

文本信息自动分类系统 ITC98(Ⅱ)^①

——基于 BP 网络的文本分类子系统

陶 兰^② 申军霞

(中国农业大学电子电力工程学院)

摘 要 设计并实现了文本信息自动分类系统 ITC98 的核心模块——基于 BP 网络的文本分类子系统。介绍系统的分类策略及根据分类问题需求确定 BP 网络结构和参数的方法。实例测试表明,系统分类精度和效率均达到要求。

关键词 文本分类; BP 网络; 分类算法

分类号 TP 391; TP 387

Intelligent Text Classifying System (Ⅱ):

BPN-based Text Classify Subsystem

Tao Lan Shen Junxia

(College of Electronic and Electric Power Engineering, CAU)

Abstract The BPN-based Text Classifying Subsystem, the core module of ITC98, is designed and carried out. The overview of the classifying algorithm and technique which regulate BPN structure and parameters are put forward. The example testing result show that the accuracy and efficiency of system classifying is satisfied.

Key words intelligent text classify; BPN; classify algorithm

人类进行文本信息分类的特点是:快速浏览,捕捉类别特征信息,以此作为类别判断依据;分类中重点注意标题、摘要、首段及段首句;若标题或摘要中有足够类别特征信息,则直接判断类别,否则进一步浏览正文;分类过程中,经验判据是类别特征信息提取及类别判断的依据。

笔者设计并实现了一个基于 BP(Back Propagation)网络的文本分类子系统。它与文献[1]介绍的自动分词子系统和编码子系统一起实现了中文文本信息的自动分类。

1 文本分类子系统工作原理

本文建造的文本分类子系统采用的分类策略是:先标题后正文,逐层提取特征信息;先学习积累分类知识,后依经验作出类别判断。本子系统的总体结构主要包括依据文本标题分类模块和 BP 网络文本信息分类器。

依据文本标题分类模块。此模块首先对已标注文献按下式计算标题类别特征信息编码值:

收稿日期:1999-01-19

①清华大学智能技术与系统国家重点实验室开放课题基金资助项目

②陶 兰,北京清华东路 17 号 中国农业大学(东校区)64 信箱,100083

$W_i = 2 \sum_{j=1}^N [2/(1+x_j)]$ 。式中: W_i 为文本标题第 i 类特征信息权重 ($1 \leq i \leq M$, M 为所分类别的总数); x_j 为第 i 类第 j 个特征信息的级别; N 为第 i 类特征信息总个数。将 W_i 与某一给定阈值 P 进行比较。这里 P 是由人类先验知识确定的, 一般较大, 以保证依据标题信息判断类别所属的正确性。若有 1 个且仅有 1 个 W_i 满足 $W_i \geq P$, 则可判定文献属于第 i 类; 否则将文献输入 BP 网络分类器, 依正文判定文献所属类别。

BP 网络文本信息分类器。在本系统中, 利用 BP 网络作为已量化文本的分类器, 选择足够的各类别样本供 BP 网络学习, 各类别样本的选择既具有代表性又兼顾普遍性, 在样本数量不太多的情况下, 尽可能包含较多的类别信息。在学习的基础上, 利用 BP 网络完成待分类文献的分类。

2 BP 网络文本分类器的设计与实现

实际应用中可根据待分类文献定义的类别数目、分类的难易度及分类的精度要求, 对 BP 网络的结构作不同的设计。下面简要说明此模块的设计与实现原则。

- 1) 输入层。输入层神经元个数与文献编码维数一致。
- 2) 输出层。输出层神经元个数由类别主题决定的分类数目确定。
- 3) 隐含层。隐含层的层数及各层神经元的数目, 可以根据问题规模大小确定。

3 系统测试及结果分析

实验内容选自 Internet 网上《计算机世界》国内外行业信息栏目, 共选择了 100 篇不同形式的文献。实验样本的篇幅从 100 字到 3 000 字不等。这些样本分别属于“企业动态”、“产品介绍”和“市场行情”等 3 个类别。例如:

产品介绍类

正文: 中国地图出版社电子地图室日前推出了两种电子地图——《中国地图集》和《香港特别行政区电子地图》。《中国地图集》光盘版包含 1997 年最新审定的中华人民共和国各省、自治区、直辖市、特别行政区……

企业动态类

正文: 为了答谢广大客户对彼岸电脑的厚爱, 彼岸电脑金秋再次降价, 每款机型降幅都在百元以上, 并在原来基础上又出新款……

市场行情类

正文: 宏基电脑公司 9 月 7 日在京宣布, 其商用台式机系列产品全面降价, 平均降幅 6.6%。降价的台式机有……

由以上例子可见, 3 个类别的文献内容相似度较高, 因而分类复杂度也较高。

实验过程和方法如下。

首先对于标题中有足够类别特征信息的文献, 本系统直接由“依标题分类模块”判断文献所属类别。例如:

标题: /台湾/Vigor/扫描仪/新品[P1]/亮相[P1]/京城

正文: /10/月/8日/, /台湾/恒瑞/集团/在/京/推出[P1]/两款/扫描仪/新品[P1]/——/

Vigor-6p/和/Vigor-12p/。/这/两种/产品/的/最大/特点/是/……。/

上例标题中有 2 个 P 类 1 级特征词,其类别权值经计算为 4,大于给定的阈值(这里的给定阈值为 3.5),因此可以由“标题分类模块”直接判定其属于 P 类文献集。

对于大部分不能通过“依标题分类模块”判定类别的文献,则输入 BP 分类器完成分类。其步骤如下:

1)确定 BP 网络的输入和输出层神经元数目。实验中根据问题精度要求,确定每类文本特征信息为 2 级,则 3 类实验样本经编码后为 6 维。由此确定 BP 网络输入神经元个数为 6。输出神经元的个数选择与所分类别数相同,为 3,每一个输出神经元对应一个类别。

2)建立样本集。选自网上的已知类别的实验样本输入自动分词子系统和编码子系统处理后,将已编码实验样本分为 2 组:一组作为训练样本集;另一组作为测试样本集。选择各类别总共 34 个样本作为训练样本。训练样本集如表 1 所示。

表 1 训练样本集

序 号	训练样本编码输入						期望输出		
	x_1	x_2	x_3	x_4	x_5	x_6	y_1	y_2	y_3
1	1.000 0	0.634 0	22.472 0	0.442 5	0.367 8	0.367 8	1	0	0
2	1.277 5	0.367 8	5.065 5	0.419 2	0.367 8	0.367 8	1	0	0
3	0.367 8	0.367 8	3.072 4	0.392 7	0.367 8	0.367 8	1	0	0
⋮									
34	1.863 5	0.381 3	0.367 8	0.381 3	1.277 5	0.374 5	0	0	1

3)确定网络参数。实验测定结果:当选择 2 个隐层,每层 4 个神经元, $\alpha=0.7$, $\beta=0.6$,个别误差为 10^{-5} ,总体误差为 10^{-6} ,学习次数为 10^4 时,系统分类精度和工作效率最佳。

4) BP 网络学习。

5)系统测试。将已编码的测试样本输入子系统以测试分类效果。将输出值中与最大一维对应的类别作为样本所属类别。测试样本的期望输出与实际输出值举例如表 2 所示。

表 2 测试样本期望输出与实际输出

序 号	期望输出			实际输出		
	y_1	y_2	y_3	y'_1	y'_2	y'_3
1	1	0	0	0.884 041	0.000 998	0.023 396
2	1	0	0	0.454 050	0.421 639	0.090 044
3	1	0	0	0.722 661	0.067 898	0.065 515
⋮						
43	0	1	0	0.361 524	0.534 879	0.089 121
44	0	1	0	0.135 674	0.745 821	0.106 354
45*	0	1	0	0.104 310	0.120 472	0.664 672
46*	0	1	0	0.516 926	0.431 195	0.099 514
⋮						
70	0	0	1	0.007 783	0.072 088	0.896 750
71*	0	0	1	0.676 300	0.110 103	0.071 746
72	0	0	1	0.021 970	0.092 823	0.822 551
73	0	0	1	0.021 970	0.092 823	0.822 551

说明:带“*”的样本,实际判断类别与期望判断类别不符。

对表2中期望输出与实际输出结果进行比较分析,可得以下结论。

1)系统的分类精度符合分类要求。系统分类精度计算式如下:

$$\text{分类精度} = (\text{实际类别判断与期望类别判断相同的数目} / \text{测试样本总数目}) \times 100\%$$

用此式对表2中的数据进行统计计算,可得此分类系统的分类精度为93.6%,基本符合分类精度要求。

2)个别样本(如表2中带有星号的样本)出现类别判断错误。其原因有以下2种可能:

a. 样本某2个类别特性相近,在分类中无法将其区分,例如样本46,对应于第2类与第3类的类别特性相近,BP分类器出现分类错误。样本46的文本举例如下:

正文:北大方正日前发布消息,方正颐和 Aha5133 笔记本电脑降价至两万元以内,市场定价为 19 977 元,表示 1997 年 7 月 1 日起香港正式回到祖国的怀抱。方正颐和 Aha5133 是方正笔记本电脑的主流机型,厚度仅为 42 毫米,重 2.7 公斤,属同档次笔记本电脑产品中的轻薄一族。

可见,此文兼有“市场行情”和“产品介绍”2类文献的特征,所以极易混淆。

b. 样本本应属于某一类,但却包含了较多的另一类的类别特征信息,BP网络分类器无法正确分类。例如样本71,其文本内容如下:

正文:由于 WAN 带宽价格的持续居高,AT&T 公司近日**宣布**对其主要服务**采取提价措施**。因为 IS 和公司间在国际网络上争夺空间的竞争愈演愈烈,这次提价对租用高性能线路的用户影响尤为显著。AT&T 公司还**提出**愿意重新协商与用户的**合同**,以帮助用户减轻一些压力。

此文本属于第3类,但文中黑体字显示为第1类的特征词,此段文本中包含了5个第1类的特征词,因而BP分类器判断类别时出现错误。

对于以上第1种情形,可以采用将此类样本加入到训练样本集中让网络学习的方法得到解决;但对于第2种情形,若将此类样本加入训练样本集,会影响其他样本的正确分类,本系统目前难以处理此类样本,而这类问题即便是人工进行分类也同样易出现误分。

因此,基于BP网络的文本分类子系统的分类精度可以达到文本分类要求。

4 结束语

本文重点介绍了基于BP网络的文本信息分类子系统的设计思路及实现方法。测试结果表明,此分类子系统满足分类精度要求。

参 考 文 献

- 1 申军霞. 文本信息自动分类系统及其应用研究:[学位论文]. 北京:中国农业大学,1998. 14~17,28~40