第35卷 第15期 计 算 机 工 程 2009年8月
Vol.35 No.15 Computer Engineering August 2009
·软件技术与数据库· 文章编号：1000—3428(2009)15—0038—03 文献标识码：A 中图分类号：TP181

基于流形学习和SVM的Web文档分类算法

王自强，钱 旭
(中国矿业大学(北京)机电与信息工程学院，北京 100083)

摘 要：为解决Web文档分类问题，提出一种基于流形学习和SVM的Web文档分类算法。该算法利用流形学习算法LPP对训练集中的高维Web文档空间进行非线性降维，从中找出隐藏在高维观测数据中有意义的低维结构，在降维后的低维特征空间中利用乘性更新规则的优化SVM进行分类预测。实验结果表明该算法以较少的运行时间获得更高的分类准确率。

关键词：文档分类；流形学习；支持向量机

Web Document Classification Algorithm Based on Manifold Learning and SVM

WANG Zi-qiang, QIAN Xu
(College of Mechanical Electronic and Information Engineering, China University of Mining and Technology(Beijing), Beijing 100083)

【Abstract】To efficiently resolve Web document classification problem, a novel Web document classification algorithm based on manifold learning and Support Vector Machine(SVM) is proposed. The high dimensional Web document space in the training sets are non-linearly reduced to lower dimensional space with manifold learning algorithm LPP, and the hidden interesting lower dimensional structure can be discovered from the high dimensional observisional data. The classification and predication in the lower dimensional feature space are implemented with the multiplicative update-based optimal SVM. Experimental results show that the algorithm achieves higher classification accuracy with less running time.

【Key words】document classification; manifold learning; Support Vector Machine(SVM)

1 概述

Web文档分类技术是信息检索和搜索引擎的关键技术基础，其主要任务是在预先给定的类别标记集合中根据文档内容判定它的类别。目前已提出了多种文档分类技术，如决策树、神经网络、k-近邻、贝叶斯方法和支持向量机(Support Vector Machine, SVM)。由于SVM[1]具有简洁的数学形式、坚实的理论基础、严格的理论分析、直观的几何解释和良好的泛化能力，因此已成为信息检索和搜索引擎的研究热点之一。

虽然SVM在文档分类方面取得了一定的成效，但是Web文档具有维数高、样本稀疏和特征不太明显的特点。而SVM在处理高维大规模数据集时，存在训练时间过长和分类准确率不高的不足。为了更好地对Web文档进行分类，本文提出了基于流形学习的SVM分类算法。该算法首先使用流形学习算法LPP[2-3]对训练集中的高维文档空间进行降维，揭示其流形分布，从中找出隐藏在高维观测数据中有意义的低维结构，然后在降维后的低维特征空间中进行分类预测。实验结果表明该算法是可行的，具有较高的分类准确率和较快的运行速度。

2 研究基础

2.1 流形学习

流形学习的目标是发现嵌入在高维数据空间中的低维流形结构，并给出一个有效的低维表示。它能够较好地解决数据处理中的“维数灾难”问题。尤其是最近提出的流形学习算法LPP能够较好地发现文档空间中的局部几何结构，并具有计算简单的优点，因此，采用基于LPP的降维方法进行文档分类是一种理想的选择。LPP[2]算法的描述如下：

Step1 构建邻接图。设G表示一个具有n个顶点的图，由点i，j之间的欧氏距离$\|x_i-x_j\|^2$确定点j是否在点i的半径ε之内(即$\|x_i-x_j\|^2\leqslant\varepsilon$)或者是点$i$的$k$个最近邻居之一。如果满足上述条件之一，则在$x_i$和$x_j$构建一条边。

Step2 确定权重。如果点i和点j相连接，则它们之间边的权重一般采用热核(heat kernel)定义法，即$S_{ij}=\exp(-\|x_i-x_j\|^2/t)$，其中，$t$为参数。有时，点$i$和点$j$之间的连接边权重也可简单地定义为$S_{ij}=1$。

Step3 计算特征映射。假设G为连接图，否则对每一个连接部分，计算下式的特征向量和特征值：

$$\boldsymbol{XLX}^{\mathrm{T}}\boldsymbol{a}=\lambda\boldsymbol{XDX}^{\mathrm{T}}\boldsymbol{a} \tag{1}$$

其中，$\boldsymbol{D}$为对角矩阵，其定义为：$\boldsymbol{D}_{ii}=\sum_j S_{ij}$，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$是拉普拉斯矩阵，可以看作是定义在图$G$节点上的操作函数。

设$\boldsymbol{a}_0,\boldsymbol{a}_1,\cdots,\boldsymbol{a}_{l-1}$是式(1)的特征向量，并按照其对应特征值由小到大排列，即$\lambda_0<\lambda_1<\cdots<\lambda_{l-1}$，则由高维数据得到的低维嵌入可表示为

$$x_i\rightarrow y_i=\boldsymbol{A}^{\mathrm{T}}x_i \tag{2}$$

其中，$\boldsymbol{A}=(\boldsymbol{a}_0,\boldsymbol{a}_1,\cdots,\boldsymbol{a}_{l-1})$。

由于LPP是通过寻找优化线形映射来发现嵌入在高维数据空间的低维流形结构，并且是基于局部邻域结构的，因此嵌入向量的求解是求解稀疏矩阵的特征向量，从而大大减少

基金项目：教育部科学技术研究基金资助重点项目(107021)
作者简介：王自强(1973－)，男，博士研究生，主研方向：数据挖掘；钱 旭，教授、博士生导师
收稿日期：2009-01-09 **E-mail：**wzqbox@yahoo.cn

了计算量，具有较快的运行速度。

2.2 SVM 分类器

为了高效地对经过 LPP 维数降维后的低维文档空间进行分类，本文采用具有良好泛化能力的 SVM 作为分类器。其实现方法如下：

设训练样本集 $\{(x_i,y_i)\}_{i=1}^{l}$，样本 $x_i\in R^n$，$y_i\in\{-1,+1\}$ 为类标签。SVM 通过求解式(3)找到一个具有最大间隔的超平面：

$$\min_{w,b,\xi}\frac{1}{2}w^{\mathrm{T}}w+C\sum_{i=1}^{l}\xi_i \tag{3}$$

约束条件为

$$y_i[(w\cdot x+b)+\xi_i-1]\geqslant 0,\ \xi_i\geqslant 0 \tag{4}$$

其中，C 为一个用于控制误差的惩罚常数；ξ_i 为非负松弛变量。

利用 Lagrange 乘子法可以把式(3)转化为其对偶形式：

$$\max W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \tag{5}$$

约束条件为

$$\sum_{i=1}^{l}\alpha_i y_i=0,\ \alpha_i\in[0,C],\ i=1,2,\cdots,l \tag{6}$$

于是对未知类别的数据点 x 可采用如下线形判决函数决定其所属类别：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i y_i(x_i\cdot x)+b) \tag{7}$$

对于非线形 SVM 的情况，通过利用非线形映射 ϕ 把输入空间映射到高维特征空间，于是核函数 $K(x_i,x_j)=(\phi(x_i)\times\phi(x_j))$ 可在特征空间进行计算，无须知道映射 ϕ 的具体形式。

用核函数代替线形 SVM 中的点积形式，于是式(5)的对偶形式可变为

$$\max W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{8}$$

约束条件为

$$\sum_{i=1}^{l}\alpha_i y_i=0,\ \alpha_i\in[0,C],\ \ i=1,2,\cdots,l \tag{9}$$

于是非线形 SVM 的判决函数变为

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i y_i K(x_i,x)+b) \tag{10}$$

由于常用的 SVM 求解方法训练代价太大，因此为了有效地优化 SVM 训练算法的性能，本文采用乘性更新方法[4]优化求解 SVM 中的二次规划问题。该方法的优点是：无须参数设置，不需要选择工作子集且具有并行更新变量的能力。实验表明该方法具有快速求解二次规划的能力，并具有较高的分类准确率。

3 Web文档分类算法

设给定文档样本集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$，其中，$x_i\in R^N$。由于向量空间模型(VSM)是文档表示的主要形式，因此本文采用 VSM 中常用的 TF-IDF 项权重向量表示每个文档 x_i，并将其长度归一化。首先利用 LPP 寻找高维矩阵 $\boldsymbol{X}$ 的低维嵌入，然后将降维后的低维嵌入输入到优化训练的 SVM 分类器中进行分类，具体算法步骤如下：

Step1 构建邻接图。设 G 表示一个具有 n 个顶点的图，其中，顶点 i 对应于文档 x_i，如果 x_i 和 x_j 互为近邻点(p 近邻)，则在 x_i 和 x_j 构建一条边。

Step2 定义图中边的权重 $\boldsymbol{S}$。如果顶点 i 和顶点 j 互连，则定义该边的权重 $S_{ij}=x_i^{\mathrm{T}}x_j$，否则 $S_{ij}=0$。权重矩阵 $\boldsymbol{S}$ 定义了文档空间的局部结构。设 $\boldsymbol{D}$ 为对角矩阵，且满足 $D_{ii}=\sum_j S_{ij}$，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$ 是拉普拉斯算子。

Step3 文档向量预处理。为了确保后面的 LPP 优化目标(式(12))不包含平凡解，本文首先将其投影到 SVD 子空间消除平凡解。设 $\boldsymbol{W}_{\mathrm{SVD}}$ 表示 SVD 的变换矩阵，则经过 SVD 投影后，文档向量 $\boldsymbol{X}$ 变为 $\tilde{x}$：

$$\tilde{x}=\boldsymbol{W}_{\mathrm{SVD}}^{\mathrm{T}}x \tag{11}$$

经过 SVD 投影后，文档向量矩阵 $\boldsymbol{X}$ 变为

$$\tilde{\boldsymbol{X}}=\{\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_n\}$$

Step4 LPP 投影。给定文档的权重向量 $\boldsymbol{S}$，LPP 投影的目标是使原始空间上相邻的数据点在投影空间上也保持相应的邻近关系，即相邻的点 $\tilde{x}_i$ 和 $\tilde{x}_j$ 经过 LPP 投影后的点 $y_i=\boldsymbol{\alpha}^{\mathrm{T}}\tilde{x}_i$ 和 $y_j=\boldsymbol{\alpha}^{\mathrm{T}}\tilde{x}_j$ 也是相邻的，而对映射后变得远的点进行惩罚，以便发现数据本身固有的内嵌流形结构。其形式化描述为

$$\arg\min_{\alpha}\sum_{i=1}^{n}\sum_{i=1}^{n}(\boldsymbol{\alpha}^{\mathrm{T}}\tilde{x}_i-\boldsymbol{\alpha}^{\mathrm{T}}\tilde{x}_j)^2S_{ij}=\arg\min_{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{L}\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\alpha} \tag{12}$$

约束条件为

$$\boldsymbol{\alpha}^{\mathrm{T}}\tilde{\boldsymbol{X}}\boldsymbol{D}\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\alpha}=1 \tag{13}$$

由拉普拉斯算子的性质可知，式(12)的解等价于计算下面方程的特征向量和特征值：

$$\tilde{\boldsymbol{X}}\boldsymbol{L}\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\alpha}=\lambda\tilde{\boldsymbol{X}}\boldsymbol{D}\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\alpha} \tag{14}$$

于是满足式(12)最小化的向量 $\boldsymbol{a}$ 可以通过求解式(14)的泛化特征向量问题获得。由于 $\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}$ 和 $\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}$ 都是对称的正半定矩阵，因此 LPP 比其他流形学习算法更加容易计算。

设 $\boldsymbol{W}_{LPP}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_k]$ 表示根据特征值 $\lambda_1<\lambda_2<\cdots<\lambda_k$ 的排列次序依次得到的式(14)的解，可得高维文档数据的低维嵌入：

$$x\rightarrow y=\boldsymbol{W}^{\mathrm{T}}\tilde{x} \tag{15}$$

其中，$\boldsymbol{W}=\boldsymbol{W}_{SVD}\boldsymbol{W}_{LPP}$。

Step5 利用优化训练的 SVM 进行分类。将原始文档空间 $\boldsymbol{X}$ 经过 LPP 投影后变成的低维空间 $Y=\boldsymbol{W}^{\mathrm{T}}\tilde{\boldsymbol{X}}$ 输入到 SVM 分类器中进行分类，其中 SVM 的训练方法采用乘性更新规则实现。

4 实验结果

为了测试本文算法的分类性能，将本文的文档分类算法(简称 LPP-SVM)与 KNN 方法、Bayes 方法和 SMO 算法[5]进行性能比较。LPP-SVM 分类算法中的参数设置为：核函数选用高斯核函数，宽度 $\sigma=0.5$，惩罚因子常数 $C=10$，在构造邻接图的近邻数中 $p=15$。对 SMO 算法和本文的 LPP-SVM 算法利用一对剩余方法(One-vs-Rest)进行多类分类[6]。

实验中的测试数据采用了文档分类领域标准测试集 Reuters-21578，遵循“ModApte”切分方式，选择文档最多的10个类别进行试验，其中包括7 194个训练文档和2 788个测试文档。文档表示采用向量空间模型，并利用TF-IDF方法进行词项加权。分类器性能评价指标使用常用的查准率(*Precision*)、查全率(*Recall*)及微平均 F_1 值，这3个评价指标的值越大，说明分类器的性能越好。它们的定义如下：

$$Precision=\frac{\text{正确分为某类的文档数}}{\text{测试集中分为该类别的文档总数}}\times 100\%$$

$$Recall=\frac{\text{正确分某的文档数}}{\text{测试集中属于该类别的文档总数}}\times 100\%$$

$$F_1=\frac{2\times Precision\times Recall}{Precision+Recall}\times 100\%$$

表 1 给出了这些分类方法在测试集上的实验结果，从中可以看出：本文 LPP-SVM 分类算法的查准率、查全率和 F_1 评价指标普遍高于常用的 KNN 方法、Bayes 方法和 SMO 分类方法。原因在于 LPP 能够较好地发现文档空间中的局部几何结构，并具有计算简单的优点，而利用乘性更新原则训练的 SVM 具有很快的训练速度和较好的分类效果，从而保证了 LPP-SVM 具有较高的分类准确率和较快的运行速度。

表 1　4 种分类方法的实验结果　(%)

分类方法	查准率	查全率	F_1
KNN	82.9	78.3	80.5
Bayes	82.1	81.7	81.9
SMO	87.2	84.1	85.6
LPP-SVM	89.6	86.3	87.9

表 2 给出了本文的 LPP-SVM 方法和 KNN 方法、Bayes 方法和 SMO 方法在不同文档大小下的运行时间比较。

表 2　4 种分类方法的运行时间　s

文档数	KNN	Bayes	SMO	LPP-SVM
1 000	13.39	12.68	9.23	8.26
2 000	24.37	21.85	15.96	12.74
3 000	40.96	35.58	26.39	21.36
4 000	59.84	43.38	35.68	31.27
5 000	75.36	59.63	48.96	42.58
6 000	98.84	75.69	67.98	51.28
7 000	120.39	108.63	97.36	59.69

从实验结果可以看出，本文分类算法 LPP-SVM 的运行时间低于 KNN 方法、Bayes 方法和 SMO 方法，并且随文档大小的增长幅度比较平缓，说明本文提出的分类算法在处理大规模数据集时具有较好的扩展性。原因在于首先利用了能保持高维文档空间中局部几何结构的 LPP 对高维文档进行降维处理，然后利用基于乘性更新原则的 SVM 优化训练算法进行分类，因此，LPP-SVM 具有较快的运行速度。

5　结束语

针对 Web 文档具有维数高、样本稀疏和特征不太明显的特点，本文提出了基于流形学习和 SVM 的文档分类算法，实验表明该算法是可行的，具有较高的分类准确率和较快的运行速度。

参考文献

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York, USA: Springer, 1995.

[2] He Xiaofei, Niyogi P. Locality Preserving Projections[C]//Proc. of Conf. on Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2003: 291-298.

[3] Cai Deng, He Xiaofei, Han Jiawei. Document Clustering Using Locality Preserving Indexing[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(12): 1624-1637.

[4] Sha Fei, Saul L K, Lee D D. Multiplicative Updates for Nonnegative Quadratic Programming in Support Vector Machines[C]//Proc. of Conf. on Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2002:1041-1048.

[5] Platt J. Fast Training of Support Vector Machines Using Sequential Minimal Optimization[C]//Proc. of Conf. on Advances in Kernel Methods-Support Vector Learning. Cambridge, USA: MIT Press, 1999: 185-208.

[6] 衣治安，吕　曼. 基于多分类支持向量机的入侵检测方法[J]. 计算机工程, 2007, 33(15): 167-169.

编辑　张正兴

（上接第 37 页）

果。可以看到，FLSI 的效果接近 LSI 的效果，且明显好于 DF 和 AW 算法。另一方面，在 FLSI 处理过的数据上再运行 LSI 的时间只有在原数据上运行 LSI 的 20%到 30%，节省了计算资源。

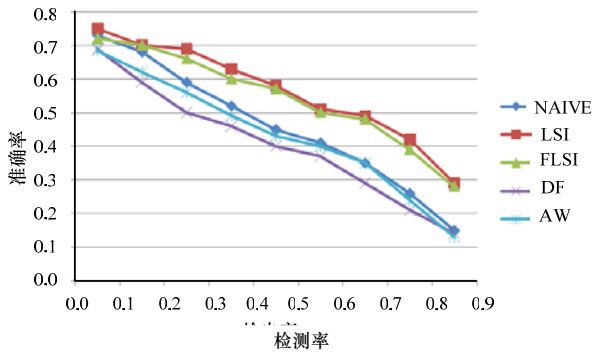

图 3　查询结果对比

5　结束语

潜在语义索引技术在应用中的最大障碍就是其中矩阵奇异值分解的高复杂度。本文针对这一问题提出了一种有效的解决方案。把特征提取和特征选择这 2 大类降维方法统一到一个形式化框架中，从而可以把 LSI 这种特征提取算法通过解空间的变化转化为一种特制提取算法。这样就降低了 LSI 的计算复杂度。同时实验结果表明这种简化并没有明显降低 LSI 的效果。

在下一步的工作中，将把这一算法应用到真正的百万级的文本数据上。此外，将尝试特征提取算法和特征选择算法之间的互相转化，以期发现更多更有用的新算法。

参考文献

[1] Scott C D, Dumais S T, Thomas K L, et al. Indexing by Latent Semantic Analysis[J]. Journal of the American Society for Information Sciences, 1990, 41(6): 391-407.

[2] 何　明，冯博琴，傅向华，等. 基于 Rough 集潜在语义索引的 Web 文档分类[J]. 计算机工程, 2004, 30(13): 3-5.

[3] Tang Chunqiang, Dwarkadas S, Xu Zhichen. On Scaling Latent Semantic Indexing for Large Peer-to-Peer Systems[C]//Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval. NY, USA: ACM Press, 2004: 112-121.

[4] Kolda T G, O'Leary D P. A Semidiscrete Matrix Decomposition for Latent Semantic Indexing Information Retrieval[J]. ACM Trans. on Inf. Syst., 1998, 16(4): 322-346.

[5] Karypis G, Han E H S. Concept Indexing: A Fast Dimensionality Reduction Algorithm with Application to Document Retrieval and Categorization[C]//Proceedings of CIKM'00. McLean, VA, USA: [s. n.], 2000: 12-19.

[6] Bingham E, Mannila H. Random Projection in Dimensionality Reduction: Applications to Image and Text Data[C]//Proceedings of KDD'01. San Francisco, CA, USA: [s. n.], 2001: 245-250.

编辑　索书志