文章编号:1001 - 9081(2009)02 - 0416 - 03

基于 KDA 和 SVM 的文档分类算法

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)
(wzqbox@ yahoo. cn)

摘 要:为了高效地解决 Web 文档分类问题,提出了一种基于核鉴别分析方法 KDA 和 SVM 的文档分类算法。该算法首先利用 KDA 对训练集中的高维 Web 文档空间进行降维,然后在降维后的低维特征空间中利用乘性更新规则优化的 SVM 进行分类预测。采用了文档分类领域两个著名的数据集 Reuters-21578 和 20-Newsgroup 进行实验,实验结果表明该算法不仅获得了更高的分类准确率,而且具有较少的运行时间。

关键词:文档分类;核鉴别分析;支持向量机;数据挖掘
中图分类号: TP181 **文献标志码**:A

Document classification algorithm based on KDA and SVM

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

Abstract: To efficiently solve Web document classification problem, a novel document classification algorithm based on kernel discriminant analysis (KDA) and SVM was proposed. The proposed algorithm firstly reduced the high dimensional Web document space in the training sets to the lower dimensional space with KDA algorithm, then the classification and predication in the lower dimensional feature space were implemented with the multiplicative update-based optimal SVM. The experimental evaluations were performed on the Reuters-21578 and 20-Newsgroup which are two well-known data sets in the field of document classification. Experimental results show that the proposed algorithm not only achieves higher classification accuracy, but also has lower running time.

Key words: document classification; Kernel Discriminant Analysis (KDA); Support Vector Machine (SVM); data mining

0 引言

为了高效地管理和组织海量信息[1],以帮助用户快速准确地查询到所需要的相关文档,基于内容的 Web 文档自动分类技术日益成为数据挖掘领域备受关注的热点研究问题。

Web 文档分类的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。有关学者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(SVM)[2]等。虽然这些方法在进行文档分类时取得了一定的效果,但是他们大多都是直接在原始文档空间分类,而 Web 文档具有维数高、样本稀疏和特征不太明显的特点。考虑到高维文档空间引起的"维数灾难"问题,我们应当首先把高维文档投影到低维特征子空间,然后在降维后的低维特征空间中利用分类器进行分类预测。基于上述考虑,本文提出了基于核鉴别分析方法(Kernel Discriminant Analysis,KDA)和 SVM 的 Web 文档分类算法,该算法首先使用核鉴别分析方法把高维文档投影到低维特征空间,然后在降维后的低维特征空间利用乘性更新规则优化[3]的 SVM 进行分类。实验结果表明该方法有效地提高了文档分类准确率,而且减少了分类器的训练时间。

1 基于 KDA 和 SVM 的文档分类

1.1 基于 KDA 的文档低维空间获取

设给定 n 个文档样本集 $X = \{x_1, x_2, \cdots, x_n\}$ 属于 c 个不同的类别,其中第 $i(1 \leqslant i \leqslant c)$ 个类别中有 n_i 个样本。核鉴别分析 KDA[4-5]的主要思想是通过构造非线性映射 φ 把输入空间 X 映射到特征空间 F:

$$\varphi: x_i \in X \rightarrow \varphi(x_i) \in F \tag{1}$$

以便在特征空间 F 中实现线性鉴别分析 LDA[6]的优化目标:最小化类内距离的同时最大化类间的距离,以达到最大限度的类鉴别分析。其形式化描述为:

$$J(\boldsymbol{w}) = \max_{\boldsymbol{w}} \frac{\boldsymbol{w}^{\mathrm{T}} S_B^{\varphi} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} S_W^{\varphi} \boldsymbol{w}} \tag{2}$$

其中,S_B^{φ} 和 S_W^{φ} 分别表示在特征空间 F 中的类间散度和类内散度,其定义如下:

$$S_B^{\varphi} = \sum_{i=1}^{c} n_i (\mu_i^{\varphi} - \mu^{\varphi})(\mu_i^{\varphi} - \mu^{\varphi})^{\mathrm{T}} \tag{3}$$

$$S_W^{\varphi} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\varphi(x_j^i) - \mu_i^{\varphi})(\varphi(x_j^i) - \mu_i^{\varphi})^{\mathrm{T}} \tag{4}$$

$$\mu_i^{\varphi} = \frac{1}{n_i} \sum_{j=1}^{n_i} \varphi(x_j^i) \tag{5}$$

$$\mu^{\varphi} = \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \tag{6}$$

其中,μ_i^{φ} 和 μ^{φ} 分别表示在特征空间 F 中的第 i 个类别中的样本均值和所有样本的均值。

根据再生核空间理论,任何解向量 $\boldsymbol{w} \in F$ 都是由空间 F 中的所有样本线张成的。可得:

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \tag{7}$$

将式(7)代入式(2)可得:

收稿日期:2008 - 08 - 12;**修回日期**:2008 - 10 - 15。 **基金项目**:教育部科学技术研究重点资助项目(107021)。

作者简介:王自强(1973 -),男,河南郑州人,博士研究生,主要研究方向:数据挖掘、模式识别; 钱旭(1962 -),男,江苏无锡人,教授,博士生导师,博士,主要研究方向:数据挖掘、信息融合。

$$J(\boldsymbol{\alpha}) = \max_{\alpha} \frac{\boldsymbol{\alpha}^{\mathrm{T}} K_B \boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}} K_W \boldsymbol{\alpha}} \tag{8}$$

其中:

$$K_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^{\mathrm{T}} \tag{9}$$

$$K_W = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\eta_j - m_i)(\eta_j - m_i)^{\mathrm{T}} \tag{10}$$

$$m_i = \left(\frac{1}{n_i} \sum_{j=1}^{n_i} k(x_1, x_j), \frac{1}{n_i} \sum_{j=1}^{n_i} k(x_2, x_j), \cdots, \frac{1}{n_i} \sum_{j=1}^{n_i} k(x_n, x_j) \right)^{\mathrm{T}} \tag{11}$$

$$m = \left(\frac{1}{n} \sum_{j=1}^{n} k(x_1, x_j), \frac{1}{n} \sum_{j=1}^{n} k(x_2, x_j), \cdots, \frac{1}{n} \sum_{j=1}^{n} k(x_n, x_j) \right)^{\mathrm{T}} \tag{12}$$

$$\eta_j = \left(k(x_1, x_j), k(x_2, x_j), \cdots, k(x_n, x_j) \right)^{\mathrm{T}} \tag{13}$$

$$k(x_i, x_j) = [\varphi(x_i), \varphi(x_j)] \tag{14}$$

因此,在 K_W 非奇异的情况下,式(8)的解向量 $\boldsymbol{\alpha}$ 是由 $K_W^{-1} K_B$ 的前 $l\left(l = \min(c-1, n)\right)$ 个最大特征向量组成的。

于是,对于一个新的测试样本点 $x \in X$,它在特征空间 F 中特征向量 $\boldsymbol{w}$ 上的投影可表示为:

$$\boldsymbol{w} \cdot \varphi(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) \tag{15}$$

对于文档分类问题,由于训练样本的维数远远大于样本数,因此矩阵 K_W 是奇异的,为了克服这一不足,我们用 $(K_W + \varepsilon \boldsymbol{E})$ 来代替 K_W 来解决这一问题,其中 ε 是一个非常小的正常数,本文中我们取 $\varepsilon = 10^{-5}$,$\boldsymbol{E}$ 是单位矩阵。

1.2　SVM 分类器

在利用 KDA 完成高维文档的降维处理后,为了高效地对降维后的低维特征空间进行分类,我们采用了具有良好泛化能力的支持向量机(SVM)[2]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(16)找到一个具有最大间隔的超平面:

$$\min_{w,b,\xi} \frac{1}{2} w^{\mathrm{T}} w + C \sum_{i=1}^{l} \xi_i \tag{16}$$

约束条件:

$$y_i[(\boldsymbol{w} \cdot x + b) + \xi_i - 1] \geqslant 0, \xi_i \geqslant 0 \tag{17}$$

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(16)转化为其对偶形式:

$$\max W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{18}$$

约束条件为:

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \alpha_i \in [0, C], i = 1, \cdots, l \tag{19}$$

于是对于未知类别的数据点 x,可采用如下线形判决函数决定其所属类别:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i (x_i \cdot x) + b \right) \tag{20}$$

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 $K(x_i, x_j) = \left(\varphi(x_i) \cdot \varphi(x_j)\right)$ 可在特征空间计算,而无需知道映射 φ 的具体形式。

用核函数代替线形 SVM 中的点积形式,于是式(18)的对偶形式可变为:

$$\max W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{21}$$

约束条件为:

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \alpha_i \in [0, C], i = 1, \cdots, l \tag{22}$$

于是,非线形 SVM 的判决函数变为:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \right) \tag{23}$$

由于 SVM 训练问题可归结为求解非负二次规划问题,常用的 SVM 训练方法存在如下不足:训练代价太大,在处理大规模高维数据时存在着收敛速度慢、分类准确率低。为了有效地提高 SVM 训练算法的性能,本文采用乘性更新规则方法[3]来优化求解 SVM 中的二次规划问题。该方法的优点是:无需参数设置、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。下面给出基于乘性更新方法的 SVM 训练方法。

由于式(18)的非负二次规划问题可归纳为如下标准形式:

$$\min F(v) = \frac{1}{2} \boldsymbol{v}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{v} + \boldsymbol{b}^{\mathrm{T}} \boldsymbol{v} \tag{24}$$

约束条件为:$\forall i, v_i \geqslant 0$。矩阵 $\boldsymbol{A}$ 是正定和半对称的。由于对 $F(v)$ 的优化为凸二次优化,加上非负 v_i 的约束,因此只能采用迭代的方法而不能采用分析的方法来获得全局最小值。

该迭代方法是通过矩阵 $\boldsymbol{A}$ 的正负元素来实现,首先在式(24)中,定义如下两个非负矩阵 $\boldsymbol{A}^+$ 和 $\boldsymbol{A}^-$:

$$A_{ij}^{+} = \begin{cases} A_{ij}, & A_{ij} > 0 \\ 0, & \text{其他} \end{cases}, A_{ij}^{-} = \begin{cases} |A_{ij}|, & A_{ij} < 0 \\ 0, & \text{其他} \end{cases} \tag{25}$$

于是有:$\boldsymbol{A} = \boldsymbol{A}^+ - \boldsymbol{A}^-$。根据以上定义的非负矩阵,迭代乘性更新规则定义如下:

$$v_i \leftarrow v_i \left[\frac{-b_i + \sqrt{b_i^2 + 4(\boldsymbol{A}^+ \boldsymbol{v})_i (\boldsymbol{A}^- \boldsymbol{v})_i}}{2(\boldsymbol{A}^+ \boldsymbol{v})_i} \right] \tag{26}$$

上述乘性更新规则的优点是实现简单,避免了常规 SVM 训练中的大量矩阵运算,且整个迭代更新过程不会违反非负性条件约束。而文献[3]中的定理 1 已证明:利用式(26)中的乘性更新规则,式(24)中的目标函数可以单调递减到全局最小值。因而保证了式(26)的更新迭代规则可以收敛到全局最小值。

对于 SVM 训练中的式(21)优化问题的具体实现,令 $A_{ij} = y_i y_j K(x_i, x_j)$,$b_i = -1$。于是,优化求解式(20)的乘性更新规则如下:

$$\alpha_i \leftarrow \alpha_i \left[\frac{1 + \sqrt{1 + 4(A^+ \alpha)_i (A^- \alpha)_i}}{2(A^+ \alpha)_i} \right] \tag{27}$$

由此可见,利用式(27)来求解 SVM 中的二次规划问题,避免了常规 SVM 训练中非常耗时的大量矩阵运算,使得训练速度得到很大提高,后面的实验结果有力地证明了这一点。

1.3　基于 KDA 和 SVM 的文档分类算法

设给定 n 个文档样本集 $X = \{x_1, x_2, \cdots, x_n\}$ 属于 c 个不同的类别。由于向量空间模型(Vector Space Model,VSM)是文档表示的主要形式,因此本文采用 VSM 中常用的 TF-IDF 项权重向量来表示每个文档 x_i。基于 KDA 和 SVM 的文档分类算法的主要思想为:首先利用核鉴别分析 KDA 把高维文档投影到低维特征空间,然后利用乘性更新规则优化的 SVM 分类器对降维的文档数据进行分类。算法的具体实现过程如下:

步骤 1　利用向量空间模型 VSM 中的 TF-IDF 项权重向量来表示每个文档 x_i。

步骤 2　通过求解式(8)获得最佳投影方向。

步骤 3　对于新的测试文档样本,利用式(15)计算出它在特征空间 F 中特征向量上的投影。

步骤 4　利用投影后的低维文档特征向量按照式(27)的乘性更新规则来优化训练式(21)中的 SVM 目标函数。

步骤5 利用式(23)中的SVM判决函数来确定待测试文档所属的类别。

2 实验结果

为了测试本文提出的基于KDA和SVM的文档分类算法(简称KDA-SVM)分类性能，我们将KDA-SVM与其他4种常用的文档分类算法：最近邻居法(K-Nearest Neighbor，KNN)、线性鉴别分析法(Linear Discriminant Analysis，LDA)、Bayes方法和经典SVM算法进行了比较。算法中的核函数采用常用的RBF核函数，模型参数采用5次交叉验证法来确定。在利用SVM进行分类时，对于超过两个类别的情况，采用一对剩余法(One-vs-Rest)[7]进行多类分类。

实验中的测试数据采用了文档分类领域两个著名的测试集Reuters-21578和20-Newsgroup。对于Reuters-21578数据集[8]，我们遵循常用的"ModApte"切分方式，选择文档最多的10个类别进行实验，其中包括了7 194个训练文档和2 788个测试文档。20-Newsgroup数据集[9]包含大致均匀地分布在20个类别上的18 846个文档，我们采用每个类别中的一半数据作为训练集，另一半作为测试集。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，这三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：

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

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

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

表1和表2分别给出了5种分类方法在测试集Reuters-21578和20-Newsgroup上的实验结果，从中可以看出：本文提出的KDA-SVM分类算法的查准率、查全率和F_1评价指标普遍高于KNN、LDA、Bayes和SVM方法。原因在于KDA-SVM方法首先使用核鉴别分析法KDA对训练集中的高维文档空间进行非线性降维，然后在降维后的低维特征空间中利用优化的SVM进行分类预测。由于核鉴别分析KDA有效地减少了文档特征空间的维数，从而大大地提高了文档分类器SVM的计算效率。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KDA-SVM具有较高的分类准确率和较快的运行速度。

表1 在Reuters-21578上的分类性能比较 %

分类方法	查准率	查全率	F_1
KNN	82.9	78.3	80.5
LDA	84.2	79.5	81.8
Bayes	85.6	81.4	83.4
SVM	87.3	85.7	86.5
KDA-SVM	90.5	87.2	88.8

表2 在20-Newsgroup上的分类性能比较 %

分类方法	查准率	查全率	F_1
KNN	78.6	72.5	75.4
LDA	81.3	77.8	79.5
Bayes	89.5	80.3	84.7
SVM	93.2	85.6	89.2
KDA-SVM	96.4	87.8	91.9

另外，我们还对这5种分类算法在不同文档大小下的运行时间进行了比较。从图中的实验结果可以看出本文提出的分类算法KDA-SVM的运行时间都低于KNN、LDA、Bayes和SVM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。原因在于我们采用的核鉴别分析法KDA有效地减少了文档特征空间的维数，避免了高维文档引起的"维数灾难"问题。另外，采用乘性更新规则训练的优化SVM避免了常规SVM训练中非常耗时的大量矩阵运算，使得训练速度得到很到提高，从而保证了KDA-SVM具有很快的运行速度。

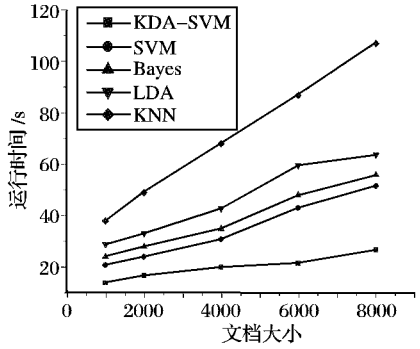

图1 在Reuters-21578上的运行时间比较

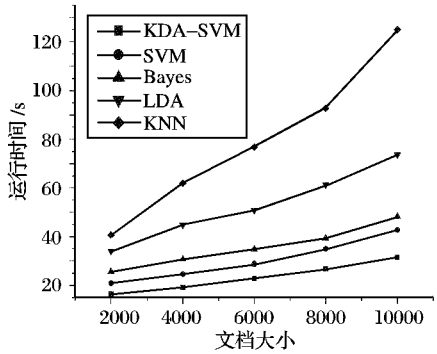

图2 在20-Newsgroup上的运行时间比较

3 结语

为了有效地解决高维文档分类时的"维数灾难"问题，本文提出了基于核鉴别分析方法KDA和SVM的Web文档分类算法。实验结果表明，该算法有效地提高了文档分类的准确率，而且具有很快的运行速度。

参考文献：

[1] SEBASTIANI F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[2] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer, 1995.

[3] SHA F, LIN Y Q, SAUL L K, LEE D D. Multiplicative updates for nonnegative quadratic programming[J]. Neural Computation, 2007, 19(8): 2004-2031.

[4] BAUDAT G, ANOUAR F. Generalized discriminant analysis using a kernel approach[J]. Neural Computation, 2000, 12(10): 2385-2404.

[5] MIKA S, RATSCH G, WESTON J, *et al*. Fisher discriminant analysis with kernels[C]//Proceedings of the 1999 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing. NewYork: IEEE Press, 1999: 41-48.

[6] DUDA R O, HART P E, STORK D G. Pattern classification [M]. Second edition. Hoboken: Wiley-Interscience, 2000.

[7] HSU C W, LIN C J. A comparison on methods for multi-class support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.

[8] LEWIS D D. Reuters-21578 text categorization collection[EB/OL]. [2008-06-22]. http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.

[9] 20 News Group[EB/OL]. [2008-06-22]. http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.htm.