

基于 HMM/BP 混合模型的文本信息抽取研究

杨红超, 肖基毅

(南华大学 计算机科学与技术学院, 湖南 衡阳 421001)

摘要:作为自然语言处理的一个分支,文本信息抽取成为了提取大量文本信息中有用信息的重要手段。介绍了目前在信息抽取领域中应用广泛的两种技术方法:HMM 和 BP 网络模型,分析了各自的优缺点,并在此基础上提出了一种基于两者的混合模型,该混合模型通过 BP 网络优秀的分类甄别能力来弥补 HMM 在分类方面的不足,而通过 HMM 强大的时域建模能力来弥补 BP 网络建模能力弱的问题,因此该模型具有强大的建模能力、分类性以及适应性强等特点。实验证明,相比传统的 HMM 以及 BP 网络模型,该混和模型在精确度和召回率上有了 10% ~ 15% 的提高。

关键词:信息抽取;隐马尔可夫模型;BP 网络

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2011)05-0115-03

Text Information Extraction Research Based on HMM and BP Network Hybrid Model

YANG Hong-chao, XIAO Ji-yi

(School of Computer and Technology, University of South China, Hengyang 421001, China)

Abstract:As a branch of natural language processing, the extraction of useful information in large text, the text information extraction became an important means. Introduce the information extraction widely used two kinds of technical methods: HMM and BP network model, analyze their advantages and disadvantages and on this basis propose a hybrid model, based on two models mentioned above. In this model, the classification by BP network capacity is to make up for deficiencies in the classification of HMM, HMM through strong time-domain modeling capabilities to make up for weak BP network modeling problem, so the hybrid model has strong modeling capabilities, classified and adaptability, etc. Experimental results show that compared to the traditional HMM and the BP network model, hybrid model in precision and recall rate is on the increase by 10% ~ 15%.

Key words:information extraction; HMM; BPN

0 引言

随着因特网的逐渐普及,网络上的文本资源正日益增加,而如何处理这些资源,使之变为有用的信息,成为目前的一个重要研究项目。目前在领域使用的主要技术中,传统的隐马尔可夫模型虽然具有考虑模式的时序性,强建模能力等优点^[1],但是其分类决策能力比较弱,需要大量先验知识等问题明显,而基本的 BP 神经网络虽然具有分类能力较强,网络的自适应性和较强的容错能力等特点,但是其需要特定的输入数据以及网络的时序性较弱等不利因素,使得两种模型在信息抽取上的应用无法达到最佳效果,文中在综合考虑了两种模型的优缺点以及二者在语音识别等领域

成功运用的基础上,提出一种结合 HMM 和 BP 网络优点的混合模型 HMMBP 网络。

文中利用该混合模型对英文论文头部进行抽取,结合论文头部的特点,首先将论文头部进行分块,然后对 HMM 进行训练,利用 Viterbi 算法计算最佳输出状态概率,并以此作为 BP 网络的输入,来进一步对相似状态进行分类,以达到最佳抽取效果。实验证明,相比传统的 HMM 以及 BP 网络模型,该混合模型在精确度和召回率上有了 10% ~ 15% 的提高。

1 HMM & BP

隐马尔可夫模型(HMM)是一个强大的统计学模型,在 20 世纪 80 年代,由 Baker 等人将其应用于语音信号处理领域,并成为主导技术。因其在该领域的成功应用和信息抽取领域的需求,90 年代末,HMM 技术被引入信息抽取领域,并逐渐得到推广。

作为在语音处理上的主流技术,HMM 其突出优点

收稿日期:2010-09-20;修回日期:2011-01-05

基金项目:湖南省科技计划项目(2008GK3090)

作者简介:杨红超(1985-),男,山东肥城人,硕士研究生,研究方向为智能信息系统与知识发现;肖基毅,教授,硕士生导师,研究方向为文本信息抽取。

是考虑了模型的时序性,对于动态时序序列有着很大的建模能力,通过这个优点,可以对语音信号的时序特点进行统计建模,进而抽取该信号的时序特点。但是对于易混淆语音的处理上,HMM 则无法更好的对其进行区分。而在信息抽取领域,HMM 的主要作用是利用 BW 算法或者 ML 算法来观察序列的统计分析,然后建立相应的统计模型,再使用 Viterbi 算法来计算某个可观察状态其最有可能的输出状态是哪一个。

通过上面的分析可以发现,HMM 的应用包括了两个方面,统计建模和分析最佳输出。这两个方面也同时说明了 HMM 的两个缺点,一是需要大量训练数据来进行统计,二是其识别只是通过其累计概率的最大值来决定相应的状态。

BP 网络是由信息的正向传播和误差的反向传播两个部分组成,可以有效地对隐含层的连接权值进行调整,因此一个优秀的 BP 网络可以充分逼近任意复杂的非线性关系,同时该网络模型保留了神经网络强大的鲁棒性、容错性、适应性等特点。在应用过程中,将待抽取信息的特征映射为相应的特征函数,然后利用各层之间的连接权值以及阈值和期望输出来对输出结果进行识别判断。BP 网络的主要缺点在于对时序的分析和映射能力不强,对动态时序的建模能力比较弱。

在语音识别领域,HMM 和 BP 网络已经得到了成功的应用:在文献[2]中则提出了用 BP 神经网络计算 HMM 的观察值矩阵的方法,解决了 HMM 适应性和鲁棒性差等问题;由 Lippmann 和 Gold 设计的用 NN 实现 HMM 中的 Viterbi 算法,更是提高了运算速度和混合网络的性能,Fulufhelo V. 等人在文献[3]中使用了高斯混合模型 GMM 和 HMM 对机器故障信息进行预测,Joachim Schenk and Gerhard Rigoll 在文献[4]中,通过一种串联的隐马尔可夫和神经网络的组合方式,即通过神经网络来进行特征抽取,然后应用到标准的 HMM 上的方法,来提高联机手写体研究领域中的特征识别。在文献[5]中,L. R. Rabiner 等人通过对单词识别中的每一个单词创建一个相应的 HMM 模板,这种多模板的抽取方式,不但减小了相似单词的抽取误差,更能提高单词的抽取效率,而该方法对于信息抽取领域也同样可以适用。

2 混合模型的提出

目前在信息抽取领域的一些研究成果:文献[6]通过使用“shrinkage”技术来改进 HMM 信息抽取模型概率的估计;文献[7]使用随机优化技术动态选择最适合的 HMM 模型结构进行信息抽取;在文献[4]中,作者提出了一种串联的方式,即通过 NN 来对特征提

取出来的部分应用到 HMM 上,以提高相似词的分类能力;文献[8]结合 HMM 模型和最大熵原理,使用 MEMM 模型(Maximum Entropy Markov Model)来实现信息抽取;文献[9]利用主动学习技术来减少训练 HMM 信息抽取模型时所需的标记数据。另外,在某些特定文本的信息抽取中,文献[10]将待抽取文本转换为相应的二进制数字,然后训练 BP 网络,以达到抽取目的。而对于 web 文本的抽取,文献[11]在分析了半结构化文档的特点后,利用 BP 网络来抽取规则,进而确定系统框架结构,实现信息的抽取。

在分析了前人方法的基础上,考虑到两种模型各自的优缺点,以及其相应的输入与输出要求,实验所使用的结合方法为:将 HMM 的最佳状态输出概率作为 BP 神经网络的输入。要抽取的文本如下所示:

A. Cau, R. Kuiper, and W. -P. de Roever.

Formalising Dijkstra's development strategy within Stark's formalism.

Proc. 5th. BCS-FACS Refinement Workshop,

1992.

对其四种状态 Author, Title, Public, Time 进行抽取,通过分析,首先将待抽取文本通过某些特殊字符(如回车,点号)进行分块,以最大限度将同一状态的单词分组在一起,然后对每一种状态设立一个 HMM,通过 BW 算法(未标记数据)或者 ML 算法(已标记数据)来对 HMM 的各个模型进行训练,在此基础上,再进行如下操作:

首先利用 HMM 中的 viterbi 算法计算出某个观察概率 $O = \{o_1, o_2, o_3, \dots, o_i\}$ 的最佳状态序列 $q^* = \{q_1^*, q_2^*, \dots, q_i^*\}$,然后再通过该最佳状态序列来计算各个模型的各个状态输出概率。部分观察序列 $O = \{o_{i1}, o_{i2}, \dots, o_{im(i)}\}$ 在状态上的输出概率记为 $x(i)$:

$$x(i) = P(q_{i,1}^* = q_{i,2}^* = \dots = q_{i,m(i)}^* = s_i, o_{i,1}, o_{i,2}, \dots, o_{i,m(i)} | \lambda)$$

如果词库中有 H 个 HMM 模型(在该文中有 4 个),每个模型有 N 个状态,所以得到的输出概率序列为 $X = \{x_{1,1}, x_{2,2} \dots x_{1,N}, x_{2,1}, x_{2,2} \dots x_{2,N}, \dots, x_{H,1}, x_{H,2}, \dots, x_{H,N}\}$ 。

而该混合模型的抽取流程图见图 1。

3 实验与分析

实验采用美国 CMU 大学 CORA 提供的数据集进行测试,该数据集共有 935 篇计算机科研论文头部数据,采用其中 500 篇作为训练数据,100 篇作为测试数据,来进行实验,将结果同单独的 HMM 和 BP 网络的结果进行对比,并采用精确度和召回率来评测算法的性能:

精确度 = (模型正确标记的单词数/所有测试的单

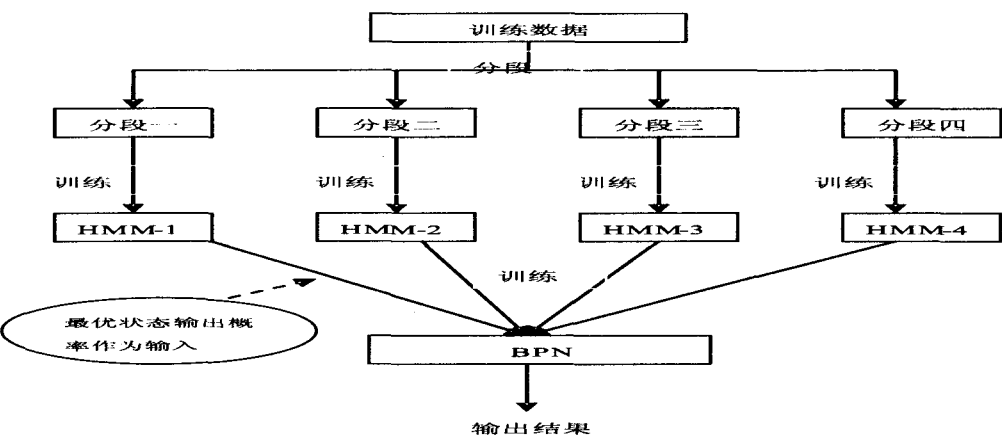


图 1 分块 HMMBP 混合模型

词数) * 100% ;

召回率 = (模型正确标记为该状态的单词数 / 手工标记该状态的单词数) * 100%

实验结果见表 1。

表 1 基于 HMM,BP,分块 HMMBP 三种模型的信息抽取比较

	HMM		BP		分块 HMMBP	
	准确率	召回率	准确率	召回率	准确率	召回率
Author	0.892	0.873	0.845	0.823	0.945	0.944
Title	0.625	0.633	0.779	0.801	0.865	0.839
Public	0.635	0.645	0.700	0.859	0.814	0.847
Time	0.827	0.859	0.934	0.893	0.942	0.951

通过上表可以看出,基于分块的 HMMBP 网络模型,比起传统的 HMM 和 BP 网络在精确度和召回率上有了-定的提高,对于具有明显抽取特征的状态,如 Author,Time,各个抽取模型均可以保持很高的抽取效率,特别是混合模型,在准确率上更接近 1,而对于具有一定混淆单词的 Title 和 Pubile 则在传统 HMM 和 BP 网络中无法更好地解决这个问题,因为在这两种状态中,具有大量的相近或者相同的输出状态,从而使得其准确率与召回率无法得到很大提高,从抽取结果中也可以发现,即使使用分块的 HMMBP 网络,其准确率和召回率也只能得到大于 15% 的提高,而无法达到最理想效果。如何进一步提高相似状态的抽取效率,将成为下一步的研究目标。

4 结束语

随着网络的发展,信息的不断增加,以及某些特殊领域的特殊需求的提高,文本信息抽取正在发挥着越来越重要的作用,文中通过对文本的分块并结合 HMM 和 BP 网络的优缺点,所提出的混合模型,明显地提高了文本抽取的效率,虽然在某些方面还不是很理想,但这正成为下一步的工作,改进混合模型,进一步提高信

息抽取的精确率和召回率。

参考文献:

[1] Leek T R. Information Extraction Using Hidden Markov Models [D]. San Diego: [s. n.], 1997.

[2] LI Weiying, Yi Kechu, Hu Zheng. Introducing neural predictor to hidden Markov model for speech recognition [C] // ICSLP. Canada: [s. n.], 1992.

[3] Nelwamondo F V, Marwala T, Mahola U. Early Classifications of Bearing Faults Using Hidden Markov Models, Gaussian Mixture Models [J]. Mel-Frequency Cepstral Coefficients and Fractals International Journal of Innovative Computing, Information and Control, 2006, 2 (6) : 1281-1299.

[4] Schenk J, Rigoll G. Novel Hybrid NN/HMM Modelling Techniques for On-line Handwriting Recognition [D]. München: Institute for Human-Machine Communication Technische University München, 2002.

[5] Rabiner L R, Lee C H, Juang B H. HMM Clustering for Connected Word Recognition [C] // Proc. of IEEE ICASSP. [s. l.] : [s. n.], 1989: 405-408.

[6] Freitag D, McCallum A K. Information extraction with HMM and Shrinkage [R]. [s. l.] : [s. n.], 1999.

[7] Freitag D, McCallum A. Information extraction with HMM structures learned by stochastic optimization [C] // Proceedings of the Eighteenth Conference on Artificial Intelligence. [s. l.] : [s. n.], 2000: 584-589.

[8] Freitag D, McCallum A, Pereira F. Maximum Entropy Markov Models for Information Extraction and Segmentation [C] // 7th International Conf. on Machine Learning. [s. l.] : [s. n.], 2000: 591-598.

[9] Scheffer T, Decomain C, Wrobel S. Active Hidden Markov Model for Information Extraction [C] // In Proceedings of the International Symposium on Intelligent Data Analysis. [s. l.] : [s. n.], 2001: 309-318.

[10] 李 帅,黄玺瑛,董家瑞.一种基于神经网络的特定文本信息提取方法 [C] // 第十届中国科协年会论文集 (1). 出版地不详:出版者不详, 2008: 420-424.

[11] 明廷波,左志宏,史永刚,等. Web 信息抽取中基于神经网络的规则学习方法 [J]. 南京大学学报 (自然科学版), 2005 (Z1) : 1-6.