结合加权特征向量空间模型和 RBPNN 的文本分类方法①

李 敏, 余正涛

(昆明理工大学 信息工程与自动化学院, 昆明 650051)

摘 要: 提出了一种结合加权特征向量空间模型和径向基概率神经网络(RBPNN)的文本分类方法. 该方法针对传统的文本特征提取方法的不足, 根据文本中特征项的位置信息和所属类别信息定义特征权重, 然后, 依据特征项的权值计算文档特征项的频数, 通过 TFIDF 函数计算特征值并得到文本的特征向量, 最后, 采用 RBPNN 网络分类, 通过最小二乘算法求解神经网络的第二隐层和输出层之间的权值, 最终训练获得文本分类模型. 文本分类实验结果表明, 该方法在文本分类中表现出较好的效果, 具有较好查全率和查准率.

关键词: 中文文本分类; 特征提取; 位置信息; 类别信息; 加权特征向量; 径向基概率神经网络

Combination of Weighted Feature Vector Space Model and the RBPNN Text Classification Method

LI Min, YU Zheng-Tao

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China)

Abstract: In this paper, a text classification method combined weighted feature vector space model and the RBPNN are presented. According to the insufficient of traditional text feature extraction method. In the method, the weigthing about text feature is given by the text feature location information and category information, and then the feature frequency is obtained. The characteristic value is calculated using the TFIDF function after that, and the characteristic vector of text is formed. Then the weights between the second network hidden layer and output layer are decided by the least squcre algorithm, so the classification model is built. The experimental results showed that, the good recall and precision are obtained. The performance of text classification method proposed is well.

Key words: Chinese text classification; feature extraction; location information; category information; weighted feature vector; radial basis probabilistic neural network;

1 引言

由于互联网信息技术的飞速发展导致电子文档大量急剧增加, 使得文本分类技术日益重要, 成为目前一个研究热点. 文本分类是指按照预先定义的主题类别, 为文档集合中的每个文档确定一个类别. 文本的有效分类关键在于对文本信息的理解以及良好的分类方法. 近年来, 已经有许多基于统计学习的文本分类方法, 如贝叶斯算法、K-NN 邻近算法、决策树算法, 支持向量机算法及神经网络等. 基于神经网络的分类模型也是一个很有效的文本分类方法, 它可以处理线性和非线性的文本分类问题, 近来受到越来越多的学者关注[1,2]. 但是大多数关于这方面的研究主要集中在一些常见且应用较广泛的神经网络中如 RBFNN、BPNN、LVQ-SOMNN、竞争型神经网络, 采用径向基概率神经网络(RBPNN)的进行文本分类研究甚少.

径向基概率神经网络(RBPNN)是由径向基神经网络(RBFNN)和概率型神经网络(PNN)两种前馈网络发展而来的, 它在训练时间和测试速度方面比 RBFNN

① 基金项目:国家自然科学基金(60863011;61175068)
收稿时间:2012-04-21;收到修改稿时间:2012-05-14

要快的多，在输出端形成了对输入样本很好的概率估计，并且克服了模式交错的影响. 总体上它很好的实现了两种网络的折衷，比前两种网络更具有优越性，其在模式识别很多问题上已经得到了很多应用，也表现出很好的效果[3,4]. 另外，传统的文本特征提取方法一般不考虑向量中各个特征项在文档中的位置信息及它所属的领域信息，只是简单的根据词频特征来构造特征向量，其对分类效果有一定的影响[5]. 本文针对传统的特征向量提取方法的不足，提出了一种考虑特征位置信息和文本类别信息的加权特征提取方法，并通过 RBPNN 建立文本分类模型，实验结果验证了提出方法的有效性.

2 中文文本特征提取及特征向量构造

文本特征提取是文本分类系统中的一个至关重要的环节，文本特征的提取直接影响分类的性能好坏. 传统的一些特征提取方法都是基于特征对文档的具有相同影响的假设前提下[6]，但由于不同特征在文本中具有不同的重要性，因此，本文应用加权求解的构造方法对特征项在文档中的位置信息以及特征项所属类别信息分别给予特征一定的权值来反映它的重要程度，并依据所给的权值来确定特征项的特征值.

假设有一个全文档的特征项集 $C=\{T_1,T_2,\cdots T_N\}$，它满足特征项集的两个重要特征：①完备性，即特征项能够体现全部文档的内容；②可区分性，根据特征项集能将目标文档与其它文档区分开. 以 $C=\{T_1,T_2,\cdots T_N\}$ 为论域用加权求解的方法构造特征向量分为两步来计算[7,8]. 算法如下：

第一步：考虑特征词在文档中的位置信息,根据特征项位置重要程度确定权值；

①特征项出现在文章的标题和关键字部分给予较大的权值 P_{11}；

②特征项出现在摘要和正文小标题中给予一定的权值 P_{12}；

③特征项出现在正文关键词句中(例如包含“目的在于……”，“因而……”，“关键在于……”等等这些词的句子中)给予适当的权值 P_{13}；

④特征项出现在引言和中心段落中给予权值 P_{14}；

⑤对于其他情况下的特征项给予合适的权值 P_{15}；

⑥将文章中和特征项相关的代词、同义词、近义词根据语义关系将其按一定的权值 P_{16} 进行替换；

对于一个特征项按照上面所述的规则进行特征项的频数计算，如果一个特征项出现在上述所说的多个地方则将它们的分别和对应的权值相乘以后叠加，计算公式如(2)，得到特征项的初步文档频数，用于进行下面的计算

$$f = f_1P_{11} + f_2P_{12} + f_3P_{13} + f_4P_{14} \tag{1}$$

式中，f 表示特征项的初步总频数，$f_i, i=1,2,3,4$ 分别表示上述①-④部分的特征项频数.

第二步：考虑特征词与文本所属类别之间的关联信息，根据特征项被包含的类别数量确定权值

在论域 $C=\{T_1,T_2,\cdots T_N\}$ 的基础上将论域 $C=\{T_1,T_2,\cdots T_N\}$ 划分为若干个子论域 $C_1=\{T_1,T_2,\cdots T_n\}, C_2=\{T_1,T_2,\cdots T_{n1}\},\cdots C_m=\{T_1,T_2,\cdots T_{nm}\}$，子论域满足以下条件：

① $C_1\cup C_2\cup\cdots\cup C_m=C$；

② $C_i\cap C_j=\Phi,\{i\neq j, i=1,2\cdots m, j=1,2\cdots m\}$

其中 m 表示待分文本的总类别数目，$n, n1, nm$ 表示子论域中特征项数. C_1 表示该论域中所有特征词只被一类文档所包含，C_m 表示该论域中的特征项被类文档所共有.

⑦分别对属于 $C_1, C_2,\cdots, C_m$ 的特征项给予不同的权值 $P_{2i}, i=1,2,\cdots,m, P_{21}\geq P_{22}\geq\cdots\geq P_{2m}$.

⑧将每篇文档的每个特证项初步得到频数，根据它所属的子论域分别乘以相应的权值 P_{2i}，计算文档的最终频数 $\tilde{f}=f\times P_{2i}$.

依据上述构造加权特征向量空间模型算法如下描述：

Step1：以特征项集为论域，按上文所述①—⑥计算每个特征项的初步文档频数；

Step2：根据特征项集的每个子论域按照⑦和⑧构造每个特征项的最终文档频数.

Step3：按下式构造 N 篇文本的特征向量：

$$\{f_T(T_{N1}), f_T(T_{N2}),\cdots f_T(T_{Nn})\},(i=1,2,\cdots N) \tag{2}$$

$$f_T(T_{Ni})=m\lg(\frac{N}{N_i}+0.5),(i=1,2,\cdots N; j=1,2,\cdots n)$$

其中，m 表示文本特征 T_i 项在文档 N 中出现的频数，N 表示全部训练文本的文本数，N_i 表示含有特征项 T_i 的文本数目.

Step4：对以上特征向量进行归一化，可得 N 篇文本的特征向量 $\tilde{T}=\{T_{N1}, T_{N2},\cdots T_{Nn}\}$

3 基于径向基概率神经网络的文本分类器构造

在得到文本的特征向量后，将数据输入 RBPNN 模型进行分类. RBPNN 的结构图如图 1 所示，主要包含三层，第一层为输入节点层至第一隐层，第一隐层是以 Paraen 为窗函数的激活函数；第二层为第一隐层至第二隐层，第二隐层是选择性的对第一隐层输出求和；第三层为第二隐层至输出层. 在数学上对于输入向量 $\vec{X}=\{x_1,x_2,\cdots x_n\}$，RBPNN 的输出层第 i 个神经元的输出值 y_i 表示为:

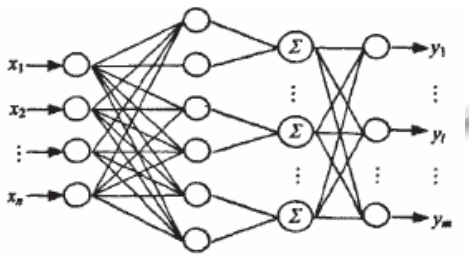

图 1 径向基概率神经网络结构

$$y_i=\sum_{k=1}^{M}w_{ik}h_k(X),i=1,2,\cdots,M \quad (3)$$

$$h_k(X)=\sum_{i=1}^{n_k}\varphi(\|X-c_{ki}\|_2) \quad (4)$$

其中，$h_k(X)$ 是第二隐层第 k 单元的输出；w_{ik} 是第二隐层第 k 个神经元和输出层第 i 个神经元的连接权重；$\varphi(\bullet)$ 是 Parzen 窗函数，c_{ki} 表示第一隐层的第 k 个类别第 i 个隐中心矢量，n_k 是第一隐层第 k 个类别的隐中心矢量数，$\|\bullet\|_2$ 是指欧拉范数，M 表示输出层的神经元数目.

从 RBPNN 网络的结构上看，只有第二隐层与输出层间的连接权重需要进行训练，目前对 RBPNN 权值训练比较好的算法是最小二乘算法，该方法对大规模样本集具有训练速度快、收敛精度高的优势. 最小二乘算法是基于误差代价函数为优化目标函数的训练算法，每步迭代的误差代价都是基于前面迭代误差的加权累加. 这里首先定义第 k 时刻的迭代加权误差目标函数为:

$$J(k)=\frac{1}{2}\sum_{i=1}^{k}\lambda^{k-1}\times tr\{[y_d(i)-y(i)]\times[y_d(i)-y(i)]^T\} \quad (5)$$

式中，λ 为加权遗忘因子；i 为输入模式矢量矩阵的次数标记；$y_d(i)$ 为第 i 个模式输入时相应的期望输出矩阵；$y(i)$ 为第 i 个模式输入时的实际输出矩阵; tr(•)为矩阵取迹运算.

最小二乘准则对权值进行估计，使输出误差与数据集正交，具体算法及步骤如下描述[9]:

Step1:给定初始权值矢量 $w(0)$,逆相关矩阵初始值 $P(0)$,高斯形状参数 $a(0)$，误差能量迭代终止值 ε,选取径向基函数中心矢量 c_{ki} $(1\le i\le N$，N 训练样本总数).

Step2: 按(4)式计算第二隐层的输出 $h_k(X)$，开始迭代, k=1;

Step3:按式(6)和(7)分别计算卡尔曼增益 $g(k)$ 与训练样本逆相关矩阵 $P(k)$ 的值.

$$g(k)=\lambda^{-1}P(k-1)h(k)(I+\lambda^{-1}h^T(k)P(k-1)h(k))^{-1} \quad (6)$$

$$P(k)=\lambda^{-1}(P(k-1)-g(k)h^T(k)P(k-1)) \quad (7)$$

Step4: 计算网络输出层第 k 步迭代时实际误差向量:

$$\varepsilon(k)=y_d(k)-h^T(k)w_i(k-1)$$

Step5: 按式(8)更新连接权矢量矩阵 $w_i(k)$ 的值

$$w_i(k)=w_i(k-1)+g(k)(y_d(k)-h^T(k)w_i(k-1)) \quad (8)$$

Step6:按式(9)和(10)更新网络的形状系数 $a_i(k)$ 的值

$$\delta_i(k)=-\frac{\partial J(k)}{\partial a_i(k)}=\lambda\delta_i(k-1)+tr\{(y_d(k)-y(k))w^T(k)\frac{\partial h(k)}{\partial a_i(k)}\} \quad (9)$$

$$a_i(k)=a_i(k-1)+\eta\delta_i(k)+(a_i(k-1)-a_i(k-2)) \quad (10)$$

Step7:按式(10)计算累积误差能量 $J(k)$ 1

$$J(k)=\lambda J(k-1)+\frac{1}{2}tr\{(y_d(k)-y(k))(y_d(k)-y(k))^T\} \quad (11)$$

Step8: 判断 $J(k)<\varepsilon$？，如果满足则训练结束，否则转 Step2.

4 实验及结果分析

实验数据采用中国期刊网上下载的文章，从中选取了 4400 篇文档 ，共有 11 个类，涵盖: 财经、医药、计算机、外语、体育、法律、农业、环境、艺术、教育、娱乐. 将下载的语料分为训练集和测试集两部分，其中训练文本 2200 篇，测试文本 2200 篇，每种类别的训练和测试文本各 200 篇.

在进行实验之前首先需对文档进行预处理，按照上文所述，先考虑文本的位置信息，根据相关的权值

(各个部分的权值如表 1 所示)计算初步的特征项频数, 再考虑特征项与文档类别的关联信息, 依据相关的权值计算特征项最终频数, 最后利用 TFIDF 函数得到文档特征向量, 特征向量的维数选取为 240. 根据得到的特征向量和文档的类别数设计 RBPNN 的网络结构, 输入节点为 240, 根据训练样本数量设计第一隐层的节点数为 2200 个, 因为第二隐层的神经元节点数和输出层的神经元节点数相同, 因而可以根据待分类的模式数确定第二隐层神经元和输出层神经元节点数为 11. 设计好网络的基本拓扑结构后, 对网络的结构参数进行设定, 将所有训练样本作为第一隐层的中心矢量 c_{ki}, 设定网络的初始化权值为 0.3, 初始相关的逆矩阵 $P(0)$, 设置 $\lambda,\eta,a,\varepsilon$ 的初始值分别为 0.95, 0.02, 0.2,0.005. 设定完结构参数后开始对网络进行迭代训练, 大约经过 2400 次迭代后网络达到收敛, 至此对网络的训练完成, 将训练好的网络进行测试, 实验在 matlab2010a 上完成[10,11].

文本分类采用 IR 中的查全率(recall)和查准率(precision)进行评价:

1)查全率(recall)= $\frac{T_n}{N}$, T_n 为通过分类算法被正确分类为 U_i 类的文本数目; N 为分类之前属于 U_i 类的文本的数目.

2)查准率(precision)= $\frac{T_n}{U_n}$, T_n 为通过分类算法被正确分类为 U_i 类的文本数目; U_n 为通过分类算法被分为 U_i 类的文本数目.

表 1　各个部分的权值取值大小

P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}
1.8	1.3	1	0.8	0.5	0.5

P_{21}	P_{22}	P_{23}	P_{24}	P_{25}	P_{26}	P_{27}	P_{28}	P_{29}	P_{210}	P_{211}
1.5	1.0	0.8	0.6	0．6	0.5	0.5	0.5	0.3	0.3	0.1

表 2　传统特征向量空间模型和加权特征向量空间模型的 RBPNN 文本分类结果

类别	传统特征向量空间模型 RBPNN 文本分类结果		加权特征向量空间模型 RBPNN 文本分类结果	
	查全率(%)	查准率(%)	查全率(%)	查准率(%)
财经	95	93.60	98.5	97.04
医药	90	94.74	94.5	99.47
计算机	93	90.73	96.5	94.61
外语	92.5	92.04	98	97.51
体育	93	94.42	97.5	98.98
法律	93.5	91.22	98.5	95.63
农业	92	94.36	95.5	98.96
环境	94	91.70	97.5	98.01
艺术	95	93.13	97	97.49
教育	92.5	90.24	98.5	95.63
娱乐	91.5	95.79	94.5	98.44
平均查全率(%)		92.91		96.95
平均查准率(%)		92.91		97.39

表 3 RBFNN、PNN、RBPNN 三种文本分类器的结果

类别	RBFNN		PNN		RBPNN	
	查全率(%)	查准率(%)	查全率(%)	查准率(%)	查全率(%)	查准率(%)
财经	94.5	93.10	96.5	95.07	98.5	97.04
医药	93	97.89	91.5	96.32	94.5	99.47
计算机	95.5	92.72	99	96.58	96.5	94.61
外语	97.5	97.01	93.5	93.03	98	97.51
体育	93.5	94.92	94	95.43	97.5	98.98
法律	94	92.16	94.5	92.20	98.5	95.63
农业	93.5	96.89	91.5	93.85	95.5	98.96
环境	95.5	93.17	97.5	95.12	97.5	98.01
艺术	96.5	96.98	93.5	92.67	97	97.49
教育	94	93.53	94.5	91.20	98.5	95.63
娱乐	93	96.88	93.5	98.42	94.5	98.44
平均查全率(%)		94.59	94.50		96.95	
平均查准率(%)		95.02	94.54		97.39	

实验中将传统的文本特征提取方法的得到的特征向量和文本加权特征提取方法得到的特征向量分别用于 RBPNN 分类器进行了实验对比, 同时还将文本加权特征提取方法得到的特征向量用于 RBFNN、PNN 两种分类器中, 用的得到的实验结果和 RBPNN 的实验结果进行比较. 表 1 所示的是各个部分的权值的取值大小, 表 2 和表 3 为实验结果. 表 1 中当各个部分的权值相等, 即各个部分权值的比值为 1:1 时, 该方法退化为传统的特征向量空间模型, 此时各个特征项对它所属文档的影响作用相同用相同, 说明权值的取值不同会影响到文本特征向量的提取和表达. 表 2 中将传统的文本特征提取方法得到的特征向量和加权文本特征提取方法得到的文本特征向量用于 RBPNN 网络进行实验, 传统特征提取方法的查全率和查准率分别为 92.91%、92.91%, 加权特征向量提取方法的查全率和查准率分别为 96.95%、97.39%.加权文本特征向量提取方法的平均查全率和平均查准率分别比传统的特征提取方法高 4.04%和 4.48%,说明加权文本特征向量的提取方法比传统的特征提取方法更有效. 这是因为本文提出的特征提出方法在特征提取时, 增强了那些在关键位置同时又是某一类文档的或者仅被少数类别文档包含的特征项的重要性, 弱化了那些普通位置同时又被大多数类别文档包含的特征项的影响, 使得同一类文档的特征向量之间具有很强的近似性, 不同类别文档的特征向量之间存在很大的差异. 表 3 比较了利用加权算法提取的文本特征向量在 RBF、PNN、RBPNN 三种神经网络中的实验结果. RBF 神经网络的平均查全率和查准率分别为 94.59%、95.02%; PNN 的平均查全率和查准率分别为 94.50%、94.54%; RBPNN 的平均查全率和查准率分别为 96.95%、97.39%. 实验结果表明 RBPNN 的分类结果最好, 说明 RBPNN 在文本分类中的性能要优于 RBF 神经网络和概率性神经网络. 另外在学习效率上 RBPNN 也是相对最好的, 训练时间不到一分钟, 而 RBF 神经网络和 PNN 分别需要 4.5 分钟左右和 3.8 分钟左右, 同时比较表 2 和表 3 中的实验结果, 可知基于加权特征向量空间模型的 RBF 和 PNN 神经网络的查全率和查准率比基于传统特征向量空间模型的 RBPNN 要高; 比基于加权特征向量空间模的 RBPNN 要低, 这个结果说明一个好的文本分类系统不仅和分类算法有关, 而且与文本特征向量的提取方法有很大的关系.

5 结语

探讨了针对文本位置信息的文本的加权特征提取方法, 并将该方法下得到的特征向量用于 RBPNN 进行实验. 实验结果表明, 结合加权特征向量空间模型

(下转第 71 页)

进行控制, 发现恶意用户可立即将该角色变为无效状态, 保证系统安全, 也提高了系统效率.

5 结语

本文针对现有 RBAC 模型的不足, 提出了一种支持空间、时间特性的角色访问控制方法—SDT-RBAC, 它能够满足基于移动用户位置的信息服务系统和空间数据库系统对访问控制模型的要求. 本文分析了空间数据的特征后, 对原有的 RBAC 模型在约束、会话等方面进行了空间扩充, 针对空间数据的特点, 提出了激活空间区域约束、激活空间角色基数约束和空间职责分离约束三类约束, 解决了有关访问控制中有关空间约束方面的需求. 下一步的工作中, 我们将对空间数据一致性维护等方面进行研究.

参考文献

1 Bacon J, Moody K, Yao W. A model of OASIS role-based access control and its support for active security. TISSEC, 2002,5(4).

2 张宏,贺也平,石志国.一个支持空间上下文的访问控制形式模型.中国科学 E 辑:信息科学, 2007,37(2):254–271.

3 Joshjbd, et al A generalized temporal role-based access control model. IEEE Trans. on Knowledge and Data Engineering, 2005,17(1):4–23.

4 Chun S, Atluri V. Protecting Privacy form Continuous High-resolution Satellite Surveillance.Technical report. CIMIC, Rutgers University, 1999.

5 Bertino E, Catania B, Damiani ML, et al. GEO-RBAC:a spatially aw are RBAC. Proc. of Symposium on Access Control Models and Teachnologies, Stockholm, 2005:29–37.

6 Belussi A, et al. An authorization model for geographical maps. Proc. of the 12th Annual ACM International Workshop on Geographic Information Systems. ACM: Washingeton DC, USA, 2004.

7 张妍,陈驰,冯登国.空间矢量数据细粒度强制查询访问控制模型及其高效实现.软件学报,2011,22(8):1872–1883.

8 Damiani M, Bertin E. Spatial Data on the web: Modeling and Management. Berlin: Springer. 2007:189–214.

9 张颖军,冯登国,陈恺.面向空间索引树的授权机制.通信学报,2010,31(9):64–73.

(上接第 89 页)

和 RBPNN 的文本分类方法具有很好的分类效果. 分类器的查全率和查准率相比较传统的特征向量空间模型的 RBPNN 分类器、加权特征向量空间模型的 RBFNN 和 PNN 分类器的结果都得到了很好的改善, 而且训练时间也相对较短. 一定程度上表明了该方法的有效性, 今后我们将进一步研究如何细化文档中特征项位置的信息及各个部分权值的优化以及 RBPNN 网络的结构参数的优化.

参考文献

1 Wang W, Yu B. Text categorization based on combination of modified back propagation neural network and latent semantic analysis. Neural Comput & Applic. 2008. 2009, (18):875–881

2 Wang Z, He YF, Jiang MH. A Comparison among Three Neural Networks for Text Classification. IEEE.2006.

3 漆随平,于慧彬,刘涛,等.基于径向基概率神经网络的气象参数状态识别.自动化仪表,2008,29(8):5–7.

4 李伦波,马广富.基于 RBPNN 的退化交通标志图像的识别算法.吉林大学学报,2008,38(6):1429–1433.

5 许增福,梁静国,由晓宇.基于 FVSM 和自组织映射网络的 Wreb 文本自动分类方法.哈尔滨工业大学学报,2004,36(9): 1168–1171.

6 庞景安.Web 文本特征提取的研究与发展.信息系统,2006 29(3):338–340.

7 郑凤萍,刘春雨.基于模糊 VSM 和 RBF 网络的文本分类方法.情报科学,2007,25(4):588–591.

8 皱娟,周经野,邓成,等.基于多重启发式规则的中文文本特征提取方法.计算机工程与科学,2006,28(6):78–80.

9 刘松,王展.基于径向基概率神经网络的人脸识别方法.计算机工程与科学,2006,28(2):83–87.

10 周开利,庄耀红.神经网络模型及其 matlab 仿真程序设计. 北京:清华大学出版社,2005.

11 matlab 中文论坛.matlab 神经网络 30 案例分析.北京:北京航空航天大学出版社,2010.