文章编号: 1000-5862(2013)03-0284-04

投影寻踪模型中投影指标的改进

万中英,王明文,揭安全,万剑怡
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要: 针对文本分类问题及投影寻踪降维的特点,对投影寻踪模型中投影指标进行改进,给出了新的投影指标. 对不同的投影指标进行相应的对比实验,实验结果表明: 改进的指标不仅充分利用投影寻踪降到超低维的特点,而且对文本分类的性能有了较大地提高.

关键词: 文本分类; 投影寻踪; 投影指标
中图分类号: TP 391　　**文献标志码**: A

0 引言

文本分类[1-3]是根据预先定义的主题类别,按照一定的规则将文档集合中未知类别的文本自动确定一个类别. 然而文本集中的单词、短语多达数万至数十万个,如果直接用来构成文本特征向量,必将带来以下问题: (Ⅰ)会产生所谓的"维数灾难",即高维空间中的稀疏样本问题; (Ⅱ)极易导致过度拟合现象,导致分类器的泛化能力有限; (Ⅲ)计算复杂度太高,不能满足实际的性能需求. 因此,必须先进行降低维数. 针对此问题已经有相关研究[4-7],采用一种投影指标进行降维,再利用 KNN 或贝叶斯方法进行分类,本文充分利用投影寻踪将数据降到 1 维的特点,提出直接以 1 维文本分类的性能指标 F_1 值作为新的投影指标进行降维. 对不同的投影指标进行相应的对比实验,实验结果表明: 改进的投影指标不仅可以加快分类速度,而且可以提高分类性能.

1 投影寻踪和投影指标

1.1 投影寻踪

投影寻踪[8-11](Projection Pursuit,PP)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,其基本思想是: 把高维数据通过某种组合,投影到低维(1 ~ 3 维)子空间上,并通过极小化(或极大化)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.

投影指标是根据分类的目标构造和优化用于寻找最优投影方向的目标函数. 它用于衡量投影到低维空间上的数据是否有意义,即要找到 1 个或几个投影方向,使它的指标值达到最大或最小值. 因此,在投影寻踪模型中投影指标的好坏直接影响投影方向的选取.

设有 n 个文本 $X \in \mathbf{R}^m$,m 为特征词的个数,$\boldsymbol{a} \in \mathbf{R}^m$ 为投影方向,则文本 X 的投影值为

$$z_i = \sum_{j=1}^{m} a_j x_{ij} = \boldsymbol{a}^{\mathrm{T}} X, i = 1, \cdots, n, j = 1, \cdots, m.$$

在已有文献中构造的投影指标为

$$Q_1(z) = B(z)/W(z),$$

其中 z 为将 m 维数据投影到 1 维的投影值,$B(z)$ 为两类中心离差,$D(z)$ 为类内散布的平均值,且 $B(z)$,$D(z)$ 定义为

$$B(z) = |E(z^{(1)}) - E(z^{(2)})|, \quad (1)$$

$$D(z) = \left[\left(\sum_{j=1}^{n_1} (z_j^{(1)} - E(z^{(1)}))^2 + \sum_{j=1}^{n_2} (z_j^{(2)} - E(z^{(2)}))^2 \right) / (n_1 + n_2) \right]^{0.5};$$

其中 $E(z^{(1)})$、$E(z^{(2)})$ 分别表示第 1 类和第 2 类投影

收稿日期: 2012-11-15
基金项目: 国家自然科学基金(60963014,61163006),江西省教育厅青年科学基金(GJJ11067)和江西省自然科学基金(20114BAB201037)资助项目.
作者简介: 万中英(1977-),女,江西南昌人,讲师,硕士,主要从事文本挖掘方面的研究.

值的均值,$z_j^{(1)}$、$z_j^{(2)}$ 分别表示第 1 类和第 2 类对应文本的投影值,n_1、n_2 分别表示第 1 类和第 2 类的文本数,且 $n_1 + n_2 = n$.

1.2　新的投影指标

在传统的投影寻踪的模型中都是采用 $Q_1(z)$ 作为投影指标,但是许多学者一直希望能找到新的投影指标,使得降维速度加快,分类效果提高. 而所建立的模型是应用于文本分类,因此考虑直接以文本分类为目标. 文本分类的主要性能评价指标是 F_1 值,从而直接以性能指标 F_1 作为投影指标来寻找最优的投影方向,这可能会达到更好的效果. 将文本降到 1 维空间,再以两类中心的中点进行分类. 这样不仅可以加快计算速度,还能提高分类性能. 为此,提出了新的投影指标

$$Q_2(z) = F_1 + B(z),$$

其中 $B(z)$ 的计算方法同(1)式. 此投影指标希望将文本投影到 1 维空间后 F_1 值越大越好,同时 2 个类的中心距离越远越好. 也可以用 F_1 值来修正原投影指标,因此得到投影指标

$$Q_3(z) = Q_1(z) + F_1.$$

2　实验

在 Reuters-21578 标准文本集上针对不同的投影指标进行了实验,该文本集中共有 135 个类别,且类别的分布是非常不均匀的. 最常见的 1 个类别(earn)有 2 877 篇正例训练文档,但有 75 个类别(大于 50%)的正例训练文档数不足 10 篇,且还有一些类别根本就没有正例训练文档. 选用前 10 个最大的类别进行了实验,如表 1 所示. 为了加快算法速度,首先采用特征选择的方法降维,本文采用的 χ^2 算法选取 500 个特征词,词的权重采用 LTC 算法. 采用 F_1 作为性能指标,整体性能指标为微平均 F_1 和宏平均 F_1.

表 1　Reuters-21578 中前 10 个类的名称及文档的个数

类别	earn	acq	money	grain	crude	trade	interest	wheat	Ship	corn
训练集正例	2 877	1 650	538	433	389	369	347	212	197	182
测试集正例	1 087	719	179	149	189	118	131	71	89	56

采用遗传算法的投影方向寻优算法(文献[6]给出了具体的算法描述),将文本降到 1 维后分别进行贝叶斯方法、KNN 方法和利用两类的中心的中点进行分类的方法.

2.1　实验 1

基于分类的目的,直接以 F_1 值作为投影指标,实验结果如表 2 所示,其中 PPNB 表示以 $Q_1(z)$ 为投影指标降维到 1 维空间后进行朴素贝叶斯分类的 F_1 值; PPKNN 表示以 $Q_1(z)$ 为投影指标降维到 1 维空间后进行 KNN 分类的 F_1 值; PPF1KNN 表示以 F_1 值为投影指标降维到 1 维空间后进行 KNN 分类的 F_1 值; PPF1NB 表示以 F_1 值为投影指标降维到 1 维空间后进行朴素贝叶斯分类的 F_1 值; PPF1ONE 表示以 F_1 值为投影指标降维到 1 维空间后以两类中心的中点进行分类的 F_1 值; Mf1 表示微平均 F_1 值; Maf1 表示宏平均 F_1 值.

表 2　以 F_1 值为投影指标的实验结果

类别	earn	acq	money	grain	crude	trade	interest	wheat	Ship	corn	Mf1	Maf1
PPNB	0.957 8	0.916 0	0.759 9	0.885 2	0.831 7	0.788 9	0.809 8	0.867 1	0.785 3	0.844 8	0.896 2	0.846 9
PPKNN	0.964 3	0.920 7	0.764 4	0.890 3	0.866 0	0.793 2	0.809 8	0.845 6	0.855 5	0.857 1	0.906 3	0.857 9
PPF1KNN	0.962 9	0.914 3	0.697 8	0.753 5	0.771 3	0.661 1	0.698 7	0.512 8	0.733 3	0.381 0	0.858 8	0.716 0
PPF1NB	0.945 4	0.911 2	0.627 5	0.750 0	0.773 1	0.666 7	0.716 7	0.476 2	0.732 7	0.333 3	0.848 8	0.707 8
PPF1ONE	0.962 9	0.914 9	0.626 9	0.678 3	0.656 5	0.511 5	0.673 1	0.476 6	0.589 9	0.338 0	0.782 4	0.665 5

从表 2 可以看出实验结果并不理想,主要原因是直接采用 F_1 值容易陷入局部最优.

2.2　实验 2

基于实验 1 的结果希望不仅要 F_1 值越大越好,也要类间距离越远越好,因此进行了实验 2,以 $Q_2(z)$ 为投影指标,实验结果如表 3 所示,其中 PPQ2KNN 表示以 $Q_2(z)$ 为投影指标降维到 1 维空间后进行 KNN 分类的 F_1 值; PPQ2NB 表示以 $Q_2(z)$ 为投影指标降维到 1 维空间后进行朴素贝叶斯分类的 F_1 值; PPQ2ONE 表示以 $Q_2(z)$ 为投影指标降维到 1 维空间后以 2 类中心的中点进行分类的 F_1 值.

表 3 以 $Q_2(z)$ 为投影指标的实验结果

类别	earn	acq	money	grain	crude	trade	interest	wheat	Ship	corn	Mf1	Maf1
PPNB	0.957 8	0.916 0	0.759 9	0.885 2	0.831 7	0.788 9	0.809 8	0.867 1	0.785 3	0.844 8	0.896 2	0.846 9
PPKNN	0.964 3	0.920 7	0.764 4	0.890 3	0.866 0	0.793 2	0.809 8	0.845 6	0.855 5	0.857 1	0.906 3	0.857 9
PPQ2KNN	0.975 8	0.940 4	0.757 2	0.919 9	0.848 0	0.740 4	0.755 4	0.855 2	0.809 2	0.890 8	0.910 6	0.851 7
PPQ2BAY	0.979 3	0.946 0	0.730 7	0.920 0	0.848 2	0.761 5	0.656 9	0.846 2	0.814 4	0.873 8	0.910 7	0.846 1
PPQ2ONE	0.969 8	0.939 8	0.761 2	0.927 3	0.846 0	0.680 0	0.773 5	0.843 5	0.836 0	0.900 0	0.901 9	0.852 0

从表 3 中看到以 $Q_2(z)$ 为投影指标的微平均 F_1 值比以 $Q_1(z)$ 为投影指标的有所提高,两者的宏平均 F_1 值相差较小. 而且发现以 $Q_2(z)$ 为投影指标降维到 1 维空间后以两类中心的中点进行分类也能得到较好的结果.

2.3 实验 3

用 F_1 值来修正原投影指标,实验结果如表 4 所示,其中 PPQ3KNN 表示以 $Q_3(z)$ 为投影指标降维到 1 维空间后进行 KNN 分类的 F_1 值; PPQ3NB 表示以 $Q_3(z)$ 为投影指标降维到 1 维空间后进行朴素贝叶斯分类的 F_1 值; PPQ3ONE 表示以 $Q_3(z)$ 为投影指标降维到 1 维空间后以两类中心的中点进行分类的 F_1 值.

表 4 以 $Q_3(z)$ 为投影指标的实验结果

类别	earn	acq	money	grain	crude	trade	interest	wheat	Ship	corn	Mf1	Maf1
PPNB	0.957 8	0.916 0	0.759 9	0.885 2	0.831 7	0.788 9	0.809 8	0.867 1	0.785 3	0.844 8	0.896 2	0.846 9
PPKNN	0.964 3	0.920 7	0.764 4	0.890 3	0.866 0	0.793 2	0.809 8	0.845 6	0.855 5	0.857 1	0.906 3	0.857 9
PPQ3KNN	0.980 7	0.959 3	0.783 6	0.783 3	0.875 6	0.798 3	0.747 8	0.826 1	0.845 2	0.907 6	0.926 5	0.869 9
PPQ3BAY	0.978 5	0.958 8	0.808 2	0.947 0	0.877 6	0.798 2	0.742 4	0.832 1	0.867 1	0.870 4	0.926 5	0.870 5
PPQ3ONE	0.981 4	0.959 4	0.793 7	0.793 7	0.872 8	0.872 8	0.795 5	0.839 2	0.887 6	0.921 7	0.928 4	0.880 6

从表 4 可以看出以 $Q_3(z)$ 为投影指标降维后进行分类的性能有较大的提高,而且从单个类别来看,最优的类别数也是最多的. 用改进的投影指标进行降维后的分类结果与常用的文本分类方法 KNN、朴素贝叶斯及 SVM 的分类结果进行比较,如表 5 所示.

表 5 与 KNN、朴素贝叶斯及 SVM 方法比较的结果

类别	earn	acq	money	grain	crude	trade	interest	wheat	Ship	corn	Mf1	Maf1
KNN	0.983 9	0.952 7	0.834 7	0.885 8	0.866 8	0.772 7	0.789 5	0.748 0	0.804 7	0.750 0	0.920 7	0.842 8
Bayesian	0.954 6	0.935 7	0.540 8	0.576 1	0.423 2	0.301 9	0.522 9	0.398 7	0.500 0	0.332 1	0.690 6	0.586 9
SVM	0.983 1	0.953 0	0.805 4	0.927 3	0.890 6	0.763 9	0.752 1	0.832 1	0.883 7	0.899 1	0.926 7	0.870 3
PPNB	0.957 8	0.916 0	0.759 9	0.885 2	0.831 7	0.788 9	0.809 8	0.867 1	0.785 3	0.844 8	0.896 2	0.846 9
PPKNN	0.964 3	0.920 7	0.764 4	0.890 3	0.866 0	0.793 2	0.809 8	0.845 6	0.855 5	0.857 1	0.906 3	0.857 9
PPQ3KNN	0.980 7	0.959 3	0.783 6	0.783 3	0.875 6	0.798 3	0.747 8	0.826 1	0.845 2	0.907 6	0.926 5	0.869 9
PPQ3BAY	0.978 5	0.958 8	0.808 2	0.947 0	0.877 6	0.798 2	0.742 4	0.832 1	0.867 1	0.870 4	0.926 5	0.870 5
PPQ3ONE	0.981 4	0.959 4	0.793 7	0.793 7	0.872 8	0.872 8	0.795 5	0.839 2	0.887 6	0.921 7	0.928 4	0.880 6

从表 5 中可知 Bayesian 方法得到的分类结果较差,这主要是因为如果某个特征词在某类文档中一次也没出现,那么这个词在该类中的先验概率就为 0,而该类测试文档中却出现了该词,这会使得测试文档在属于该类的概率为 0,从而导致错分的结果. 因此 Bayesian 方法得到的分类结果并不理想.

并且从表 5 还可以看出,采用改进投影指标的方法得到的分类结果是最优的.

2.4 实验 4

为检验改进投影指标的方法在中文文档集中的性能,在相同情况下用复旦文档集进行了实验. 在复旦数据集中训练集和测试集的总文档数分别为 8 213 和 6 163,前 9 大类的训练集正例及测试集正例如表 6 所示.

表 6 复旦文档集中前 9 个类的名称及文档的个数

类别	Economy	Sports	Computer	Politics	Agriculture	Environment	Art	Space	History
训练集正例	1 369	1 204	1 019	1 010	847	805	510	506	466
测试集正例	1 127	980	591	989	635	371	286	248	468

以 $Q_3(z)$ 为投影指标降维后进行分类的结果与常用的文本分类方法 KNN、朴素贝叶斯及 SVM 的分类结果进行比较,如表 7 所示.

表 7　在复旦文档集上的结果比较

类别	Economy	Sports	Computer	Politics	Agriculture	Environment	Art	Space	History	Mf1	Maf1
KNN	0.883 8	0.908 7	0.932 1	0.849 4	0.907 7	0.862 1	0.758 1	0.755 9	0.658 8	0.861 4	0.842 6
Bayesian	0.441 9	0.794 0	0.824 9	0.724 0	0.381 9	0.664 2	0.363 8	0.470 1	0.351 9	0.526 6	0.591 2
SVM	0.904 0	0.947 4	0.954 3	0.864 9	0.929 0	0.940 4	0.787 8	0.861 9	0.736 4	0.894 5	0.881 5
PPNB	0.862 7	0.925 0	0.908 8	0.854 7	0.913 0	0.910 4	0.750 4	0.836 9	0.732 1	0.868 1	0.857 4
PPKNN	0.860 8	0.923 9	0.934 0	0.857 9	0.906 4	0.912 1	0.751 3	0.831 5	0.743 2	0.871 0	0.858 8
PPQ3KNN	0.870 6	0.918 7	0.936 8	0.850 0	0.907 4	0.909 6	0.734 5	0.808 9	0.680 6	0.865 6	0.848 7
PPQ3BAY	0.875 9	0.917 0	0.939 1	0.820 5	0.903 9	0.907 2	0.712 7	0.806 2	0.679 9	0.860 7	0.842 7
PPQ3ONE	0.873 2	0.917 5	0.937 7	0.861 3	0.908 5	0.903 7	0.740 5	0.792 6	0.703 9	0.863 1	0.856 2

从表 7 可以看出在复旦文档集上 SVM 算法的性能是最好的. 但除去 SVM 和 Bayesian 方法外其它的分类效果都大致相同. 由此可见,此投影指标应用于英文文档集比用于中文文档集要好.

3　总结

本文就投影寻踪模型中投影指标进行了改进,该指标以文本分类效果达到最优为目标. 在 Reuters-21578 标准文本集上采用原投影指标和新的投影指标进行了对比实验,实验结果表明新的投影指标取得了更好的效果. 尤其是将文本投影到 1 维空间后仅需根据两类的中心距离的中点就可以进行分类,这样不仅简单方便,而且得到了较好的分类效果. 最后在复旦文档集上进行验证,发现在中文文档集上的效果没有显著地提高.

4　参考文献

[1] 谭松波. 高性能文本分类算法研究 [D]. 北京: 中国科学院计算技术研究所,2006.

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

[3] 尚文倩. 文本分类及其相关技术研究 [D]. 北京: 北京交通大学,2007.

[4] Wan Zhongying, Wang Mingwen, Liao Haibo. Orthogonal projection feature extraction and its application to text classification [J]. Journal of Computational Information Systems,2008,4(3):1289-1297.

[5] 万中英,王明文,廖海波,等. 维数约简在网页分类中的应用 [EB/OL]. [2012-06-12]. http://cpfd.cnki.com.cn/Article/CPFDTOTAL_ZGZR20041001028.htm.

[6] 万中英,王明文,廖海波. 基于投影寻踪的中文网页分类算法 [J]. 中文信息学报,2005,19(4):60-67.

[7] 万中英,廖海波,王明文. 遗传-粒子群的投影寻踪模型 [J]. 计算机工程与应用,2010,46(20):210-212.

[8] 成平,李国英,陈忠琏,等. 投影寻踪讲义 [M]. 北京: 中国科学院系统科学所,1986.

[9] Mandelzweig M, Demko A B, Dolenko B, et al. A projection method for the visualization of high-demensional biomedial datasets [EB/OL]. [2003-05-07]. http://citeseer.ist.psu.edu/669694.html.

[10] 付强,赵小勇. 投影寻踪模型原理及其应用 [M]. 北京: 科学出版社,2006.

[11] Emmanuel A, Luis O J. Unsupervised feature extraction using projection pursuit [EB/OL]. http://www.censsis.neu.edu/Education/StudentResearch/2001/posters/arzuaga-cruz_e!.pdf.

The Projection Index´s Improvement in Projection Pursuit Model

WAN Zhong-ying, WANG Ming-wen, JIE An-quan, WAN Jian-yi

(College of Computer Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: Based on the text classification problem and PP features, the projection index of PP model has been improved. The projection index has been comparative experiment, the experimental results show that the projection index not only has made full use of projection pursuit to ultra low dimensional characteristics, but also has been improved greatly on the performance of text classification.

Key words: text classification; projection pursuit; projection index

(责任编辑: 冉小晓)