第14卷 第36期 2014年12月
1671—1815(2014)36-0092-06

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 14 No. 36 Dec. 2014

基于投影寻踪的kNN文本分类算法的加速策略

张 永 孟晓飞
(兰州理工大学计算机与通信学院,兰州 730050)

摘 要 传统的 k 近邻(k-nearest neighbors,kNN)文本分类中,由于文本被表示成向量空间模型后维数非常高,且训练文本的数目巨大,kNN分类算法通常被视为是一种虽然有效,但并非高效的文本分类算法。针对传统kNN分类算法效率低下的问题,提出了一种基于投影寻踪思想的kNN分类算法加速策略。基本思想是:通过投影的方法缩减训练集的规模,同时在寻找 k 近邻过程中对文本进行降维处理,从两方面着手降低算法的计算开销。实验数据表明,优化后的kNN算法比传统kNN算法在时间性能上有较大的提升,同时保证了分类的精度。

关键词 kNN 文本分类 投影寻踪 降维 训练集缩减
中图法分类号 TP391.3; **文献标志码** A

目前文本分类方法主要包括 k 近邻(kNN)、支持向量机(SVM)、决策树、关联规则、贝叶斯等传统的分类算法和基于软计算的模糊逻辑、神经网络、粗糙集和遗传算法等。

由于kNN算法[1]思想简单、分类效果好,在文本分类中有着广泛的应用。使用kNN算法时,首先要计算测试文本与所有训练文本的相似度,根据文本间的相似度,寻找出测试文本的 k 个近邻文本。因此,kNN算法存在缺陷:处理相似性计算的计算开销巨大,尤其当特征空间的维数特别高、训练集非常大的时候。

投影寻踪(projeetion pursuit,PP)[2]是一种探索性数据分析方法。它通过某种组合,将高维数据投影到低维(一至三维)子空间上,并通过极大(小)化某个投影指标,寻找出能反映原高维数据结构或特征的投影,在低维空间上对数据结构进行分析,以达到研究和分析高维数据的目的。

以往投影寻踪在文本分类的应用中,主要是通过数据分析的方法对文本特征进行降维处理。本文提出的优化策略在特征降维的基础上,基于一个简单的原理:空间中邻近的点映射到某一方向上也是邻近的。通过投影的方法对原始训练集进行缩减,从而大大降低了计算开销。从两方面着手对kNN文本分类算法进行加速优化。

首先,通过遗传算法挑选出若干个最优的投影方向;其次,把文本投影到各一维的投影方向上,寻找出各方向上测试文本可能的近邻文本,由各方向所有可能的近邻文本组成最终的查询集,从而缩减训练集的规模;最后,在低维的投影空间中计算测试文本的k近邻,从而达到特征降维的目的。实验表明,通过选择适当的参数,优化后的kNN算法在保证分类精度的同时,有效提高了算法的效率。

2014年8月16日收到
第一作者简介:张 永(1963—),男,甘肃陇南人,硕士生导师,教授。研究方向:智能信息处理,数据挖掘,图像识别与处理等。E-mail:mengxiang3721@163.com。

1 相关研究

如何降低kNN算法的计算复杂度,目前已经提出了众多改进方法,这些方法可以归为以下三类:

1.1 基于特征降维的改进

为了清理噪声数据,降低计算开销,需要对样本进行特征降维,特征降维方法分为特征选择和特征抽取。特征选择的关键是选择特征评估函数,目前比较常用的特征评估函数包括:文档频率(DF)、互信息(MI)、信息增益(IG)、统计(CHI)等。Gupta等[3]使用粗糙集方法为文本分类进行特征选择;宋枫溪等[4]提出了低损降维,其降维效果与互信息、χ^2 统计相当,并优于文档频率。

常用的特征抽取方法包括:主成份分析(PCA)、Fishe线性判别分析、潜在语义分析(LSA)和投影寻踪(PP)等。钟将等[5]使用潜在语义分析方法对文本特征空间进行降维处理。廖海波等[6]通过投影寻踪的方法,把高维空间的数据投影到低维空间,进行特征维数约简。

1.2 基于缩减训练样本的改进。

较小的训练集意味着更少的相似度计算开销。优化方法分为两类: 一是对训练文本进行剪裁以减小训练集规模; 二是在原训练集中选取或生成一些代表样本,代替原训练样本。李荣陆等[7]针对类偏斜问题提出基于密度的剪裁方法。周勇等[8]提出一种基于聚类中心改进的 kNN 算法,通过聚类算法把训练集分簇,计算每个簇的中心点,代表本簇的训练样本。Jiang J Y[9]通过模糊相似度计算把样本分簇,寻找测试样本的 k 近邻时,只在与其隶属度大于某一阈值的簇中进行计算。

1.3　基于 k 近邻搜索方法的改进。

此类方法通过加快寻找 k 近邻的速度来对 kNN 算法进行优化。Guo 等[10]通过在训练集上构建多个类似索引结构的 kNN 模型簇,提高算法效率。邓箴等[11]利用模拟退火改进 kNN 算法中查找 k 近邻的过程并取得良好效果。

空间中邻近的点映射到某个方向上也是邻近的。因此,当寻找某一测试样本点的 k 近邻点时,大多数无关的点就可以被忽略掉,这大大降低了计算开销。正是利用这一点对传统 kNN 算法进行优化。

2　基于投影寻踪的 kNN 算法的加速策略

2.1　kNN 文本分类算法简介

kNN 算法对一篇测试文本 d_x 进行分类的过程是: 首先计算 d_x 与训练文本集中每个文本的相似度,相似度一般采用余弦相似度度量,依据相似度找到最相似的 k 个训练文本,把这 k 个文本的类别作为 d_x 的候选类别; 然后把相似度作为 k 近邻文本所属类别的权重,将各类中近邻文本的类权重之和作为该类别和测试文本的相似度,再把测试文本 d_x 归到权重最大的类别中。

测试文本 d_x 和训练文本 d_i 的相似度计算公式为

$$\mathrm{sim}(d_x,d_i)=\frac{\sum_{k=1}^{M}w_{xk}w_{ik}}{\sqrt{\sum_{k=1}^{M}w_{xk}^{\ 2}}\sqrt{\sum_{k=1}^{M}w_{ik}^{\ 2}}} \tag{1}$$

式(1)中,M 为特征向量维数,w 为特征权重。

测试文本 d_x 和类 c_j 的权重计算公式为

$$p(d_x,c_j)=\sum_{d_i\in kNN}\mathrm{sim}(d_i,d_x)\,y(d_i,c_j) \tag{2}$$

式(2)中,$y(d_i,c_j)$ 是类别属性函数,当 d_i 属于 c_j 时,$y(d_i,c_j)=1$; 当 d_i 不属于 c_j 时,$y(d_i,c_j)=0$。

分类决策函数为

$$f=\mathrm{argmax}c_j\,[p(d_x,c_j)] \tag{3}$$

2.2　投影寻踪及其算法描述

投影寻踪最早是由 Kruskal 于 20 世纪 70 年代初提出的,是用来分析和处理高维观测数据,尤其是非正态非线性高维数据的一种新兴统计方法。投影寻踪模型中三个基本的概念: 线性投影、投影指标和最优化投影方向。

2.2.1　线性投影

线性投影是对高维数据进行投影降维的手段。利用线性投影将 M 维向量空间中的数据映射到 m 维投影空间后,在投影空间中,数据点的个数不变,但维数由 M 维降低为 m 维,投影寻踪方法正是利用线性投影研究数据在低维空间的散布特征,从而找到其在高维空间的结构特征。

$$Z=\frac{da}{\|a\|}=\frac{\sum w_ia_i}{\sum a_i^{\ 2}};\ i=1,2,\cdots,M \tag{4}$$

式(4)中,d 代表文本特征向量,a 代表投影的方向向量。

2.2.2　投影指标

投影指标是用于衡量投影到低维空间上的数据是否有意义的目标函数。

随机变量 X 在投影方向 A 上的投影指标表示为 $Q(F_A)$,实际上 Q 是一个 m 维空间上的泛函,将空间函数转变成某一确定的数值,也可表示称 $Q(AX)$。在使用优化算法优化投影指标时,投影指标即为目标函数,其具体形式视不同需求而定。

在文本分类中,希望: 局部上,投影点要密集,凝聚成若干个类; 整体上,投影点要散开,类与类之间要尽可能的分离。

本文参照文献[12]构造投影指标如下

$$Q(Z)=B(Z)/W(Z) \tag{5}$$

以各类一维投影数据均值作为类中心的度量,以标准差作为类内离差的度量。求得 $\max Q(Z)$,使得投影点局部密集,整体分散。

中心离差: $B(Z)=|E(Z^{(1)})-E(Z^{(2)})|$ (6)

式(6)中,$E(Z^{(i)})$ 是第 $i(i=1,2)$ 类投影的均值。

类内离差: $D(Z)=$

$$\sqrt{\left\{\frac{\sum_{j=1}^{n_1}[Z_j^{(1)}-E(Z^{(1)})]^2+\sum_{j=1}^{n_2}[Z_j^{(2)}-E(Z^{(2)})]^2}{n_1+n_2}\right\}} \tag{7}$$

式(7)中,$Z_j^{(i)}$ 是第 i 类$(i=1,2)$ 对应文本的投影值,$n_1+n_2=n$,n 为训练文本数。

需要注意的是: 训练是针对单个类别进行的,也就是说,把属于该类的文本看成正例,其他文本全部看作为负例,为各类计算出一个最优的投影方向。

2.2.3 投影方向

原始空间中有无数的单位向量，但并不是所有方向都适合作为投影方向。我们应该选择那些能够有效代表原始空间的方向，尽可能的保留原始信息。

2.3 分类工作流程

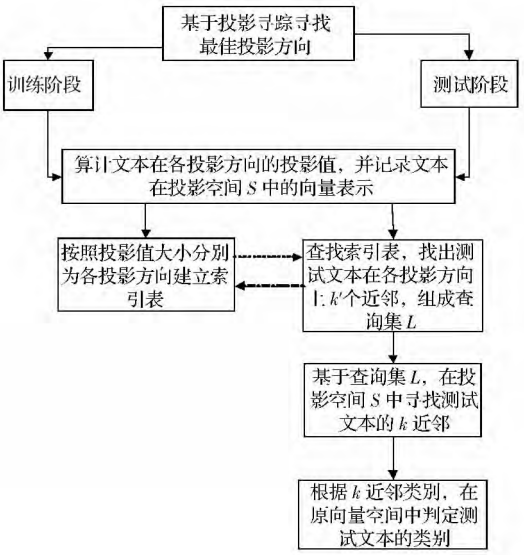

图 1 分类流程

Fig. 1 The classification process

不同的投影方向反映着不同的数据结构特征，所谓最佳投影方向是能够最大可能体现高维数据的某类特征结构的方向，从信息论角度而言，最佳投影方向是对数据信息利用最充分，信息损失量最小的方向，优化投影方向归根到底是找出某种意义下好的投影指标。

算法 1 确定最优投影方向

根据定义的投影指标 $Q(Z)$，采用遗传算法优化投影方向 a，算法步骤如下:

(1) M 维原向量空间中，随机选出 m 组(m 为训练集中类别数) 初始投影方向 $\boldsymbol{a}_j(1 \leqslant j \leqslant m)$，$\boldsymbol{a}_j = (a_{j1}, a_{j2}, \cdots, a_{jM})$，$\| \boldsymbol{a}_j \| = 1$。

(2) 根据公式(4) 分别计算出 m 组投影方向上 n 个文本的投影值 $Z_i(i = 1, 2, \cdots, n)$。

(3) 根据公式(6) 计算中心离差 $B(Z)$，公式(7) 计算类内离差 $D(Z)$，根据公式(5) 分别计算 m 组投影方向对应的投影指标 $Q(Z)$。

(4) 将各投影方向按照其对应的投影指标 $Q(Z)$ 升序排列，选取前 $m/2$ 组投影方向进行交叉操作，前 $0.01m$ 组投影方向进行变异操作。得到 $m + m/2 + 0.01m$ 组投影方向，计算新产生投影方向的投影指标，并按升序排序。

(5) 取步骤(4) 中得出的前 m 组投影方向，从骤(2) 开始循环计算。

(6) 直到投影指标不再增加时，算法收敛，从中选出投影指标最大的方向即为最佳投影方向。

得出最优投影方向后，把各文本向其作投影，在各一维投影方向上计算测试文本可能的近邻，并最终得到缩减后的查询集。优化的 kNN 算法由算法 2 给出。

算法 2 优化的 kNN 分类算法

(1) 使用算法 1 取得最优的 m 个投影方向。

(2) 训练文本 $d_i(1 \leqslant i \leqslant n)$ 向 m 个最优投影方向分别作投影，并根据公式(1) 计算文本 $d_i(1 \leqslant i \leqslant n)$ 在各最优投影方向上的投影值 $Z_1^{(i)}, Z_2^{(i)}, \cdots, Z_m^{(i)}$，其对应在投影空间 S 中可表示成 $(Z_1^{(i)}, Z_2^{(i)}, \cdots, Z_m^{(i)})$。

(3) 为每个投影方向建立一个索引表。索引表中的数据按照文本在此方向上的投影值升序排列。

(4) 同理，计算测试文本 d_x 在各最优投影方向上投影值。

(5) 查找各投影方向的索引表，使用二分查找算法，找出测试文本 d_x 对应各方向上 k' 个近邻。例如第 $j(1 \leqslant j \leqslant m)$ 个方向上 k' 近邻 $L_j = \{d_{j1}, d_{j2}, \cdots, d_{jk'}\}$。总共 m 个投影方向，所以测试文本 d_x 所有可能近邻集合为 L，$L = \bigcup_{j=1}^{m} L_j$，$|L| \leqslant mk'$。

(6) 在查询集 L 上执行 kNN 分类算法。首先，在投影空间 S 中寻找测试文本 d_x 的 k 个近邻。其次，在原向量空间中根据找到的 k 个近邻的类别，使用式(2)、式(3) 判定 d_x 的类别。

3 实验结果与分析

实验采用中科院 ICTCLAS 分词系统对文本进行分词处理，采用 χ^2 特征选择选出 500 个特征词，采用优化后的 kNN 分类器对测试文本集进行分类。

3.1 数据集及评价指标

采用复旦大学中文语料库作为实验数据集。语料库包含 9 个类别，如表 1 所示。

表 1 实验数据

Table 1 Experimental data

类别	训练文本数	测试文本数
C_1 - 环境	480	150
C_2 - 计算机	500	140
C_3 - 艺术	440	120
C_4 - 历史	460	130
C_5 - 教育	450	120
C_6 - 经济	520	150
C_7 - 政治	490	130
C_8 - 军事	460	130
C_9 - 体育	500	130
总计	4 300	1 200

对文本分类效果进行评估标准有查准率,查全率。两者反映了分类质量的两个不同方面。

$$\text{查准率} = \frac{\text{分类正确的文本数}}{\text{实际分类的文本数}},$$

$$\text{召回率} = \frac{\text{分类正确的文本数}}{\text{应有文本数}};$$

$F1$ 值是考察查准率和召回率的综合指标;

$$F1 = \frac{\text{查准率} \times \text{召回率} \times 2}{\text{查准率} + \text{召回率}}。$$

Macro-$F1$ 表示 m 个类别 $F1$ 值得平均值。

3.2　参数设置及算法复杂度分析

为了实验描述方便,经本文优化后的算法为 PPkNN 算法,该算法中 k' 的选取尤为关键,k' 选取的过小会造成近邻点的漏选,k' 选取的过大会使得分类时间过长,当 k' 增加到一定程度时,PPkNN 算法的精度与 kNN 分类精度相同。图 2、图 3 分别验证了取不同 k、k' 值时,算法的 Macro-$F1$ 值及算法执行时间的变化情况。

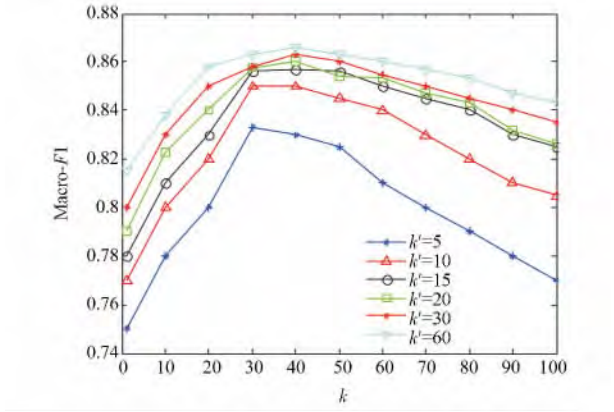

图 2　PPkNN 算法的 Macro-$F1$ 值

Fig. 2　Macro-$F1$ of PPkNN

由图 2 可知,k' 越大代表原向量空间中越多真正的近邻文本被包含在集合 L 中,所以算法的 Macro-$F1$ 值会随着 k' 的递增而递增。

与传统 kNN 算法类似,算法随着 k 的递增,Macro-$F1$ 会先增加后降低,而随着 k' 的增大,这种变化趋势会越来越缓慢。实验数据显示,当 $k' \geqslant 60$ 时,PPkNN 算法的 Macro-$F1$ 值近似于传统 kNN 的 Macro-$F1$ 值。

设 $t(x)$ 表示一组 x 维的样本相似度计算的时间开销。则传统 kNN 算法时间复杂度 $O(\text{kNN}) = nt(M)$,n 表示训练文本数。M 表示原向量空间中文本的维度。

改进后 PPkNN 算法的时间复杂度 $O(\text{PkNN}) = mk't(m) + kt(M)$,$mk'$ 表示 m 个方向上测试文本的所有可能的近邻数。等式右侧前半部分表示在投影空间中寻找测试样本的 mk' 个可能近邻所需的时间开销; 等式右侧后半部分则表示在原向量空间中,使用公式(2)判定测试文本类别的时间开销。

由于 $k, m, k' << n, M$,所以优化后 PPkNN 算法的时间复杂度要远小于传统 kNN 算法。

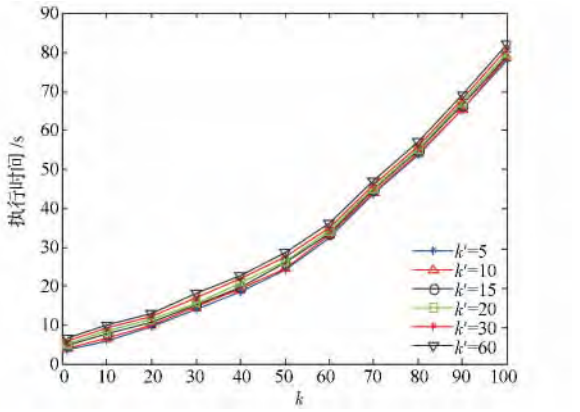

图 3　PPkNN 算法的执行时间

Fig. 3　Execution time of PPkNN

由图 3 可以看出,算法执行时间随着 k 的递增而急剧增加; 此外之外算法执行时间还会轻微收到 k' 的影响。由于 $t(m) << t(M)$,所以算法执行时间主要还是依赖于参数 k。

3.3　实验对比及分析

实验对比了传统 kNN 分类算法及两种改进的加速 kNN 算法与本文提出的优化算法的 Macro-$F1$ 值和执行时间两项性能指标。两种改进算法分别是: 样本裁剪 kNN 算法[7] 和 LSI 降维 kNN 算法[5]。样本裁剪 kNN 算法的加速思想是: 对类中心区域高密度文本进行裁剪,降低区域密度,从而缩减训练集。LSI 降维 kNN 算法则是通过潜在语义分析的思想对文本进行降维处理。两种改进算法分别代表了两种传统 kNN 算法的加速策略。

对比实验中,首先对参数进行了优化设定。样本裁剪 kNN 算法中参数 MinPts 设置为 15,LowPts 设置为 10。PPkNN 算法中,k' 设置为 30。

由图 4 可以看出,传统 kNN 算法的 Macro-$F1$ 值最高,其次为文本算法 PPkNN。其它两种改进算法的精度损失要比 PPkNN 算法大。实验数据显示,k 值设置为 30 时,PPkNN 算法的 $F1$ 值降低约 0.7%,其他两种改进算法分别降低约 1.1% 和 1.6%。

图 5 显示了四种算法随 k 值增加,分类执行时间的变化情况。由图可知,PPkNN 算法的加速效果明显高于其他改进算法。当 k 取值为 30 时,PPkNN 算法与传统 kNN 算法的加速比达到 21.8,由此可知,本文提出的优化算法加速效果要明显优于以往的加速算法。

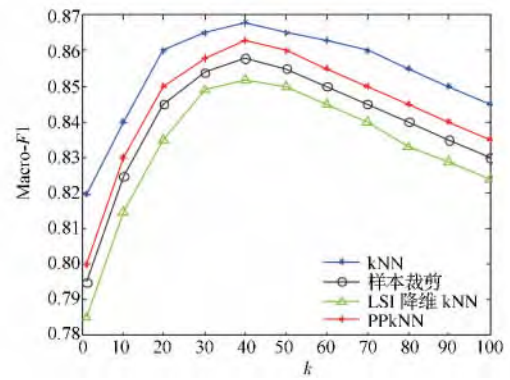

图 4 不同算法 Macro-*F*1 值对比

Fig. 4 Comparison of Macro-*F*1

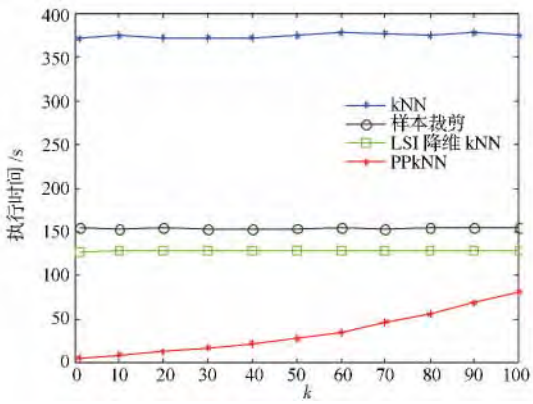

图 5 算法执行时间对比

Fig. 5 Comparison of execution time

4 结束语

针对传统 kNN 算法分类效率低下的缺陷，提出了一种应用投影的方法缩减训练集规模及特征降维处理的优化策略。实验数据显示，通过选取适当的参数，本文提出的优化策略可以有效提高 kNN 算法的分类效率，并且当数据集规模越大、特征空间维度越高时，加速效果越优。

PPkNN 算法虽然加速效果显著，但仍存在一些有待完善之处。首先，PPkNN 算法通过结合训练集缩减和降维来提高算法效率，牺牲了一部分分类的精度。其次，PPkNN 算法中存在多参数设置问题。如何更好的设置各参数，使得分类精度与加速效果达到一个最优平衡点是下一步研究的重点问题。

参 考 文 献

1 Guo G，Wang H，Bell D，*et al*. kNN model-based approach in classification. On The Move to Meaningful Internet Systems 2003: CoopIS，DOA，and ODBASE. Springer Berlin Heidelberg，2003: 986—996

2 Peter H. On projection pursuit regression. Ann Statist，1989; 17(2): 573—588

3 Gupta K M，Moore P G，Aha D W，*et al* . Rough set feature selection methods for case-based categorization of text documents. Proceedings of the 1st International Conference on Pattern Recognition and Machine Intelligenc，2005; 3776 LNCS: 792—798

4 宋枫溪，高秀梅，刘树海，等. 统计模式识别中的维数削减与低损降维. 计算机学报，2005; 28(11): 1915—1922
Song F，Gao X，Liu S H，*et al*. Dimensionality reduction in statistical pattern recognition and low loss dimensionality reduction. Chinese Journal of Computers，2005; 28(11): 1915—1922

5 钟 将，刘荣辉. 一种改进的 KNN 文本分类. 计算机工程与应用，2012; 48(2): 142—144
Zhong J，LIU R. Improved KNN text categorization. Computer Engineering and Applications，2012; 48(2): 142—144

6 廖海波，万中英，王明文. 基于投影寻踪回归文本自动分类的模型. 清华大学学报(自然科学版)，2005; 45(9): 1823—1827
Liao H，Wan Z，Wang M. Automated text classification model based on projection pursuit regression. Tsinghua Univ(Sci & Tech)，2005; 45(9): 1823—1827

7 李荣陆，胡运发. 基于密度的 kNN 文本分类器训练样本裁剪方法. 计算机研究与发展，2004; 41(4): 539—545
Li R，Hu Y. A density-based method for reducing the amount of training data in kNN text classification. Journal of Computer Research and Development，2004; 41(04): 539—545

8 Yong Z. An improved kNN text classification algorithm based on clustering. Journal of Computers，2009; 4(3): 230—237

9 Jiang J. Y，Tsai S. C，Lee S. J. FSKNN: multi-label text categorization based on fuzzy similarity and *k* nearest neighbors. Expert Syst，2012; 39(3): 2813—2821

10 Guo G，Wang H，Bell D，*et al*. Using kNN model for automatic text categorization. Soft Computing-A Fusion of Foundations，Methodologies and Application，2006; 10(5): 423—430

11 邓 箴，包 宏. 用模拟退火改进的 KNN 分类算法. 计算机与应用化学，2010; 27(3): 303—307
Deng Z，Bao H. Based on the simulated annealing algorithm for improved KNN categorization algorithm. Computers and Applied Chemistry，2010; 27(3): 303—307

12 万中英，王明文，廖海波. 一种新的投影寻踪计算方法及在文本分类中的应用. 第三届全国信息检索与内容安全学术会议论文集，2007
Wang Z，Wang M，Liao H. A new PP algorithm and its application to text classification. The Third National Conference on Information Retrieval and Information Content Security，2007

(下转第 102 页)

Conversion Method of Axial Pulling Force in Anchoring Segment of Rocks and Oblique Rally of Anchor End

XIE Dai-xing, SU Chun-tian
(Institute of Karst Geology, Chinese Academy of Geological Sciences, Karst Dynamics Laboratory, Ministry of Land and Resources & Guangxi Autonomous Region, Guilin 541004, P. R. China)

[Abstract] In the anchorage of rock, anchor pullout of parent rock and oblique rally of anchor end were in consistent and had a angle, there were not conversion method at present. To solve the conversion relationship between the anchor pullout of parent rock and oblique rally of anchor end, the principle of force balance and mathematical derivation, such trigonometric function, and method of test are accorded. To infer the conversion formula of anchor pullout of parent rock and oblique rally of anchor end in different angles. the formula could be converted anchor pullout of parent rock and oblique rally of anchor end in different angles.

[Key words] axial pulling force oblique pulling fixed anchor hold pullout test conversion method

(上接第 96 页)

Accelerated *k*-nearest Neighbors Text Classification Algorithm Based on Projection Pursuit

ZHANG Yong, MENG Xiao-fei
(School of Computer & Communication, Lanzhou University of Technology, Lanzhou 730050, P. R. China)

[Abstract] In the traditional k-nearest neighbor (kNN) text classification, the text is represented as the vector space model. As the feature vector dimension and the number of training texts is very large, the *k*-nearest neighbors algorithm is considered as an effective, but not efficient, classification algorithm for text categorization. Aiming at the problem of low classification efficiency, an accelerated strategy is proposed for the traditional kNN based on projection pursuit. The basic idea is to reduce the size of the training set by the projection method, while in the process of looking for *k*-nearest neighbor reduce the dimension of the text. Two-pronged approach to reduce the computational overhead of the algorithm. Experimental results show that the proposed strategy greatly improves the time performance of the traditional kNN, with little degradation in accuracy

[Key words] kNN text classification projection pursuit dimensionality reduction training set reduction