文章编号: 1000-8608(2007)01-0131-05

基于改进 BP网络的中文歧义字段分词方法研究

张 利*[1], 张立勇[1], 张晓淼[1], 耿铁锁[2], 岳宗阁[3]

(1.大连理工大学 电子与信息工程学院, 辽宁 大连 116024;
2.大连理工大学 国有资产处, 辽宁 大连 116024;
3.大连理工大学 附属医院, 辽宁 大连 116024)

摘要: 文本挖掘中中文歧义字段的自动分词是计算机科学面临的一个难题. 针对汉语书写时按句连写,词间无间隙,歧义字段分词困难的特点,对典型歧义中所蕴含的语法现象进行了归纳总结,建立了供词性编码使用的词性代码库. 以此为基础,通过对具有特殊语法规则的歧义字段中的字、词进行代码设定,转化为神经网络能够接受的输入向量表示形式,然后对样本进行训练,通过改进 BP神经网络的自学习来掌握这些语法规则. 训练结果表明: 算法在歧义字段分词上达到了 93.13% 的训练精度和 92.50% 的测试精度.

关键词: 文本挖掘; 歧义字段; 自然语言处理; 神经网络
中图分类号: TP391.1 **文献标识码**: A

0 引 言

中文文本的大字符集上的连续字串特点,使其在计算机自动切分上存在一定的难度[1],分词理论与方法的设计直接影响到中文自动分词系统的实现及效果. 近年来,语言学界、人工智能领域和情报检索界的学者们对中文自动分词与自动标引进行了大量的研究与实践,找到了多种解决中文分词的方法. 同时,各种分词系统也不断建立,分词系统在运行速度、准确度等方面都已具有了研究应用的价值[2~4]. 目前常用的分词算法有正向最大匹配法(MM)[5]、逆向最大匹配法(RMM)[6]、逐词遍历法[7]、最佳匹配法(OM法)[8]、双向扫描法等. 虽然目前分词领域的发展已经有了一定突破,但是这些分词方法在面对语言不断变化时仍显得适应性不强,还需继续进行研究,以期形成更加完善的分词方法.

鉴于神经网络具有很强的自组织、自学习性,在对典型歧义语法现象进行研究的基础上,本文利用神经网络进行文本歧义字段分词,使之应用于中文文本的词语切分,形成独自的输入转换规则和输出解释规则. 通过对样本的测试训练,最终实现将目标误差率控制到限定范围内,进而形成一种有效的分词方法.

1 神经网络分词技术

1.1 BP网络模型

BP网络模型如图 1所示,包括输入层、中间层(隐层)和输出层. 设输入层为 I,即有 I个输入信号,其中的任一输入信号用 i表示;隐层为 J,即有 J个神经元,其中任一神经元用 j表示;输出层为 K,即有 K个输出神经元,其中任一输出神经元用 k表示. 输入层与隐层的连接权值为 w_{ji},隐层与输出层的连接权值为 w_{kj}.

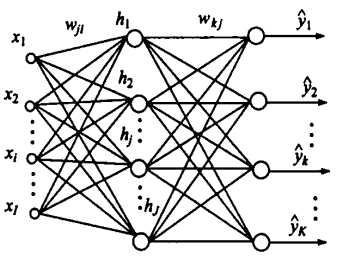

图 1 带有一个隐层的 BP网络
Fig. 1 BP network with a hidden layer

1.2 神经网络分词特点

(1) 知识以并行、分布的数值方式存储,处理

收稿日期: 2005-12-17; 修回日期: 2006-11-09.
基金项目: 国家自然科学基金资助项目(60573172).
作者简介: 张 利*(1971-),男,博士,副教授.

机制为动态演化过程;

(2) 字词或抽象概念与输入神经元对应,切分方式与输出神经元对应,即存在一个输入、输出逻辑概念到输入、输出模式的转换;

(3) 可适应不断变化的语言现象,包括结构的自组织和权值自学习;

(4) 新知识的增加不影响神经网络处理的速度,这与基于规则的产生式方法有很大区别;

(5) 知识表达简洁、学习功能强;

(6) 知识库容易维护和更新;

(7) 分词速度快、精确度较高.

2 神经网络分词的实现

利用神经网络原理进行分词,基本点是要解决分词知识的输入、学习和理解. 由于 BP及其改进算法可解决诸如学习、分类和模式识别等问题,本文采用 BP网络进行分词训练.

2.1 数据预处理

为了使神经网络能够接受外部数据,就要进行数据预处理. 首先从输入文字流中取出语句,然后对语句中的字、单词进行编码,把字、词变成神经网络能够识别、学习和存储的字符形式,送至神经网络. 具体实现时,对于一些固定的特殊后缀词、助词采用 Unicode码进行表示. 对于不同类型单词所对应的词性进行代码设定,根据对句法分析的经验给出歧义字段的分词规则. 典型的歧义字段语法规则如下:

(1) 歧义字段若和前面字构成名词,则首字切分,否则自己成词.

如表 1所示,样例 11、12可以利用这条规则进行切分处理.

(2) 若歧义字段为动词短语,则如果前面接动词则自行成词,前面若是名词则首字单切. 样例 1、13、14、17可以利用这条规则进行切分处理.

(3) 如果交集字串与其直接后继字组成形容词,则将该歧义词的首字单切,否则确认该歧义词为词.

样例 4、10可以利用这条规则进行切分处理.

(4) 如果歧义字串中的直接前趋词是数词,则歧义字段的首字单切.

样例 2、9可以利用这条规则进行切分处理.

(5) 如果歧义字段的前趋词中有介词,则歧义字段的首字单切,否则该歧义字段成词.

样例 3、8可以利用这条规则进行切分处理.

(6) 如果歧义字段的后继词有趋向动词或助词,则尾字单切,否则该歧义字段成词.

样例 5、6、7可以利用这条规则进行切分处理.

(7) 如果歧义字段后继名词的义项中含有“数学式子”或“扣子”之类义素,则歧义字段的尾字单切,否则该歧义字段成词.

样例 18可以利用这条规则进行切分处理.

表 1 目前国内公开发表汉语分词论文中的典型语句

Tab. 1 Typical sentences in Chinese segmentation papers published at present

No.	样例	No.	样例
1	他 /吃 /切面 /了	11	一群 /蜂 /似的 /散开
2	这里 /真 /热闹	12	东西 /拍卖 /完 /了
3	他 /打 /梯上 /下来	13	把 /图形 /拼接 /起来
4	物理 /学 /起来 /很难	14	他 /一阵 /风 /似的 /跑 /了
5	学生 /活 /下去 /吃力	15	校 /学生会 /发表 /公告
6	这支 /歌 /太 /平淡无味	16	该 /学生 /会 /发表 /文章
7	球拍 /卖 /完了	17	球 /打 /桌上 /滚 /下
8	他 /烤 /白薯	18	他 /学会 /了 /解 /方程
9	他 /吃 /烤白薯	19	今天 /空位子 /很多
10	他 /切 /面 /了	20	如今 /天空 /总是 /很 /蓝

2.2 初始值的选择

由于系统是非线性的,初始值的选取对于学习过程能否得到收敛的结果关系很大. 一个重要的要求是: 初始权值在输入累加时使每个神经元的状态值接近于零,权值一般随机选取,而且要比较小. 输入样本也同样希望进行归一化处理,使那些比较大的输入仍落在传递函数梯度较大的地方.

具体数值设置应注意以下几个因素:

(1) 学习开始时,各隐含层连接权系数的初值应以设置较小的随机数较为适宜.

(2)采用 S型激发函数时,输出层各神经元的输出只能趋于 1或 0,不能达到 1或 0.

(3) 在设置各训练样本时,期望的输出分量 d_{pk} 不能设置为 1或 0,以设置为 0.9或 0.1较为适宜.

(4)学习速率 Z的选择,在学习开始阶段,Z选较大的值可以加快学习速度. 学习接近收敛时,Z值必须相当小,否则权系数将产生振荡而不收敛. 平滑因子 T的选值取 0.9左右.

2.3 输入和输出层的设计

输入神经元可根据需要求解的问题和数据表

示方式确定. 输出层的维数可根据使用者的要求确定. 如果将 BP网络用做分类器,类别模式一共有 m个,那么输出层神经元的个数为 m或 $\log_2 m$.

2.4　隐层的设计

隐层神经元的数目选择往往需要根据设计者的经验和多次实验来确定,因而不存在一个惟一的解析式来表示. 以下 3个公式可用于选择隐层单元数:

(1) $\sum_{i=0}^{n} C_{n_1}^{i} > k$,其中 k为样本数, n_1为隐单元数, n为输入单元数,如果 $i > n_1$, $C_{n_1}^{i} = 0$;

(2) $n_1 = \sqrt{n+m}+a$,其中 m为输出神经元数, a为 [1, 10]之间的常数;

(3) $n_1 = \log_2 n$.

2.5　改　进

(1) 基本 BP算法采用梯度下降法向学习目标逼近

BP算法在学习过程中有时能量函数会陷于局部极小状态,这是由于对激活函数的求导引起的. 通常激活函数选取 Sigmoid函数形式,函数特性曲线如图 2所示.

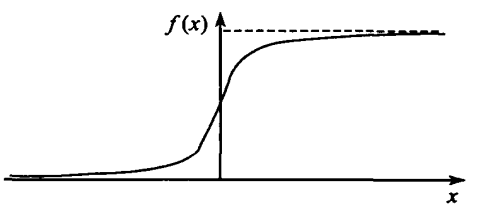

图 2　神经元饱和特性曲线

Fig. 2　Curve of saturation characteristic of the nerve cell

当 x取值很大或很小时,神经元处于饱和区. 这样,在饱和区算法对网络权值的修正量趋于零,网络陷入局部极小点,收敛速度极为缓慢. $f'(x)$易趋于零是网络陷入局部极小点的重要原因,因此,将 $f(x)$修改为 $f^{*}(x) = f(x) + \eta$, $\eta \in (0.05, 0.1)$,从而使导数得到提升,不再趋于零而是趋于一个小的数 η.

另外,在 BP算法中,学习速率 α在一定程度上决定了网络的收敛速度. α过小会导致收敛速度缓慢; α过大却又会导致在极值点附近振荡的可能性加大,乃至反复振荡而难以收敛. 因此,采用"矩方法" 对算法进行改进,加上矩量项

$$\Delta w(t+1) = -\alpha \frac{\partial E}{\partial w} + \beta \Delta w(t) \qquad (1)$$

其中 $\beta \in (0, 1)$为矩参数. 这样,在通过能量函数曲面上的平坦区域时,可以认为 $\Delta w(t+1) \approx \Delta w(t)$,则网络的学习速率近似为

$$\alpha^{new} = \alpha^{old}(1-\beta) \qquad (2)$$

从而加快了网络在这一区域的学习速度. 而在振荡比较剧烈的区域,矩量项便可忽略不计,从而相对于平坦区域减小了学习步长,避免了振荡加剧. 从总体效应来看,矩方法的引入实现了学习步长的动态变化,提高了网络的自适应能力,加快了收敛速度.

(2) 基本 BP算法以及改进的 BP算法都是对样本集进行学习训练最终得到最佳权值

当样本数量非常大,而且需要经常加入新样本进行样本集扩充,基本 BP算法和各种改进 BP算法都必须对网络重新进行训练,训练时间明显增加.

BP的学习过程实质上是误差梯度的下降过程. 可以求出权值的更新量,这时新的权值为 $w_{new} = w_{old} + \Delta w$,同时这时的总误差 E也可以求出,当 E达到规定的最小范围时,可以求出此时各层的权值大小.

这时,如果加入一个新的样本,传统的方法是将权值重新初始化,对网络重新进行训练. 这样做使得已有的训练结果完全打乱,需大量的再训练时间. 本文将 N个样本训练得到的最优解权值记忆下来,作为下一次训练的联想记忆权值,在加入一个新样本时用 N个样本训练得到的最优解权值代替重新初始化的权值.

原来的总能量函数为

$$E_k(w) = \sum_{m=1}^{M} E_k^m(w) = \frac{1}{2}\sum_{m=1}^{M} (y_k^m - \hat{y}_k^m)^2 \qquad (3)$$

现在的总误差 $E = E_k + \Delta E_{k+1}$,由于前面的 N个样本对系统贡献的误差 E_k趋于零,权值记忆后新的系统误差相对于重新初始化权值得到的初始时刻误差明显减小.

3　实验结果及分析

文本样例选取如下.

样例 a 他吃烤白薯.

样例 b 他马上回来.

样例 c 物理学不好.

样例 d 没有人能行.

样例 e 这歌太难听.

下面针对上述 5个样例进行网络训练,具体

过程如图 3所示.

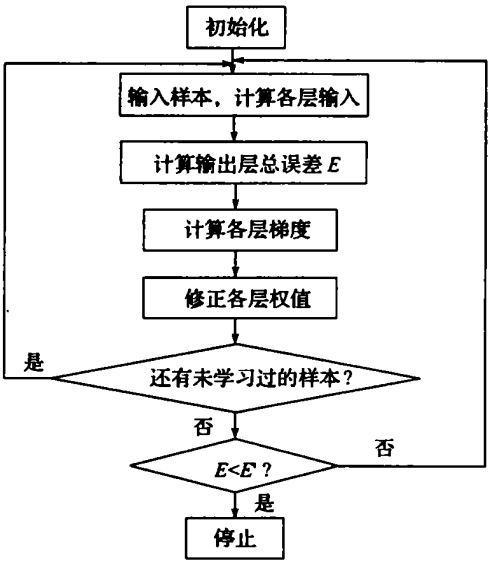

图 3 样本训练流程图

Fig. 3 The flow chart of samples training

本文采用的 BP网络输入层有 20个神经元,隐含层有 8个神经元,输出层有 5个神经元. 隐含层和输出层均采用 S型激活函数. 训练次数分别为 200、500和 1 000次. 从得出的输出值中可以看出分词后的网络输出结果.

表 2~ 6分别列出了上述 5句样例在训练 200、500和 1 000次的网络输出值.

表 2 样例 a的网络输出

Tab. 2 The network output of Sample a

训练次数	神经元输出				
	1	2	3	4	5
200	0.966 1	0.300 2	0.086 8	0.084 1	0.072 5
500	0.970 3	0.322 8	0.071 1	0.074 8	0.069 9
1 000	0.984 7	0.336 0	0.068 5	0.060 1	0.050 2

表 3 样例 b的网络输出

Tab. 3 The network output of Sample b

训练次数	神经元输出				
	1	2	3	4	5
200	0.160 1	0.019 9	0.511 0	0.089 9	0.040 3
500	0.168 7	0.018 1	0.633 3	0.073 3	0.038 8
1 000	0.172 4	0.017 7	0.711 0	0.061 1	0.029 6

表 4 样例 c的网络输出

Tab. 4 The network output of Sample c

训练次数	神经元输出				
	1	2	3	4	5
200	0.005 8	0.112 7	0.263 1	0.049 9	0.068 5
500	0.004 7	0.120 3	0.271 9	0.034 1	0.054 4
1 000	0.003 3	0.133 9	0.280 0	0.025 5	0.047 9

表 5 样例 d的网络输出

Tab. 5 The network output of Sample d

训练次数	神经元输出				
	1	2	3	4	5
200	0.049 6	0.010 9	0.338 9	0.069 0	0.047 6
500	0.034 5	0.006 5	0.347 0	0.056 6	0.039 9
1 000	0.023 9	0.003 1	0.351 1	0.048 3	0.020 2

表 6 样例 e的网络输出

Tab. 6 The network output of Sample e

训练次数	神经元输出				
	1	2	3	4	5
200	0.452 2	0.461 7	0.278 8	0.037 7	0.025 5
500	0.466 9	0.477 0	0.280 5	0.026 9	0.019 9
1 000	0.478 3	0.489 2	0.290 1	0.014 7	0.010 6

从以上各表中样例的输出值可以发现,切分处的值均大于 0.1,而非切分处的值则小于 0.1. 神经网络经过大量的学习训练,具备了一定的自学习性能,它通过 0、1之间的输出值数据来智能化地描述分词规则,若定性化地将切分处的标志设为 1,非切分处的标志设为 0,则可以将大于 0.1的数值定为“1”,小于 0.1的数值定为“0”.

从数据上可以看出,随着学习次数的增加,切分处的值也在随着不断增加,而非切分处的值则随着学习次数的增加而不断减少. 样本经过训练后,各指标如表 7所示. 取未参加训练的样本进行测试,结果如下: 测试样本数 5;待切分点总数 40;正确切分个数 37;错误切分个数 3;正确率 92.50% .

表 7 样本训练后各指标

Tab. 7 The indexes of trained samples

训练次数	样本个数	待切分点总数	正确切分个数	错误切分个数	正确率
200	20	160	126	34	78.75%
500	20	160	138	22	86.25%
1 000	20	160	156	4	93.13%

4 结　语

本文采用一种自然语言理解与神经网络相结合的分词算法,在分词过程中,首先进行单词代码库的设计,接下来进行输入向量的表示,然后输入网络进行训练,最后通过判断输出值是否落在规定区间内来判定切分点的切分状态. 通过选取具有代表性歧义结构的语句作为样例,对样本进行训练后,达到输出节点输出值与目标向量误差控制在较小范围内的目的. 该分词方法提供了一种新的输入、输出逻辑概念到输入、输出模式的转换方式,成功地解决了由于字间组合方式无穷多而无法训练的难题. 应用于歧义词语切分上,取得了很好的分词效果.

参考文献:

[1] 娄　[illegible],宋　柔,李　卫. 现代汉语分词系统通用接口设计与实现 [J].中文信息学报,2001, **15**(5): 1-7

[2] ZHANG Mao-yuan, LU Zheng-ding. A Chinese word segmentation based on language situation in processing ambiguous words [J]. **Inf Sci**, 2004, **162**(3-4): 275-285

[3] 林　珊. 中文分词在邮件过滤系统中的应用 [J]. 华南理工大学学报,2004, **32**(z1): 112-116

[4] 刘　群,张华平,俞鸿魁. 基于层叠隐马模型的汉语词法分析 [J]. 计算机研究与发展,2004, **41**(8): 1421-1429

[5] 孙茂松. 汉语自动分词研究评述 [J]. 当代语言学,2001, **3**(1): 22-32

[6] 马玉春,宋瀚涛. Web中文文本分词技术研究 [J]. 计算机应用,2004, **24**(4): 134-135

[7] HUANG Xiao-hong, LUO Zhen-sheng, TANG Jian. Quick method for Chinese word segmentation [C] // **Proceedings of the IEEE International Conference on Intelligent Processing Systems**. Piscataway: IEEE, 1998

[8] WANG Ye, HUANG Shang-teng. Chinese word segmentation based on apriori and adjacent characters [C] // **International Conference on Machine Learning and Cybernetics**. Piscataway: IEEE, 2005

Research on ambiguous words segmentation algorithm based on improved BP neural network

ZHANG Li*[1], ZHANG Li-yong[1], ZHANG Xiao-miao[1], GENG Tie-suo[2], YUE Zong-ge[3]

(1.School of Electr. and Inf. Eng., Dalian Univ. of Technol., Dalian 116024, China;
2.Nat. Assets Adm. Office, Dalian Univ. of Technol., Dalian 116024, China;
3.Hosp. of Dalian Univ. of Technol., Dalian 116024, China)

Abstract In the text mining, the technology of Chinese automatic word segmentation is a difficult problem that the computer science has to face. Aiming at the characteristics of Chinese writing, such as no space between words, continuous writing in sentences and difficulty of segmenting the ambiguous words, the grammatical phenomena are summarized which lie in the typical ambiguity, and the codes library of different parts of speech used for coding is built up. On this basis, words in ambiguity fields with special grammatical rules are set with codes and transformed to the representation form of inputting vector which can be accepted by the neural network. Then the samples are trained and the grammatical rules can be obtained by improving the self-learning of BP neural network. After a lot of training through adopting the BP network, the algorithm reaches 93.13% of training precision and 92.50% of test precision on ambiguous words segmentation.

Key words text mining; ambiguous words; natural language processing; neural network