

结合聚类思想神经网络文本分类技术研究 *

朱云霞^{1,2}

(1. 南京大学 信息管理系, 南京 210093; 2. 南京人口管理干部学院 信息科学系, 南京 210042)

摘 要: 针对传统的基于神经网络文本分类算法收敛速度慢等缺点,在分析了文本分类系统的一般模型,以及在应用了互信息量的特征提取方法提取特征项后,提出了一种基于样本中心的径向基神经网络文本分类算法;并引入了聚类算法的核心思想,改进误差反向传播神经网络分类算法收敛速度较慢的缺点。实验结果表明,提出的改进算法与传统的 BP 神经网络分类算法相比,具有较高的运算速度和较强的非线性映射能力,在收敛速度和准确程度上也有更好的分类效果。

关键词: 文本分类; 神经网络; 聚类算法; 互信息量

中图分类号: TP183 **文献标志码:** A **文章编号:** 1001-3695(2012)01-0155-03

doi:10.3969/j.issn.1001-3695.2012.01.044

Text classification algorithm research based on clustering and neural network

ZHU Yun-xia^{1,2}

(1. Dept. of Information Management, Nanjing University, Nanjing 210093, China; 2. Dept. of Information Science, Nanjing Population Management Cadre College, Nanjing 210042, China)

Abstract: This paper analyzed the general model of text categorization system technology, after the application of the mutual information feature extraction,proposed a text classification algorithm based on a sample center radiate basis function (RBF) neural network, and introduced the core idea of clustering algorithm, to improve the back-propagation (BP) neural network classification algorithm convergence slower shortcomings. Experimental results show that,compared with BP network,RBF network not only has higher speed and stronger nonlinear mapping capacity,but also has faster convergence speed and better accuracy on classification effect. Combining the clustering thought,RBF text classification algorithm has larger theory research value and practical application prospect.

Key words: text classification; neural network; clustering algorithm; mutul information

文本分类是中文信息处理的一个重要的研究领域。其目标是在分析文本内容的基础上,给文本分配一个或多个比较合适的类别,从而提高文本检索、存储等应用的处理效率。目前国内外基于内容信息过滤的研究主要集中在核心算法上,基本上可以概括为用户模板的构建及其算法研究和用户模板与文本的匹配技术两个方面。这两个方面是文本信息过滤的两大关键技术。

目前已有大量的统计分类技术应用于文本自动分类中,这些传统的分类方法有着不同的缺陷。传统的支持向量机方法具有较好的分类精度,但其训练时间较长^[1];向量距离分类法^[2]具有算法简单、分类速度快等特点,但是过于依赖经过平均运算得到的中心向量,因而分类精度不高^[3];传统的神经网络对文本分类规则的判别具有独特的优势,但文本特征维数过高导致神经网络不易收敛,学习时间太长^[4]。所以目前很难找到一个执行效率高、精确率和召回率都很理想的算法^[5]。本文提出了一种基于样本中心的径向基神经网络文本分类算法,并引入了聚类算法的核心思想,改进了误差反向传播神经网络分类算法收敛速度较慢的缺点。

1 文本分类原理

文本分类系统的任务是:在给定的分类体系下,根据文本

的内容自动地确定文本关联的类别。自动文本分类即根据统计模式识别思想,将文本表示成特征向量;然后用训练文本对事先选定的分类器进行训练,直接或间接地提取出蕴涵在训练文本中有关各个文本类的统计特性,并根据这些特性确定出分类准则;最后依据这些准则对未知文本进行分类决策。

一个典型的文本分类系统如图 1 所示。

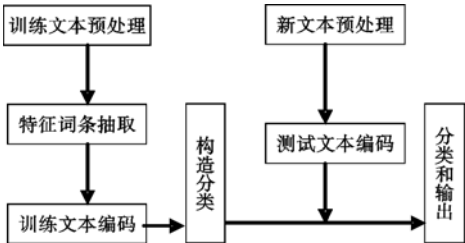


图1 文本分类系统

在对文档进行特征提取前,采用中国科学院的 ICTCLAS 分词系统进行分词处理。目前,在文本信息处理问题上,文本的表示主要采用向量空间模型。向量空间模型的基本思想是以向量的形式来表示文本,如相似度:

$$\text{sim}(d_i, C_j) = \frac{\sum_{k=1}^m (w_{jk} \times W_{ik})}{\sqrt{\sum_{k=1}^m w_{jk}^2} \times \sqrt{\sum_{k=1}^m W_{ik}^2}} \quad (1)$$

其中: d_i 属于 $C_j(j=1,\cdots,m)$, d_i 表示要处理的文本, C_j 是 d_i 所属的类, m 是分类体系的类别数; W_k 为 d_i 所属的类 C_j 的第 k 个特征相的权重; n 为 C_j 包含的特征项数目。

2 样本特征提取

向量空间模型表达文档效果的优劣直接依赖于特征项的选取以及权重的计算, 如何选择和提取文本的特征是提高分类器的效率和分类准确性的关键和基础性工作。常用的特征提取方法有潜在语义索引、文档频数、信息增益、期望交叉熵、互信息、文本证据权、CHI 统计^[6]等。

采用词和类别的互信息量 $\log\left(\frac{P(W/C_j)}{P(W)}\right)$ 作为特征项抽取的判断标准。其中:

$$P(W/C_j) = \frac{1 + \sum_{i=1}^{|D_j|} N(W, d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D_j|} N(W_s, d_i)} \quad (2)$$

为词条 W 在类别 C_j 中占的比重; $|D_j|$ 为 C_j 类的训练文本数; $N(W, d_i)$ 为词 W 在 d_i 中的词频; $|V|$ 为 C_j 类的总词数; $\sum_{s=1}^{|V|} \sum_{i=1}^{|D_j|} N(W_s, d_i)$ 为所有词在该类的词频和。而 $P(W)$ 与计算式(2)相同, 只是把所有的训练样本组成一个“总类”, 就是计算词条在总类中的比重^[7]。

$$P(W) = \frac{1 + \sum_{i=1}^{|D_j|} N(W, d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D_j|} N(W_s, d_i)} \quad (3)$$

其中: $|D_j|$ 表示的为总类的训练文本数; $N(W, d_i)$ 为词 W 在 d_i 中的词频; $|V|$ 为总类的词条数; $\sum_{s=1}^{|V|} \sum_{i=1}^{|D_j|} N(W_s, d_i)$ 为所有词在总类的词频和。

从上面的定义可以看出: 互信息的意义是词条与类别的相关程度。通常可以认为, 互信息越大, 词条与相应类别的相关性也越大。

3 改进的神经网络文本分类

神经网络是一种模仿人类神经系统的数学模型。它采用了大量简单的处理单元即神经元来处理信息, 神经网络运用多种算法来修改这些神经元的权值和阈值, 再经过多次训练达到理想结果, 目前在文本分类领域得到广泛应用的是前馈式神经网络。但由于表征文本的特征项众多, 利用神经网络预测如果输入量特别多, 而神经网络不能将输入信息空间维数简化, 过量冗余数据易造成神经网络训练过度, 影响了文本分类的精度。

利用神经网络作文本分类的一般方法是将文档 d_i 采用空间向量模型表示成一个向量, 作为神经网络的输入, 其中, 输入层的每个节点对应着输入向量的每个特征, 输出层的节点对应着某个文档是否属于类别 C_j 。

径向基函数(RBF)网络是在借鉴生物局部调节和交叠接受区域知识的基础上提出的一种采用局部接受域来执行函数映射的人工神经网络。在人的大脑皮层区域中, 局部调节及交叠的感受域是人脑反应的特点。基于感受域的这一特点, Moody 和 Darken 提出了一种神经元网络结构, 即 RBF 网络。RBF 神经网络是由输入层、隐层和输出层三层神经元构成的

典型前向神经网络。输入层节点传递输入信号到隐含层, 隐含层节点由像高斯函数那样的辐射状作用函数构成, 而输出层节点通常是简单的线性函数。

RBF 分类算法基本思想是用径向基函数作为隐层神经元的基, 构成隐层空间, 实现输入空间到输出空间的映射。

构成 RBF 神经网络隐层的径向基函数是一种局部分布、中心对称衰减的非负非线性函数:

a) 局部分布是指当且仅当输入向量落在输入空间一个很小的指定区域内隐层神经元的径向基函数才会作出有意义的非零回应。

b) 中心对称衰减是指与径向基函数中心点 c_j 径向距离相同的输入都会产生相同的输出, 并且输入向量与 c_j 越近, 隐层神经元的回应越大。

RBF 神经网络的隐层径向基函数选用如下非线性高斯函数:

$$\varphi(x) = \exp\left(-\frac{x^2}{\sigma^2}\right) \quad (4)$$

隐层神经元采用径向基函数使得 RBF 神经网络模拟了人脑中局部调整、相互覆盖感受域的特性。因此, RBF 神经网络是一种局部逼近神经网络。现已证明了 RBF 神经网络能以任意精度逼近任一连续函数, 并且可以根据问题确定相应的网络结构, 训练速度快, 不存在局部最小问题。RBF 神经网络的特性使其在越来越多的领域取代 BP 神经网络。目前 RBF 神经网络广泛应用在模式识别、函数逼近、信号处理以及自动控制等多个领域。

RBF 神经网络的输出层对隐层神经元的输出进行线性加权组合, 并增加一个偏移量 w_0 , 可表示为

$$f_j(x) = \sum_{i=0}^H w_{ij} \varphi_i(x); j=1, 2, \cdots, J \quad (5)$$

其中: H 和 J 分别表示隐层和输出层神经元个数; $x \in \mathbf{R}^n$ 表示输入向量; w_{ij} 为隐层第 i 个神经元和输出层第 j 个神经元之间的连接权值。

典型的聚类过程主要包括数据(或称之为样本或模式)准备、特征选择和特征提取、接近度计算、聚类(或分组)、对聚类结果进行有效性评估等步骤。K-means 聚类算法是聚类分析中使用最为广泛的算法之一, 该算法取定 k 类和选取 k 个初始聚类中心, 按最小距离原则将各样本分配到 k 类中的某一类, 之后不断地计算类心和调整各样本的类别, 最终使各样本到其判属类别中心的距离平方之和最小。算法步骤如下:

a) 针对 n 个样本, 任选 k 个样本作为初始聚类中心($z_1, z_2, \cdots, z_k$);

b) 对每个样本 x_i 找到离它最近的聚类中心 z_r , 并将其分配到 z_r 所标明的类 u_r ;

c) 采取平均的方法计算重新分类后的各类心;

d) 计算 $D = \sum_{i=1}^n [\min_{r=1, \cdots, k} d(x_i, z_r)]^2$;

e) 若 D 值收敛, $\text{return}(z_1, z_2, \cdots, z_k, U)$, 并终止本算法, 否则转至 b)。

对于大数据集 K-means 算法是相对可伸缩的和高效率的, 因为它的时间复杂度是 $O(nkt)$, 其中 n 是样本数, k 是聚类数, t 是迭代次数。通常 $k < t$, 且 $t < n$ 。用 K-means 聚类算法时, 对于类内紧密、类间远离的聚类结构, 它的聚类效果较好。

4 实验结果与分析

本文采用中国科学院计算技术研究所中文自然语言处理开发平台发布的由复旦大学计算机信息与技术系国际数据库中心自然语言处理小组所提供的中文语料库,对文本分类器进行分类测试。对 581 个文本样本进行信息编码,得到 10 维文本的信息编码向量 581 个,其中 140 个作为训练样本,其余 441 个作为测试样本,在 MATLAB 环境下分别进行 BP 和 RBF 神经网络的分类算法实现;再利用 K-means 聚类方法作为 RBF 神经网络分类算法的核心思想,进行 RBF 分类,并与 BP 分类算法比较。文本分类系统的最主要的两个指标是查准率和查全率,查准率(精度)是衡量检索系统信号噪声比的一种指标,即检出的相关文献与检出的全部文献的百分比,表示为:查准率 = (检索出的相关信息量/检索出的信息总量) × 100%。使用泛指性较强的检索语言(如上位类、上位主题词)能提高查全率,但查准率下降。查全率(召回率),是衡量某一检索系统从文献集合中检出相关文献成功度的一项指标,即检出的相关文献与全部相关文献的百分比,表示为:查全率 = (检索出的相关信息量/系统中的相关信息总量) × 100%。使用专指性较强的检索语言(如下位类、下位主题词)能提高查准率,但查全率下降。

所谓的分类正确就是指自动分类结果与人工分类结果吻合。取其中三类的统计结果如表 1 所示。

表 1 三种网络分类算法统计结果

网络	查全率			查准率			训练时间 /s
BP	0.9492	0.9444	0.9574	1.0000	0.9714	0.9783	22.782
RBF	0.8305	0.8194	0.6170	0.8909	0.8676	0.9667	19.156
RBF + 聚类	0.8136	0.7639	0.5319	0.8889	0.8871	0.8929	0.890

从表 1 中可以看出,RBF 网络分类结果不如 BP 网络的好,但使用给定样本特征值的平均值作为聚类中心的径向基网络分类结果较好,且训练时间很短。这证明了本文提出的算法要比传统的神经网络文本分类算法要好得多。对以上三类样本在利用结合聚类的 RBF 文本分类算法时,修改高斯函数宽度参数,不同高斯函数宽度参数对实验结果的影响如图 2

所示。

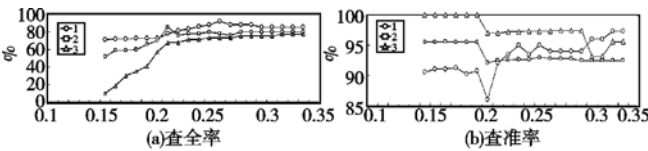


图2 不同高斯函数宽度参数对结果的影响

5 结束语

本文以神经网络为基础,借鉴了聚类算法的思想,采用样本中心作为 RBF 分类算法的核心,并与 BP 神经网络分类算法进行了比较。从实验结果可以看出,在收敛速度和分类效果上,结合聚类的 RBF 文本分类算法好于 BP 神经网络分类算法,充分体现了改进后 RBF 分类算法的简洁和时效性。径向基函数的宽度参数会影响分类的准确程度和实验误差,查全率和查准率不能同时提高,随着宽度参数的增大,误差会变小。结合了聚类思想的 RBF 文本分类算法有一定的理论研究价值和实际应用前景。

参考文献:

[1] 包剑,冀明,冯军. 基于模糊支持向量机的文本分类[J]. 辽宁工程技术大学学报:自然科学版,2010,29(5):974-977.

[2] 刘金岭,严云洋. 基于上下文的短信文本分类方法[J]. 计算机工程,2011,10(2):511-514.

[3] 史晶蕊,郑玉明,韩希. 人工神经网络在文本分类中的应用[J]. 计算机应用研究,2005,22(10):218-221.

[4] DEBOLE F, SCABASTIANI F. An analysis of the relative hardness of recuters-21578 subsets [J]. Journal of the American Society for Information Science and Technology,2004,56(6):584-596

[5] AHN B S, CHO S S, KIM C. The integrated methodology of rough set theory and artificial neural network for business failure prediction[J]. Expert Systems with Applications, 2000,18(2):65-74.

[6] 熊忠阳,黎刚,陈小莉,等. 文本分类中词语权重计算机方法的改进与应用[J]. 计算机工程与应用,2008,44(5):187-190.

[7] 张玉芳,彭时名,吕佳. 基于文本分类 TFIDF 方法的改进与应用[J]. 计算机工程,2006,32(9):76-78.

(上接第 154 页)

参考文献:

[1] 李志纯,黄海军. 先进的旅行者信息系统对出行者选择行为的影响研究[J]. 公路交通科技,2005,22(2):95-99.

[2] BELL D E. Regret in decision-making under uncertainty[J]. Operations Research,1982,30(5):961-981.

[3] ZEELENBERG M, PIETERS R. Comparing service delivery to what might have been: behavioral responses to disappointment and regret [J]. Journal of Service Research,1999,2(1):86-97.

[4] COOKE A D J, MEYVIS T, SCHWARTZ A. Avoiding future regret in purchase-timing decisions[J]. Journal of Consumer Research, 2001,27(4):447-459.

[5] SIMONSON I. The influence of anticipating regret and responsibility on purchase decisions[J]. Journal of Consumer Research,1992, 19(1):105-118.

[6] ZEELENBERG M, BEATTIE J. Consequences of regret aversion 2: additional evidence for effects of feedback on decision making[J]. Or-

ganizational Behavior and Human Decision Processes,1997,72(1):63-78.

[7] SHEFRIN H, STATMN M. The disposition to sell winners too early and ride losers too long[J]. Theory and Evidence Journal of Finance,1985,40(3):777-790.

[8] 李新路. 后悔厌恶心理对投资者行为影响的实证分析[J]. 河北经贸大学学报,2006,27(2):40-44.

[9] TORRES F J, HUBER M. Learning a causal model from household survey data using a Bayesian belief network [J]. Transportation Research Board Annual Meeting,2003,1836:29-36.

[10] JANSSENS D, WETS G, BRIQS T, et al. Improving the performance of a multi-agent rule-based model for activity pattern decisions using Bayesian networks[J]. Journal of the Transportation Research Board,2004,1894(1):75-83.

[11] 李晓毅,徐兆棣,孙笑微. 贝叶斯网络的参数学习研究[J]. 沈阳农业大学学报,2007,38(1):125-128.

[12] 林元烈,梁宗霞. 随机数学引论[M]. 北京:清华大学出版社,2003:331-343.