

基于 KPCA 和 RBF 网络的文本分类研究

杨 俊, 陈贤富

(中国科学技术大学 电子科学与技术系, 安徽 合肥 230027)

摘 要: 基于词空间的分类方法很难处理文本的高维特性和复杂相关性, 为此文中提出了基于核的主成分分析和径向基神经网络的文本分类算法. 首先利用核主成分分析选择合适的核函数从高维特征空间中提取文本向量的主成分, 实现了文本输入空间的降维和语义特征空间的抽取, 然后在语义特征空间中训练径向基神经网络分类器, 并利用训练得到的分类器进行文本分类工作. 实验结果表明: 核主成分分析不仅实现了降维, 而且能在大幅减少径向基神经网络训练时间的基础上显著提高其分类精度.

关键词: 文本分类; 特征选择; 核主成分分析; 径向基神经网络

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2010)03-0122-04

Text Categorization Based on KPCA and RBF Neural Network

YANG Jun, CHEN Xian-fu

(Department of Electronic Sci. and Tech., University of Science and Technology of China, Hefei 230027, China)

Abstract: It is difficult for methods based on word spaces to handle with the high dimensionality characteristic and complex correlation of the texts vectors. To solve this problem, a algorithm based on kernel principal component analysis and RBF neural network is proposed. First, this new algorithm employs KPCA with a appropriate kernel function to find the principal components of the input vectors in the high dimensional feature space, which effectively reduces the dimensionality of input vectors and gets the semantic feature space. Then, we train a RBF neural network in the semantic feature space. The experiment results show that the new method can effectively reduce the dimensionality of the data sets and notably enhance the classification precision while reduces the training time of the RBF networks.

Key words: text categorization; feature selection; KPCA; RBF neural networks

1 引言

文本分类技术是信息检索和文本挖掘的重要基础, 其主要任务是在预先给定的类别标记集合下, 根据文本内容判定它的类别^[1]. 其可分为 3 个步骤: 文本的向量表示、特征降维、分类器模型的构造.

目前广泛使用向量空间模型来表示文本, 相关研究主要集中在以什么语义单元作为特征项及计算项的权重两个问题上. 大部分研究仍以词或 n -gram 作为项, 以项的频率为基础计算权重, 如著名的 TF-IDF 公式^[1].

由于自然语言中存在大量一词多义和多词同义现象, 导致由向量空间模型(VSM)得到的文本特征

向量具有高维度、复杂相关性和非线性等特性^[1]. 传统的特征选择算法如文档频率(DF), 信息增益(IG)等基于文本特征间相互独立的前提假设, 对每个特征进行独立评估并计算权值, 按权值大小排序, 然后根据预定的阈值或特征数目选取最佳特征子集^[2]. Cover 指出, 即使假定各个原始特征之间相互独立, 利用一定的最优原则得到的各个最优特征构成的特征子集也未必就是最优特征子集^[3]. 因此有必要采取一种综合评估方法对原始特征进行变换, 将输入空间中大量相关的词转化为少数不相关的特征集合.

由 Cover 定理可知: 输入空间中的非线性可分模式, 可以通过非线性函数 ϕ 映射到特征空间, 使得

模式在特征空间中以较高的概率线性可分^[3].而Scholkopf等引入了核方法^[4],使得特征空间中两个向量的内积可在输入空间中利用一个核函数计算得到,不必显式计算从输入空间到特征空间的映射,从而可以利用已有的线性求解算法对那些高维的且复杂相关的非线性输入模式实施降维,而文本特征非常适于用核方法来进行处理.

通过与核方法有机融合而形成的基于核的主成分分析算法(KPCA)特别适合于处理高维非线性问题,且相对于特征选择能提供更多的信息^[4].

2 基于KPCA和RBF网络的文本分类

2.1 基于KPCA的特征降维

其基本原理是:设变换 ϕ 实现输入空间 R^n 到特征空间 F 的映射即 $x_k \rightarrow \phi(x_k), \phi(x_k) \in F, k = 1, 2, \dots, M$.则 F 空间中的协方差矩阵为:

$$\bar{C} = \frac{1}{M} \sum_{i=1}^M \phi(x_i) \phi(x_i)^T \quad (1)$$

求 $\bar{C}$ 满足 $\lambda \geq 0$ 的特征值和特征向量 $v \in F$,
 $\lambda v = \bar{C}v$ (2)

对 $\lambda \neq 0$ 对应的特征向量 v ,存在系数 $\alpha_i (i = 1, 2, \dots, M)$,使得

$$v = \sum_{i=1}^M \alpha_i \phi(x_i) \quad (3)$$

同时由式(2)得

$$\phi(x_k) \cdot \lambda v = \phi(x_k) \cdot \bar{C}v \quad (4)$$

联立式(3)(4)两式,并定义一个 $M \times M$ 的矩阵 K ,其中 $k_{ij} := (\phi(x_i) \cdot \phi(x_j))$,可得

$$M\lambda\alpha = K\alpha \quad (5)$$

得 q 个非零的特征值 $\lambda_1 > \lambda_2 > \dots > \lambda_q$ 和对应的特征向量 $\alpha^1, \alpha^2, \dots, \alpha^q$.

再由式(3)求得 $\bar{C}$ 的 q 个特征向量 $v^1, v^2, \dots, v^q$,比值 $\lambda_i / \sum_{i=1}^q \lambda_i$ 的大小反映了 v^i 分量在 $\bar{C}$ 中的重要程度,通过预先设定比值 $\theta (0 < \theta \leq 1)$,使得

$$\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^q \lambda_i} \geq \theta, \text{便可从中挑选出 } m \text{ 个主成分 } (m < q)$$

并实现输入空间的降维.测试样本在 F 空间向量 v^k 上的投影为:

$$p_k(x) = (v^k, \phi(x)) = \frac{1}{\sqrt{M\lambda_k}} \sum_{i=1}^M \alpha_i^k k(x_i, x) \quad (6)$$

由式(6)可见,核函数的优势在于并不要求知道具体的 ϕ ,而只要求已知核函数的具体形式.

2.2 RBF神经网络

RBF网络是一种性能优良的前向神经网络:①网络所有的信息都等势分布贮存于网络内的各神经元,具有很强的鲁棒性和容错性;②网络可以充分逼近任意复杂的非线性关系,适于分类曲面复杂的文本分类问题;③网络采用并行分布的处理方法,使得快速进行大量运算成为可能.但RBF对输入特征的要求较高,当特征选择不当或特征维数过大时,网络的学习能力和泛化能力都可能出现.因此,利用RBF进行文本分类,关键是提取出既能全面反映文档类别信息,又有利于神经网络学习的特征.图1是RBF神经网络的结构.

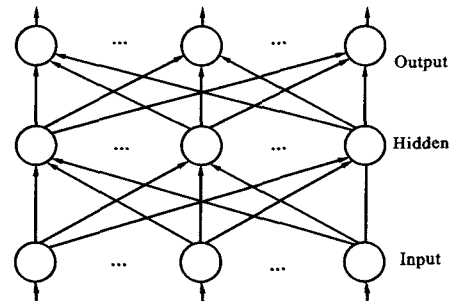


图1 RBF神经网络结构

2.3 算法流程

步骤1:把训练文本集表示成文本向量集 $x_1, x_2, \dots, x_M$,得到输入空间.

步骤2:选定核函数,并由式(1)~(6)求得 $\bar{C}$ 满足 $\lambda \geq 0$ 的特征值和特征向量 v ,通过设定 θ 的值以控制降维幅度,实现从输入空间到特征空间的转化.

步骤3:利用降维后的模式对RBF网络进行训练.

步骤4:对测试集进行分类测试.

3 实验设计和结果分析

3.1 实验评价标准

文本分类系统的评价标准包含两个指标:准确率(r)和查全率(p).两者反映了分类质量的两个不同方面,必须综合考虑,因此存在一种新的评估标准: F_1 测试值,其计算公式为:

$$F_1 = \frac{2rp}{r+p} \quad (7)$$

另有微平均和宏平均两种计算 F_1 值的方法.微平均统一计算全部类的 F_1 值;宏平均计算每一类的 F_1 后取算术平均值.

根据文献[2],微平均 F_1 被广泛应用于文本分类器评价,其计算公式如(8),文中使用它来评价分类效果.

$$\text{micro-}F_1 = \frac{2 \sum_{i=1}^m p_i \sum_{i=1}^m r_i}{m(\sum_{i=1}^m p_i + \sum_{i=1}^m r_i)} \tag{8}$$

3.2 实验设计

实验流程如图 2 所示.

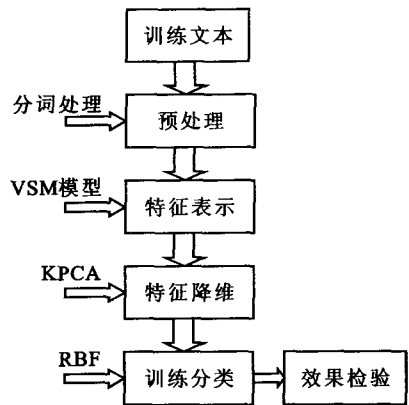


图 2 实验流程图

训练集采用“ApteMod”版本的 Reuters-21578 语料集,选择其中样本最多的 8 个类别,将其划分为训练集和测试集两部分以供实验,具体情况见表 1.

表 1 8 个类别的训练集和测试集文档数目

	训练集	测试集
Earn	2709	1044
Acq	1488	719
Money-fx	465	168
Grain	389	189
Crude	394	130
Trade	337	106
Interest	289	131
Wheat	208	65

文中利用开源文本分类系统 Rainbow 进行实验,在其中嵌入 KPCA 和 RBF 模块. Rainbow 中包含 525 个停用词,分词后去除停用词并合并同根词,得到的词条总数为 13673.

特征表示采用 VSM,以词作为特征项,词权重的计算采取 TF-IDF 公式.

根据 Mercer 定理,函数 k 可以作为核函数的充分必要条件是:从给定输入空间中的有限数据集 $x_1, x_2, \dots, x_k$,求得的 k 是半正定的. 根据这一定理,3 类核函数(多项式核函数、高斯核函数、

Sigmoid 核函数)被广泛应用于 KPCA. 文中使用多项式核函数: $k(x, x_i) = [(x \cdot x_i) + 1]^q$.

考虑到 RBF 训练时间较长,对全部 8 个类别只建立一个分类器. 基函数采用高斯基函数:

$$g_j(x) = \exp(-\frac{\|x - c_j\|^2}{\sigma_j^2}).$$

3.3 实验结果分析

表 2 是通过改变多项式核函数中参数 q 的值和特征维数所得的实验结果. 由文献[2]知,DF 是目前最优秀的特征选择算法之一,而 KNN 是目前分类精度最高的单分类器之一. 最后两行是分别采用基于 DF 降维的 RBF 分类器和 KNN 分类器所得的分类结果,用于和文中算法进行比较[5].

表 2 改变 q 和降维后文本特征维数的微平均 F_1

	100	200	500	1000	2000	4000
$q = 1$	0.691	0.732	0.780	0.795	0.794	0.801
$q = 2$	0.745	0.796	0.829	0.825	0.819	0.818
$q = 3$	0.791	0.843	0.852	0.854	0.852	0.853
$q = 4$	0.792	0.844	0.860	0.857	0.850	0.851
$q = 5$	0.803	0.839	0.856	0.859	0.857	0.855
DF + RBF	0.735	0.818	0.830	0.851	0.854	0.851
DF + KNN	0.774	0.831	0.844	0.855	0.861	0.860

当 q 取 1 时,核函数退化为线性函数,KPCA 退化为 PCA. 由表 2 可见,其微平均 F_1 显然低于 DF 算法和 $q > 1$ 时的 KPCA 算法,这和 Li 等的实验结果相符合[6]. 在他的实验中,PCA 没有取得令人满意的效果. 这是因为 PCA 是一种线性变换方法,没能充分考虑到原始文本特征的非线性所致.

由表 2 和图 3 可见:当 q 取 3 时,分类器取得较好的性能. 当 q 再提高时,带来的性能提升有限,计算时间复杂度却显著提高. 此时通过比较 KPCA + RBF 和 DF + RBF 可以发现:当维数低于等于 500 时,KPCA 的表现要好于 DF,这是因为 KPCA 通过非线性映射,去除原始输入特征的复杂相关性,从而仅利用较少的维数就可以表示原始输入空间的大部分信息,反观 DF 在维数低于 500 时却丢失了大量信息,导致分类精确度下降幅度较大;而当维数高于 500 时,两者效果趋于相当.

由图 3 可见:当 q 取 3 时,KPCA + RBF 取得了和 DF + KNN 相当的分类精度. 这说明 KPCA 特征抽取是成功的,其能提取出全面反映文档类别信息,且有利于 RBF 神经网络学习的特征,使其获得了能够和 KNN 分类器相媲美的分类精度. 与此同时,前者的单篇文档平均分类时间只有后者的 1% 左右,在分类效率上 RBF 神经网络发挥了其并行分布运

算的优点.

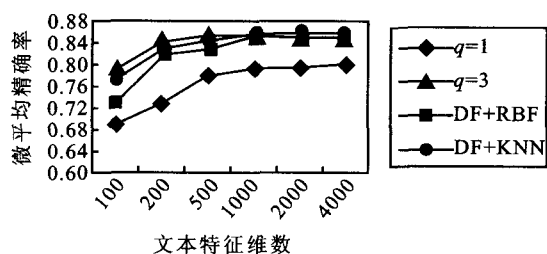


图3 微平均 F_1 坐标图

由表2可见,当特征维数高于500后,RBF分类精度并没有获得显著提升,但伴随着特征维数升高,其训练时间呈几何级数增长.可见,如果利用KPCA把特征维数压缩到500维以下,不但可以获得不逊于KNN的分类精度和好于DF的降维效果,而且可以显著降低RBF的训练时间,这正是结合KPCA和RBF神经网络的优势所在^[7].

4 结束语

特征降维是目前文本分类所面临的主要问题之一.文中提出了一种基于KPCA和RBF的文本分类新方法.实验结果表明,KPCA不仅有效实现了文本输入空间的特征降维,而且在维数低于1000的情况下取得了比DF特征选择算法更好的分类性能,同时取得了和KNN分类器相当的分类精度;如果能充分利用KPCA的这种特性,就可以在不损失分类精度的前提下显著减少RBF分类器的训练时间,从而在发挥RBF神经网络一系列优点的前提下克服其在训练集特征维数过高情况下训练时间过长的缺

点.

文中只在reuter-21578上进行了实验.在RBF神经网络训练过程中还存在过训练的缺点,这是需要改进的地方.另外,将通过对更多的语料测试集的研究来进一步明确KPCA+RBF算法在文本分类领域的能力和应用潜力.

参考文献:

- [1] 苏金树.基于机器学习的文本分类技术研究进展[J].软件学报,2006(9):1848-1859.
- [2] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]// Proceedings of the 14th ICML. USA: Nashville 1997:412-420.
- [3] Cover T M. The best two independent measurements are not the two best[J]. IEEE Transactions on Systems, Man and Cybernetics, 1974(4):116-117.
- [4] Scholkopf B, Smola A, Muller K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998,10(5):1299-1319.
- [5] 崔志明,谢春丽.基于WEB的文本挖掘研究[J].微电子学与计算机,2002,19(10):51-53.
- [6] Li YH, Jain A K. Classification of the text document[J]. The Computer Journal,1998,41(8):537-546.
- [7] 王俊英,郭景峰,霍铮.中文文本分类系统的设计与实现[J].微电子学与计算机,2006,23(9):262-265.

作者简介:

杨俊男,(1983-),硕士研究生.研究方向为自然语言处理、机器学习.

陈贤富男,(1963-),博士,副教授.研究方向为机器学习、智能信息处理.

(上接第121页)

- [5] 赖鑫生,张明义.基于渗透原理迁移策略的并行遗传算法[J].计算机学报,2005,7(7):1146-1152.
- [6] 刘文强,张晓伟.多峰值全局优化的混合算法[J].计算机工程与应用,2008,44(33):43-45.
- [7] 范虹,孟庆丰,张优云.用混合编码遗传算法实现匹配追踪算法[J].西安交通大学学报,2005,39(3):295-299.
- [8] 云庆夏.进化算法[M].北京:冶金出版社,2000.

作者简介:

李喜艳女,(1983-),硕士研究生.研究方向为自动机理论和信息安全.

周清雷男,(1962-),教授,博士生导师.研究方向为自动机理论和信息安全.

张文宁女,(1982-),硕士研究生.研究方向为软件工程.