

文章编号: 1003-0077(2018)11-0049-06

基于小波分析的特征提取文本分类方法研究

朱 晋¹, 怀丽波¹, 崔荣一¹, 尹 慧²

(1. 延边大学 计算机科学与技术学院 智能信息处理研究室, 吉林 延吉 133002)

(2. 延边大学 计算机科学与技术学院, 吉林 延吉 133002)

摘 要: 该文提出了基于小波分析的文本特征提取方法, 对传统 TF-IDF 向量空间模型下的特征向量进行了该文的的小波变换、逆小波变换。使用 KNN 分类方法检验这两空间下的文本分类准确率。实验结果表明, 该文的小波变换方法在减少了 TF-IDF 向量空间模型近一半的维度下在各种实验条件中都能和向量空间模型保持一致的分类准确率; 该文的逆小波变换方法在大幅度降低 TF-IDF 向量空间模型维度的基础上, 同实验中其他特征提取方法相比, 在特定条件下有着卓越的特定文本类别分类优势, 这也在一定程度上检验了压缩感知理论的正确合理性。

关键词: 压缩; 小波分析; TF-IDF; KNN; 分类正确率; 压缩感知

中图分类号: TP391

文献标识码: A

Feature Extraction for Text Classification Based on Wavelet Analysis

ZHU Jin¹, HUAI Libo¹, CUI Rongyi¹, YIN Hui²

(1. Intelligent Information Processing Lab., Department of Computer Science and Technology,
Yanbian University, Yanji, Jilin 133002, China)

(2. Department of Computer Science and Technology, Yanbian University, Yanji, Jilin 133002, China)

Abstract: This paper presents a method for text feature extraction based on wavelet analysis with the TF-IDF vector space as the input. KNN method is employed to examine text classification accuracy in two spaces. The experiment results show that the wavelet transformation method reduces almost half of vector space dimensions while maintaining the same classification accuracy of classical vector space model and the proposed inverse wavelet transformation exhibits excellent advantages of large dimension reduction for specific text categories, which testify the correctness and rationality of the compressive sensing.

Keywords: compressing; wavelet analysis; TF-IDF; KNN; classification accuracy; compressive sensing

0 引言

文本分类是分析待定文本的特征, 并与已知类别中文本所具有的共同特征进行比较, 然后将待定文本划归为特征最接近的一类并赋予相应的分类号^[1-2]。通常用一组词条作为属性向量构成特征向量空间。文本的原始特征向量空间包含全部的词条属性, 具有高维、稀疏的特点, 但并不是所有属性对

分类决策都有贡献, 冗余的属性不但对决策无任何贡献, 反而会降低决策的执行效率。因此需要在不降低系统性能的前提下, 对高维文本特征空间进行有效地降维, 提取出最佳分类特征属性集合^[3]。

数据压缩一直是小波分析的重要应用领域之一, 并由此带来了巨大的社会效益和经济效益^[4]。本文对向量空间模型下的特征向量进行了本文的小波变换、逆小波变换, 使文本特征空间维度有所减小, 期望达到提取文本特征、进行有效文本分类的目的。

收稿日期: 2018-03-16 定稿日期: 2018-08-26

基金项目: 国家语委“十二五”科研规划项目(YB125-178); 吉林省科技发展计划项目(20140101186JC); 吉林省教育厅科研项目(JJKH20180896KJ)

1 相关工作

1.1 文本特征处理

文本表示一般采用向量空间模型,该模型是由 G. Salton 提出的^[5]。该模型不考虑词的顺序,将文本简化为一个 BOW(Bag-of-Words),并表示为特征权重的向量。除此之外的文本表示有基于高阶词统计、基于特征概率分布、将文本理解为信号序列、二维视图等模型,但应用都十分局限^[6]。向量空间模型主要以词作为特征,以词频矩阵为基础计算权重。常用的特征提取方法有文档频率、信息增益、互信息、卡方检验、期望交叉熵、TF-IDF 方法和特征降维^[7]等。

现有的特征降维技术很多:停用词表,停用词的区别作用不大,从词典里去掉停用词可以达到降维目的,但现今停用词表依旧不够健全完善且特殊情况下停用词对提取一篇文档的特征还是有作用的;独立成分分析(ICA),用 ICA 将输入文本空间映射到相应的独立成分空间,这种方法产生的计算空间小^[2];主成分分析(PCA),将高维的词语特征一文档空间转换为一个低维度的正交矩阵,从中选择最有辨别能力的特征,最终得到最佳的分类特征子集^[3];奇异值分解(SVD),使用 SVD 对特征文本矩阵进行降维,解决了同义词和多义词问题,降低了文本分类的计算量^[8]。

1.2 小波分析

小波分析能有效地从非平稳信号中提取出有用信息,从根本上克服了傅立叶分析只能以单个变量描述信号的缺点^[9]。小波分析在信号分析、神经网络、模式识别、语音合成、方程求解等方面取得了重要成果。小波变换可以起到压缩数字信号的作用:小波变换后数据可以截断,仅存放小部分最强的小波系数,就能保留近似的压缩数据^[10]。常见小波函数有 Haar 小波, Mexican hat 小波, Morelet 小波, Daubechies 小波等^[11]。

1.3 文本分类

自从 20 世纪 90 年代开始,文本分类主要为基于统计和机器学习的方法,这种方法相对于知识工程方法,在准确率和稳定性方面都有明显的优势。在构建分类器过程中,分类器是自动建立的,分为学习过程和分类过程,学习过程是基于训练集学习到

一个分类器,而分类过程则是利用学习到的分类器预测新数据的类别。整个过程不需要专家参与,分类的时间开销和人力投入都很少,准确率却得到了提高。研究人员尝试了大量的机器学习算法,包括:支持向量机、朴素贝叶斯、KNN、决策树、Rocchio、最大熵模型^[12]等。

1.4 压缩感知

压缩感知理论首先由 Candès、Romberg、Tao 和 Donoho 等人在 2004 年提出,文献直到 2006 年才发表。Candès 证明了只要信号在某一个正交空间具有稀疏性,就能以较低的频率采样信号,而且能以高概率重构该信号^[13]。压缩感知理论可以高效地采集稀疏信号的信息,通过非相关性感知测量值,此特性使得压缩感知广泛地应用于现实生活中。压缩感知理论解决了信息采集和处理技术目前遇到的瓶颈,带来了革命性的突破,受到各国学者的广泛关注,从医学成像和信号编码到天文学和地球物理学均有应用^[14]。

2 基本理论

2.1 TF-IDF 特征提取方法

单词权重计算最为有效的实现方法是 TF-IDF,它是 Salton 在 1988 年提出的。它的计算如式(1)所示。

$$W(t_i, d_j) = tf(t_i, d_j) \times idf(t_i, d) \quad (1)$$

其中, $W(t_i, d_j)$ 是特征项 t_i 在文本 d_j 的权重取值; $tf(t_i, d_j)$ 是特征项 t_i 在文本 d_j 中出现的频率,用于计算该词描述文档内容的能力; $idf(t_i, d)$ 是特征项 t_i 在文本集 d 中出现文本频率数的反比,称为反文档频率,用于计算该词区分文档的能力。TF-IDF 法认为一个单词出现的文本频率越小,它区别不同类别的能力就越大,所以引入了反文本频率 IDF 的概念,以 TF 和 IDF 的乘积作为特征空间坐标系的取值测度^[3]。

2.2 Mallat 算法

Mallat 于 1987 年把多分辨率思想引入小波分析中,提出了塔式分解算法,即 Mallat 算法,该算法在实际应用中减少了小波变换的复杂度。分解式子可表示为式(2):

$$c_{j+1} = \sum_{m \in Z} c_j(m) h(m - 2n)$$

$$d_{j+1} = \sum_{m \in Z} c_j(m)g(m-2n) \quad (2)$$

重构式子表示为式(3)：

$$c_j = \sum_{m \in Z} c_{j+1}(m)h(n-2m) + d_{j+1}(m)g(n-2m) \quad (3)$$

信号的小波分解和重建可通过子带滤波的形式来实现^[15]。

2.3 KNN 法

KNN 分类算法能够确定待分类样本与训练样本之间的相似程度,从而确定与待分类样本距离最近的 K 个训练样本。其最关键的因素是相似性度量方法,最常用的相似性度量方法是余弦相似度,如式(4)所示。

$$\text{sim}(X, Y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n (x_i)^2} \sqrt{\sum_{i=1}^n (y_i)^2}} \quad (4)$$

其中, X, Y 代表两个文档表示向量。对于一个待分类文本 x , 根据相似性度量函数从整个训练集中找到与文本 x 最相似的 K (K 是预先设定的一个整数) 个文本, 然后根据 K 个近邻文本所属的类别给 x 的候选类别评分^[16]。

2.4 基于小波分析的文本特征提取方法

2.4.1 相关理论分析

向量空间模型简化了文本处理, 不同词语之间的组合可能会达到它们排列的效果。但其缺点是随着文本集扩充、词典单词增多, 向量维度会迅速增加。单个文本向量很难占有词典里的大部分词, 故有很多维度权重值为 0, 产生了向量的高维度、稀疏性现象。因此我们需要对传统向量空间模型中的向量进行降维处理, 可以把文本向量看成数字信号。而小波分析理论对数字信号处理具有很强的优势。现有的理论和实践表明, 变换后的数字信号能高度还原到原始信号, 且小波分析能独到地捕捉到局部化细节, 这就使在小波变换空间进行本文操作变成了可能。

本文的小波变换是将一维离散小波变换产生的低频向量和高频向量进行了对应分量的简性相加。一维离散小波变换后的低频和高频向量维度相同且约为原始向量的一半, 因此达到了降维目的。这可解释为我们在分析语言内容时会从正向和反向进行理解。低频信息类比我们正向理

解, 这在整体理解上确实占了很大比重; 高频信息类比我们语言内容的反向理解, 相应地所占比重小。正向加反向理解便是我们对语言内容的整体把握。在小波空间比重不同, 但是实际上比例应该一致, 故本文进行的逆小波变换是采用尺度函数对高频信息和低频信息进行变换再进行简性相加。更进一步解释, 信号一般都符合高斯分布, 所以本文的逆小波变换提取了上述变换的中间若干维, 从而达到降维的目的。

2.4.2 小波分析法对特定分类类别的优势

本文对特定训练集进行了相关统计：

- 1) 某类别数量小于 1KB 的文本数 $S_1^i, 1 \leq i \leq n$;
- 2) 某类别总文本数 $S_2^i, 1 \leq i \leq n$;
- 3) 某类别特有特征词数 S_3^i (非精确值, 误差 ± 10), $1 \leq i \leq n$;
- 4) 某类别综合因子 S^i , 自定义表达式如式(5)所示。

$$S^i = (S_1^i/S_2^i) \times (S_2^i/T) \times (S_3^i/D) \quad (5)$$

其中, n 为训练集的文本类别总数, T 为训练集总文本数, D 是词典中单词数。

根据压缩感知理论, 正交、高稀疏空间的信号进行变换会以高概率还原到原始信号。本文中 DWT 空间符合初始条件: 1) 正交的 DBN 小波; 2) 高稀疏: 后期测得小波空间低频部分、高频部分的零值过半的向量均占了各自总向量的 90% 以上, 某类别的稀疏程度可由上文提到的类别综合因子描述。所以在做本文提出的逆变换后有机会还原出原始信号的重要部分。类综合因子作为本文逆小波变换特定分类优势的判断标准。

通过以上分析, 本文对向量空间模型进行了本文采用的 DWT、IDWT 方法, 用文本分类的准确程度检验该文本特征提取方法的有效性。以下是算法步骤:

Step 1 对数据集进行分词, 构建出词典, 获得 TF-IDF 特征空间向量;

Step 2 利用式(2)等对已获得的 TF-IDF 向量进行一维离散小波变换, 得到尺度系数和小波系数;

Step 3 对尺度系数和小波系数进行对应分量相加, 得到本文提出的小波空间各向量;

Step 4 对 Step 3 得到的尺度系数和小波系数均利用式(3)进行对应的尺度函数还原再相加, 对得到的向量提取其中间若干维度, 获得本文提出的逆小波空间各向量;

Step 5 利用测试文本计算在两类空间下进行 KNN 分类的相似性;

Step 6 比较所有测试向量判定标签和其真实所属标签,计算准确率。

3 实验部分

本文分类实验是为了验证特定分类组的实验准确率,故对于每个特定分类组的实验没有掺杂负样本。

3.1 实验过程

采用中国科学院 NLPPIR 分词系统对数据集进行分词。使用 SVD 方法、ICA 方法、PCA 方法、本文 DWT 小波变换方法、本文 IDWT 逆小波变换方法对向量空间 VSM 下的特征向量(TF-IDF)分别进行降维特征提取。再使用 KNN 方法(余弦距离作为相似性度量标准)对各个空间进行分类,测得各空间的分类准确率。由于前期所做文本实验发现选用 DB20 小波、K=8 时优于同水平实验结果,故本文实验参数亦如此设定。SVD 空间、ICA 空间、PCA 空间、IDWT 空间的维度一致。

本文共进行两大组实验。第一组实验测试各训练空间针对内部样本时的分类性能(测试集和训练集来自同一样本集);选用复旦大学新闻组语料库,该语料库包含 20 个类别、9 804 篇文本。舍去英文单词、标点符号、数字、部分停用词、极低频词,获得特征词共计 27 950 个。每类中的文本数如表 1 所示。

表 1 各类别文本数

类别	文本数	类别	文本数
农业	1 021	教育	59
艺术	740	电子	27
交通	25	能源	32
计算机	1 357	环境	1 217
经济	1 600	历史	466
心理	44	法律	51
政治	1 024	文学	33
太空	640	医药	51
运动	1 253	军事	74
运输	57	矿业	33

分别随机抽取样本集的 13/19、11/17、7/13、1/2、

1/3、1/5、1/7 作为测试集,剩余作为训练集,先进行各类空间的转化(每组 SVD、ICA、PCA、IDWT 降维尺度由每个 ICA 训练集的最多有效主元个数决定),再进行共计 7 次各空间分类实验。

第二组实验验证本文提出的小波分析法在外来样本中的优越性:选用数据堂网站新闻组语料库,该语料库包含 10 个类别、2 815 篇文本。在每个类别中随机抽出 10 篇,其余作为训练集。在训练集中舍去英文单词、标点符号、数字、部分停用词、极低频词,获得特征词即词典单词数 D 共计 9 700 个。SVD、ICA、PCA、IDWT 空间的维度为 2 000。表 2 是统计出的具体信息。

表 2 训练集的统计信息

统计数 类别	S ₁	S ₂	S ₃	S _i
环境	38	191	2 053	0.003 0
计算机	31	190	2 165	0.002 5
交通	103	204	540	0.002 1
教育	40	210	942	0.001 4
经济	107	315	1 072	0.004 4
军事	123	239	689	0.003 2
体育	249	440	639	0.006 0
医药	73	193	407	0.001 1
艺术	69	238	859	0.002 3
政治	345	495	420	0.005 5

该组实验选取复旦新闻组语料库中的对应新闻类,进行各空间中各类别的分类实验。

3.2 实验结果及分析

以第一组实验的实验条件和实验结果形成表 3。横向表头代表各个训练空间,纵向表头代表上文所随机抽取的 VSM 向量的在每组 SVD、ICA、PCA、IDWT 空间降维后的维度数。分类准确率(单位:%)相关结果如下。

表 3 内部样本下各类特征提取方法的文本分类

分类方法 降维尺度	SVD	ICA	PCA	VSM	DWT	IDWT
2 903	85.82	85.93	85.97	85.96	85.75	71.32
3 234	85.42	86.03	85.83	85.83	85.61	73.68
4 129	86.04	85.93	85.85	85.85	85.83	76.59

续表						
分类方法 降维尺度	SVD	ICA	PCA	VSM	DWT	IDWT
4 450	86.39	86.90	86.74	86.74	86.43	77.58
5 777	86.11	86.75	86.90	86.90	86.78	78.64
6 811	87.46	87.46	87.71	87.71	87.40	81.23
7 196	86.87	87.01	86.87	86.87	86.58	81.44

当测试集减少到一定阶段之前(本组实验中是原始数据集的 1/5),训练集会增长到一定程度,测试样本能充分利用训练集,多个空间的分类准确率整体呈增长趋势;当减少到某一阈值后,训练集的分类多样性趋于稳定、可利用率变低,导致分类误差越来越大。总体来看,传统的 VSM 空间分类准确率已有一个很高的水平;DWT 空间在维度降为 VSM 空间维度的一半后仍能 and VSM 空间的分类准确率保持一致(平均分类误差 0.2%);IDWT 空间在从原始 29 750 维降到相应低维度,有效避免了维度灾难的发生,却也付出了分类准确率降低的代价,同时可以看出,随着本文训练集的增加,IDWT 空间的分类准确率呈直线增长,这说明该空间前文所提阈值上限要高于其他空间;SVD、ICA、PCA 空间不但在维度上锐减,而且在分类准确率上亦有个好的效果,和 VSM 空间的平均分类误差为 0.30%、-0.02%、-0.001%。

值得一提的是,SVD、ICA、PCA 方法能有效提取训练样本最具有分类效果的特征(去除冗余特征)。但这些降维特征提取方法严重依赖于训练集的特性,所以会有偏差:如某特征在训练集中是重要分类特征,但随着样本的增多,可能就发现是冗余特征。而小波分析空间先于训练集的存在,故鲁棒性较强。这在下组实验中有所体现。

以第二组实验条件和实验结果形成表 4。分类准确率(单位:%)结果如下所示。

表 4 外来大样本下各类特征提取方法的文本分类

类别 分类方法	环境	计算机	交通	教育	经济
SVD	4.36	5.67	12.28	10.17	12.50
ICA	3.29	0.81	12.28	5.08	9.06
PCA	4.27	7.89	36.00	10.17	33.31
VSM	4.11	7.30	40.00	8.47	31.31
DWT	2.96	11.05	32.00	8.47	23.13

续表					
类别 分类方法	军事	体育	医药	艺术	政治
IDWT	3.20	1.10	20.00	3.39	8.69
SVD	21.62	24.10	7.84	5.41	15.43
ICA	12.16	26.74	9.80	5.27	29.00
PCA	82.43	28.81	7.84	92.57	31.15
VSM	82.43	28.89	7.84	92.30	31.54
DWT	83.78	28.97	13.73	90.81	30.76
IDWT	74.32	36.07	11.76	85.00	39.26

该表是在大样本测试集下进行分类所得结果。综合分类准确率规律基本沿袭之前实验体现出的结果。但在交通类新闻中 VSM 空间分类准确率最高,SVD 空间、ICA 空间、PCA 空间低于 VSM 空间的分类准确率,这也验证了 SVD、ICA、PCA 方法依赖于训练集的特性;体育类、政治类新闻的分类准确率 IDWT 空间最高,由表 2 发现这两类的类综合因子都很高(均大于等于 0.005 5);IDWT 空间分类准确率较低的其他新闻类的类综合因子也都很低。在 SVD、ICA、VSM、PCA、DWT 空间中具有特征不足、特征值小的政治类向量(参考表 2)会在这几个空间的分类准确率低。政治类和体育类平均能提高 VSM 空间的约 8.5% 准确率。

综上发现:(1) DWT 空间能降低高稀疏、高维的 VSM 空间近一半的维度,且分类准确率在各个实验环境下基本与 VSM 空间保持一致,这是因为本文的小波变换能保留下 VSM 空间的重要分类特征,故其与 VSM 空间的分类误差较小,但其降维尺度固定,基本为 VSM 空间的一半;

(2) IDWT 空间在类综合因子大的条件下具有明显的分类优势,能在很低维度下超越实验中其他典型空间特定类别的分类准确率。

此外,同多数特征降维方法相比,小波分析方法不严重依赖于样本的统计特征,尤其是本文的逆小波变换方法,且小波分析方法只涉及卷积运算(如 SVD、ICA、PCA 方法需要矩阵运算),所以在实现难易程度上亦有所区别。

4 结束语

本文提出了一种基于小波分析的特征提取文本分类方法。实验表明本文提出的小波空间在各个环

境中同传统向量空间下的分类误差基本一致,且能减少向量空间近一半的维度;本文提出的逆小波空间在特定条件下能对特定分类类别有更高的准确率,低维下很多特征提取方法会丢失掉分类重要特征,而根据压缩感知理论可知,高稀疏正交的本文小波空间向量能有高概率还原出最原始特征向量的重要特征部分。接下来的工作有检验小波分析法的特征提取效率是否在实验中具有一定优势,如何扩大本文逆小波空间的特定条件使其特定优势更大化。

参考文献

- [1] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-47.
- [2] 何海斌,李新福,赵蕾蕾. 基于 CCIPCA 和 ICA 降维的文本分类研究[J]. 计算机工程与应用, 2008, 44(29): 150-152, 167.
- [3] 李建林. 一种基于 PCA 的组合特征提取文本分类方法[J]. 计算机应用研究, 2013, 30(8): 2398-2401.
- [4] 朱希安,等. 小波分析的应用现状及展望[J]. 煤田地质与勘探, 2003, 31(02): 51-55.
- [5] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.
- [6] 孟佳娜. 迁移学习在文本分类中的应用研究[D]. 大连: 大连理工大学博士学位论文, 2011.
- [7] 路永和,梁明辉. 遗传算法在改进文本特征提取方法中的应用[J]. 现代图书情报技术, 2014, 4: 48-57.
- [8] 何亮亮. SVD 在文本分类中的应用[D]. 广州: 华南理工大学硕士学位论文, 2012.
- [9] 李莎莎. 复小波变换像空间的等距变换与反演公式[D]. 长春: 吉林大学博士学位论文, 2014.
- [10] 韩家炜[美]等著. 数据挖掘: 概念与技术[M]. 范明等译. 北京: 机械工业出版社, 2012: 7.
- [11] 崔治. 小波分析在超声检测信号处理中的应用研究[D]. 长沙: 湖南大学硕士学位论文, 2012.
- [12] 凤丽洲. 文本分类关键技术及应用研究[D]. 长春: 吉林大学博士学位论文, 2015.
- [13] 石光明,等. 压缩感知理论及其研究进展[J]. 电子学报, 2009, 37(5): 1070-1081.
- [14] 尹宏鹏,等. 压缩感知综述[J]. 控制与决策, 2013, 28(10): 1441-1445, 1453.
- [15] 虞湘宾,董涛. 一种离散小波变换的快速分解和重构算法[J]. 东南大学学报(自然科学版), 2002, 32(4): 564-568.
- [16] 杨杰明. 文本分类中文本表示模型和特征选择算法研究[D]. 长春: 吉林大学博士学位论文, 2013.
- [17] 何海斌,李新福,赵蕾蕾. 基于 CCIPCA 和 ICA 降维的文本分类研究[J]. 计算机工程与应用, 2008, 44(29): 150-152, 167.
- [18] 李建林. 一种基于 PCA 的组合特征提取文本分类方法[J]. 计算机应用研究, 2013, 30(8): 2398-2401.



朱晋(1993—),硕士研究生,主要研究领域为数据挖掘。

E-mail: 8627222762@qq.com



崔荣一(1962—),博士,教授,研究生导师,主要研究领域为自然语言处理。

E-mail: cuirongyi@ybu.edu.cn



怀丽波(1973—),通信作者,硕士,副教授,硕士生导师,主要研究领域为优化理论与方法、数据挖掘。

E-mail: huailibo@ybu.edu.cn