=(v_0,v_1,\cdots,v_{d-1})$。

从以上 LDE 算法过程可以看出，该算法具有以下优点：

(1)LDE 算法在寻找高维数据空间的低维映射过程中，同时考虑了局部数据结构和样本类标签，因此映射后的数据能更加适合分类的要求；

(2)算法采用了线性映射的方法，从而很好解决了传统流形学习算法无法直接得到新测试数据的低维嵌入表示问题；

(3)算法的低维嵌入向量是通过求解稀疏矩阵的特征向量得到的，因此具有计算量小、运行速度快的优点。

2.2 线性回归

线性回归是一种经典的统计分析方法，旨在寻找一个线性函数，拟合变量及其响应变量之间的关系。对于 n 个变量及其响应变量 $\{x_i,y_i\}_{i=1}^n$，线性回归要求最小化平方误差和：

$$\min\sum_{i=1}^n(y_i-v^Tx_i)^2 \tag{1}$$

同时，不失一般性，假设 $\{x_i\}$ 和 $\{y_i\}$ 具有零均值，即 $\sum_{i=1}^n x_i=0$，$\sum_{i=1}^n y_i=0$，式(1)可写为：

$$J_{LR}(w)=\min(y-v^TX)(y-v^TX)^T \tag{2}$$

对 v 求偏导并令其等于零，可得 $v_{LR}=(XX^T)^{-1}Xy^T$。当取响应变量 y 为类标签时，线性回归可用于分类问题，并且与 LDA 有着紧密的联系[10,11]。尤其是在两分类问题中，v_{LR} 等效于两类情况下的 LDA。

3 基于流形正则化的文档分类算法

3.1 MLD-RLSC 算法

在文档分类中，尽管文档样本有很高的维度，但可认为其本质上是分布于一个低维的流形空间。此外，在实际应用中，获得样本的类别信息往往非常困难。如何利用样本的类别信息帮助学习分类，具有重要的意义。因此，可以利用训练样本的类别信息，通过估计样本空间几何边缘分布，将其作为正则项来帮助学习平滑的流形子空间。

本文采用基于线性回归的流形正则化框架[12]，结合局部鉴别嵌入(LDE)算法的优势，提出一种基于流形正则化的文档分类算法 MLD-RLSC(Multiple Local Discriminant Regularized Least Squares Classification)。其主要思想为：通过构建训练样本的最近邻关联图来估计数据空间的边缘几何结构，然后将该结构作为正则化项引入到分类问题中，通过对多变量的线性回归增加此正则化项与合适的指示矩阵，并结合 kNN 分类器，得到针对多类问题的 MLD-RLSC 算法。

设给定 l 个有标签训练样本 $\{x_i,y_i\}_{i=1}^l$，依据 LDE 算法构建近邻图，并得到类内和类间的权值矩阵。为估计未知函数 f，将图结构作为正则化项得到如下目标函数：

$$J(f)=\min\sum_{i=1}^l(y_i-f(x_i))^2-\alpha\sum_{i,j=1}^{l+u}(f(x_i)-f(x_j))^2w_{ij} \tag{3}$$

其约束条件为 $\sum_{i,j}\|f(x_i)-f(x_j)\|^2w_{ij}=1$。

式中，$f=[f(x_1),f(x_2),\cdots,f(x_n)]$。当 f 为线性函数时，即 $f(x)=v^Tx$，目标函数变为：

$$J(f)=\min\sum_{i=1}^l(y_i-v^Tx_i)^2-\alpha v^TX(D'-W')X^Tv \tag{4}$$

对于两类问题，即令 $y_i\in\{0,1\}$，计算式(4)可得到：

$$v=(XX^T-\alpha X(D'-W')X^T+\beta X(D-W)X^T)^{-1}Xy^T \tag{5}$$

式中，W 和 W' 分别表示图 G 和 G' 中的边权值矩阵，D 和 D' 表示对角矩阵，其定义为 $d_{ij}=\sum_j w_{ij}$ 和 $d'_{ij}=\sum_j w'_{ij}$，$X=[x_1,x_2,\cdots,x_l]$，且 $y=[y_1,y_2,\cdots,y_l]$，α 和 β 为正则化参数，可通过交叉验证的方法选取。

对于多类问题，仅仅得到一个投影向量，不能充分描述样本复杂的多类结构，因此通过多变量线性回归进行流形正则化并采用合适的指示矩阵 Y，可以建立学习目标函数，具体形式为：

$$J(f)=\min\sum_{i=1}^l(y_i-v^Tx_i)^2-av^TX(D'-W')X^Tv \tag{6}$$

其约束条件为 $\sum_{i,j}\|f(x_i)-f(x_j)\|^2w_{ij}=1$。求解如上目标函数，可得 $V=(XX^T-\alpha X(D'-W')X^T+\beta X(D-W)X^T)^{-1}XY^T$，其中 $V=[v_1,v_2,\cdots,v_c]$，$Y=(Y_{ji})\in R^{c\times n}$。为简化计算过程，令 Y_{ji} 的取值为：如果 $y_i=j$，则 $Y_{ji}=1$；否则 $Y_{ji}=0$。

由此可以看出，高维文档到低维特征空间的映射可表示

为 $x_i \to z_i = V^T x_i$。可计算得到训练样本和测试样本经 V 映射后的低维向量表示，结合 kNN 分类器即可得到针对多类问题的 MLD-RLSC 算法。

通过以上的分析推理，可以总结出 MLD-RLSC 算法具有以下优点：

(1)可针对多类的情况，能充分利用训练样本的标签信息；

(2)由于同时利用了样本空间的全局和局部几何结构，MLD-RLSC 兼具 LDA 和 LDE 的优点，能够更好地表示样本空间的特点；

(3)能够得到 c 个映射向量，样本 x 被映射到低维空间，可简化计算，提高分类效率。

3.2 算法流程

设给定文档样本集 $X=\{x_1, x_2, \cdots, x_n\}$，其中 $x_i \in R^N$。由于向量空间模型(VSM)[13]是文档表示的主要形式，因此本文采用 VSM 中常用的 TF-IDF[14]权重向量表示每个文档 x，提取词干，去停用词，归一化，得到文档的向量集合。然后进行降维处理，最后对降维后的高维文档数据进行分类判定。算法的流程图如图 1 所示。

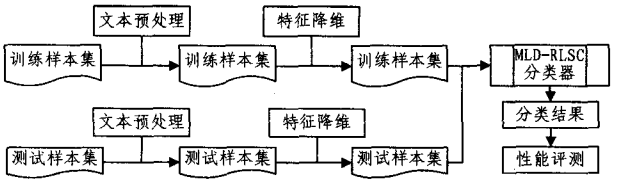

图 1 算法流程图

在流程图中，训练集和测试集的特征映射及分类是算法的主要部分，具体步骤如下：

Step1 文档向量预处理。为了保证使优化目标不包含平凡解，减少噪声对分类结果的影响，首先采用 PCA 方法对文档向量进行预处理，消除平凡解。经过 PCA 投影后，文档向量矩阵 X 变成 $\widetilde{X}=\{\widetilde{x}_1, \widetilde{x}_2, \cdots, \widetilde{x}_n\}$。

Step2 构建邻接图。设 G 和 G' 均表示一个具有 n 个顶点的无向图，且每个顶点 x_i 的类别标记为 c_i。构造图 G 的方法如下：如果顶点 x_i 和 x_j 具有相同的类标签(即 $c_i=c_j$)，且 x_j 是 x_i 的 k 个最近邻居之一，则在顶点 x_i 和 x_j 之间构建一条边。同理，构造图 G' 的方法如下：如果顶点 x_i 和 x_j 具有不同的类标签(即 $c_i \neq c_j$)，且 x_j 是 x_i 的 k 个最近邻居之一，则在顶点 x_i 和 x_j 之间构建一条边。

Step3 计算图 G 和 G' 中边的权值。为简化计算，在图 G 中，如果顶点 x_i 和 x_j 之间有一条边，则赋权值 $w_{ij}=1$，否则 $w_{ij}=0$。同理，如果图 G' 中，顶点 x_i 和 x_j 之间有一条边，则赋权值 $w'_{ij}=1$，否则 $w'_{ij}=0$。

Step4 计算特征映射。依据上述多类优化目标式(6)，求解可得

$$V=(XX^T-\alpha X(D'-W')X^T+\beta X(D-W)X^T)^{-1}XY^T \tag{7}$$

式中，$X=[x_1, x_2, \cdots, x_n]$。高维文档到低维特征空间的映射可表示为 $x_i \to z_i = V^T x_i$，其中 $V=(v_0, v_1, \cdots, v_c)$。计算可得训练样本和测试样本经 V 映射后的低维向量表示。

Step5 利用 kNN 分类判定。将原始高维文档集 $X=\{x_1, x_2, \cdots, x_n\}$ 经 MLD-RLSC 映射后所得的低维文档特征集 $Y=\{y_1, y_2, \cdots, y_n\}$，通过 kNN 分类器来确定测试文档所属的类别。

4 实验结果

为了测试本文算法 MLD-RLSC 的分类性能，将 MLD-RLSC 算法与其他基于特征降维的文档分类算法进行了比较，分别为：(1)直接采用文档频率(DF)[15]方法进行特征选择并利用 kNN 进行分类，作为实验比较基线，简称为 Baseline；(2)首先采用主成分分析对文档进行降维，然后利用 kNN 进行文档分类的算法，以下简称为 PCA-kNN；(3)首先采用线性判别分析对文档进行降维，然后利用 kNN 进行文档分类的算法，以下简称为 LDA-kNN。

实验系统的硬件环境为 CPU Inter Core2 E6550，主频为 2.33GHz，内存为 2G、2.34GHz；操作系统为 Windows XP；开发环境为：Microsoft Visual Studio 2005，代码采用标准的 C++STL 编写。本文在测试数据采用了分类领域标准测试数据集 Reuters-21578，遵循“ModApte”切分方式，选择文档最多的 5 个类别(acq, earn, trade, money-fx, crude)进行实验。为了减小样本类别数目的不一致带来的偏差，对样本数较多的两类 acq 和 earn 选择部分具有代表性的样本作为训练集。处理后的样本集合共包括 1991 个训练文档和 804 个测试文档。文档集类别如表 1 所列。

表 1 文档集类别表

文档类别	训练样本	测试样本
acq	420	250
earn	450	118
trade	353	300
money-fx	400	90
crude	368	46

在实验的所有方法中，文档表示均采用向量空间模型，并利用 TF-IDF 方法进行词项加权。经过剔除停用词、提取词干等文档预处理后的原始特征集维数为 10180。其中文档频率(DF)的阈值设定为 0.01。在 MLD-RLSC 方法中，正则化参数选取 $\alpha=0.01$，$\beta=0.01$，特征选择维度为文档类别数 5，其他方法均选择分类性能保持最好维度的实验结果进行比较。

分类算法的性能评价指标使用常用的准确率(*Precision*)、召回率(*Recall*)以及 F 值，这 3 个平均指标的值越大，说明该算法分类性能越好。它们的定义分别是

$$Precision=\frac{\text{正确分类的文档数}}{\text{测试集中分为该类的文档数}}\times 100\% \tag{8}$$

$$Recall=\frac{\text{正确分类的文档数}}{\text{测试集中属于该类的文档数}}\times 100\% \tag{9}$$

$$F=\frac{2\times Precision\times Recall}{Precision+Recall}\times 100\% \tag{10}$$

本文使用准确率、召回率和 F 值指标来衡量不同算法在每一个类别上的分类性能。为了评估算法在整个数据集上的性能，有两种平均的方法可供使用，即宏平均(macro_average)和微平均(micro_average)。宏平均是每一个类别的性能指标的算术平均值，而微平均是每一个实例的性能指标的算术平均。对于单个实例而言，它的准确率和召回率是相同的，

因此准确率和召回率的微平均是相同的。根据式(10)可知，对于同一个数据集，这3个指标均相同。

实验中，我们采用 Baseline，PCA-kNN，LDA-kNN 在测试集上的实验结果与本文提出的 MLD-RLSC 算法进行比较，具体的实验结果如表2所列。

表2　4种算法的分类性能比较

Method	宏平均			微平均
	Precision	Recall	F	F
MLD-RLSC	0.94421	0.9534	0.9474	0.9527
PCA-*kNN*	0.9080	0.9213	0.9122	0.9403
LDA-*kNN*	0.8668	0.8595	0.8568	0.8744
Baseline	0.7739	0.7522	0.6917	0.7438

从表2可以看出，MLD-RLSC 方法在准确率、召回率和 F 值等各项评价指标上均高于其他3种分类算法。原因在于本算法能利用训练样本的标签信息，在降维过程中保持类内的紧凑性和类间的分离性，估计样本几何流形结构，从而保证了很好的分类性能。

此外，为了说明本文所提出的 MLD-RLSC 算法的高效性，我们通过实验进一步对这4种分类算法的运行时间进行了比较，实验结果如表3所列。

表3　4种算法的运行时间比较

算法	运行时间
MLD-RLSC	36.7348s
LDA-kNN	37.1256s
PCA-kNN	321.9318s
Baseline	241.1980s

从表3可以看出，MLD-RLSC 算法的运行时间与 LDA-kNN 算法相当，远低于其他两种分类算法。原因在于：MLD-RLSC 能够有效地避免高维文档引起的“维数灾难”问题，缩小了分类器的搜索范围，从而保证了算法具有很快的运行速度。

结束语　Web 文档具有高维度、样本稀疏和特征不太明显的特点，传统的特征提取方法通常是保存文档空间的全局结构，实现从高维到低维的线性映射。针对传统文档分类算法在处理高维数据集存在的测试时间过长、分类准确率不高的问题，本文提出基于流形正则化的文档分类算法 MLD-RLSC，该算法能够充分利用训练样本的类别信息帮助学习样本空间的局部几何结构，以提取有效的低维特征并进行分类。实验表明，本文方法具有很好的分类性能和较快的测试运行速度。

参考文献

[1] Ian T J. Principal Component Analysis [J]. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1-3): 37-52

[2] 杨健，杨静宇，叶晖. Fisher 线性鉴别分析的理论研究及其应用[J]. 自动化学报，2003(04)

[3] Martinez A M, K A C. PCA versus LDA[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(2): 328-340

[4] 李波. 基于流形学习的特征提取方法及其应用研究[D]. 合肥：中国科学技术大学，2008

[5] Roweis S T, Saul L K. Nolinear Dimensionality Reduction by Locally Linear Embedding[J]. Science, 2000, 290(5500): 2323-2326

[6] He Xiao-fei. *Locality Preserving Projections* [M]. Cambridge, USA: MIT Press, 2003

[7] He Xiao-fei. Neighborhood Preserving Embedding [C]// Tenth IEEE International Conference on Computer Vision. 2005, 2: 1208-1213

[8] Chen H T. Local Discriminant Embedding and Its Variants[C]// IEEE. 2005

[9] Yang Y, Liu X. A re-examination of text categorization methods [C]// ACM. USA, 1999

[10] Duda R O. Pattern Classification [M]. New York: Wiley-Interscience, 2000

[11] Ye Jie-ping. Least squares linear discriminant analysis[C]// ICML'07. Corvalis, Oregen: ACM Press, 2007

[12] Belkin M. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples[J]. The Journal of Machine Learning Research, 2006, 7: 399-434

[13] Gerald S, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620

[14] 林永民，吕震宇，赵爽，等. 文本特征加权方法 TF·IDF 的分析与改进[J]. 计算机工程与设计，2008(11)

[15] Yang Y. Noise reduction in a statistical approach to text categorization[C]// ACM, Proc. of SIGIR. 1995

(上接第182页)

[4] 谢丽星，孙茂松，佟子健，等. 基于用户查询日志和锚文字的汉语缩略语识别[C]// 中国计算机语言学研究前沿进展，2009. 烟台：清华大学出版社，2009：551-556

[5] 武子英，郑家恒. 现代汉语缩略语自动识别的方法研究[J]. 计算机工程与设计，2007，28(16)：4052-4054

[6] Tian Guogang, Cao Cungen, Liu Lei, et al. MFC: A Method of Co-referent Relation Acquisition from Large-Scale Chinese Corpora[C]//Proceedings of the ICNC'06-FSKD'06. Xi'an, China, 2006: 1256-1261

[7] Guang Jiang, Cao Cungen, Sui Yuefei, et al. A General Approach to Extracting Full Names and Abbreviations for Chinese Entities from the Web[C]//IFIP Advances in Information and Communication Technology. Manchester, UK, 2010: 271-280

[8] 支流，段慧明，朱学锋，等. 中文缩略语知识库建设[C]// 第三届学生计算语言学研讨会论文集，2006. 沈阳：中文信息学会，2006：316-320

[9] 鲍明凌，亢世勇. 基于数据库的现代汉语新词语缩略语的研究[C]// 第一届学生计算语言学研讨会论文集，2002. 北京：中文信息学会，2002：233-238

[10] Han Jia-wei, Kamber M. 数据挖掘概念与技术(第2版)[M]. 范明，孟小峰，译. 北京：机械工业出版社，2007：188-200

[11] 田国刚. 受限中文语料的自监督文本知识获取研究[D]. 北京：中国科学院计算技术研究所，2007

[12] 卢汉，曹存根，岳小莉. 一种根据实体的汉语简称识别出实体全称的方法和系统[P]. ZL200710119513. 中国，2007