第 42 卷 第 6A 期
2015 年 6 月

计 算 机 科 学
Computer Science

Vol. 42 No. 6A
June 2015

一种基于动态协同神经网络的文本作者身份分类研究

张爱华

(军事经济学院 武汉 430035)

摘 要 为了提高文本作者身份分类的准确度,提出一种动态协同神经网络算法。该算法利用了协同神经网络训练速度快、抗造声强等特点,并采取了注意参数动态调整的策略。通过原型模式向量与实验模式向量间的相似性动态地选取合适的注意参数,在演化过程中对误识别的模式进行自适应纠正。与平衡注意参数条件下的识别效果进行对比校验,结果表明,该算法在很大程度上提高了网络的自学习能力,从而改善了作者身份分类的精度和鲁棒性。

关键词 动态协同神经网络,作者身份分类,注意参数,序参量

中图法分类号 TP391.1 文献标识码 A

Dynamic Synergetic Neural Network Algorithm for Authorship Classification of Texts

ZHANG Ai-hua

(Military Economy Academy, Wuhan 430035, China)

Abstract To improve the accuracy of authorship classification of texts, a dynamic synergetic neural network algorithm was proposed in this study. Specifically, this algorithm makes use of the characteristics of fast training and strong noise-immunity, and dynamically adjusts the attention parameters. Thus, the initial mis-identified patterns are adaptively corrected by measuring similarity between the prototype pattern and the testing pattern in evolution process. Compared with classification result under balanced attention parameter, the experimental result demonstrates that self-learning ability of the network is significantly improved in the dynamic synergetic neural network algorithm, thus the classification performance and robustness are improved.

Keywords Synergetic neural network, Authorship classification, Attention parameter, Order parameter

1 概述

在线信息通常以网络评论、聊天记录、电子邮件等形式存在,使用较多非正式的语言和符号,信息内容很丰富。而网络中经常产生一些不良行为,如利用发电子邮件或在论坛发帖等方式散布色情、虚假、诽谤等不良信息。通常,这些信息大多具有匿名的特点,这为网络信息的安全监管提出了较大的挑战。目前,追踪匿名文本信息的主体身份已成为数据挖掘和信息安全领域的研究热点[1],该研究的关键在于准确地辨识匿名文本的作者身份[2]。

作者身份分类研究的关键技术在于特征集的选取和分类模型的构建。目前,特征集的选择主要集中于电子文档的写作风格抽取上,如 N 元单词或字符组合[3],句法、标点等统计特征[4];而分类模型的研究主要集中于单分类器[5](如神经网络、决策树、贝叶斯等算法)的改进和集成分类算法[6,7]的研究上。

本研究提出了一种基于动态协同神经网络的文本作者身份分类方法。在预处理中,我们组合了大量的写作风格特征来构成全局特征。在识别过程中,采取了一种动态调整注意参数的网络优化策略,对训练误差进行自适应校正。实验证明:优化后的网络结构具有更强的自学习能力和更好的识别效果。

张爱华(1979—),硕士生,讲师,主要研究方向为数字模型与计算。

2 协同神经网络

80 年代末期,Haken[8]提出了协同学原理运用于模式识别的新概念,将模式识别的过程对应于一个动力学的过程[9]。设想一个虚拟粒子在有势地形图上移动,当粒子进入某个吸引谷底时,与之相应的模式就被识别出来。模式识别可被认为是一个协同系统,假设 q 为协同演化系统的参量,设系统具有 M 个分量:

$$q=(q_1,q_2,\cdots,q_M) \tag{1}$$

对待识别模式 q 可以构造一个动力学过程:

$$q=\sum_{k=1}^{M}\lambda_k(v_k^+q)v_k-B\sum_{k\neq k'}v_k(v_{k'}^+q)^2(v_k^+q)-Cq(q^+q)+F(t) \tag{2}$$

使 q 经过中间状态 $q(t)$ 进入到一个原型模式 v_k,即该模式与 $q(0)$ 最为接近,这个过程可描述为:

$$q(0)\to q(t)\to v_k \tag{3}$$

如果将向量 q 在原型向量上分解,有

$$q=\sum_{k=1}^{M}\xi_k v_k+w,\ v_k\cdot w=0 \tag{4}$$

其中,ξ_k 为序参量,w 为剩余向量。序参量的引入可以使网络行为得到简化:

$$\xi_k=v_k^+q \tag{5}$$

v_k^+ 为 v_k 的伴随向量,它满足正交关系式:

$$(v_k^+, v_j)=\delta_{kj}, \delta_{kj}=\begin{cases}1, & k=j\\ 0, & k\neq j\end{cases} \tag{6}$$

序参量的时间演化方程为：

$$\dot{\xi}_k=\xi_k(\lambda_k-D+B\xi_k^2) \tag{7}$$

$$D=(B+C)\sum_{k'}\xi_{k'}^2 \tag{8}$$

输入层对应状态向量各分量的初始值，共 N 个神经元，序参量层有 M 个神经元，每个神经元的初始状态由式(5)从输入层获得，连接权值由 v_k^+ 决定。D 使各单元相互作用和竞争，当 $t\to\infty$ 时，输出该层表现为最终识别出的模式。

3 基于动态协同神经网络的作者身份分类方法

3.1 文本作者身份分类框架

根据协同神经网络的基本原理，本节提出一种基于动态协同神经网络的文本作者身份分类的方法(见图 1)。该方法首先对得出的初始序参量进行演化，演化过程使用自适应的动态参数调整法，即本文所提出的改进方法；然后，在分类过程中，我们采用 SCAP 算法[10]求解原型向量 $v_k(k=1,2,\cdots,M)$，即求出每个作者类别的特征向量的平均值，来用伪逆法求出伴随向量。分类过程需经过训练和分类两阶段，步骤如下：

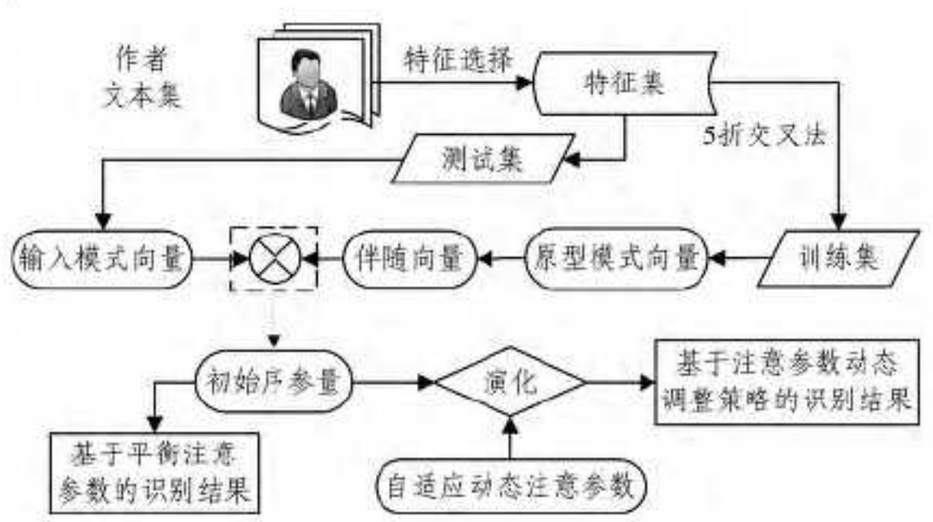

图 1 基于协同神经网络的网文本作者分类模型

(1)训练阶段

a)读入训练样本的全局特征向量，计算出满足归一化和零均值的原型模式向量 v_k；

b)求出原型模式向量 v_k 的伴随向量 v_k^+，并存储伴随向量矩阵。

(2)分类阶段

a)网络输入层读入待识别模式的特征向量 $q(0)$，使其满足归一化和零均值；

b)输入层模式特征向量 $q(0)$ 与网络权值相乘，即 $\xi_k(0)=v_k^+\cdot q(0)$，求出网络中间层序参量 ξ_k 的初始值；

c)根据式(9)描述动力学方程执行序参量 ξ_k 的演化：

$$\begin{aligned}&\xi_k(n+1)-\xi_k(n)=\gamma(\lambda_k-D+B\xi_k^2(n))\xi_k(n)\\&D=(B+C)\sum_{k'}\xi_{k'}^2(n)\end{aligned} \tag{9}$$

γ 为迭代步长，它决定了协同神经网络的稳定性。

d)判断序参量 $\xi_k(n)$ 演化的稳定性，若演化过程已稳定，则判断最终序参量模值为 1 的类别为测试样本的所属作者类别；否则转 c)继续调整。

3.2 动态协同神经网络

在式(8)中，λ_k 称为注意参数，它控制了模式演化的最终结果和速度。由初始的关于序参量 ξ_k 的动力学方程：

$$\dot{\xi}_k=\lambda_k\xi_k+\tilde{N}_k(\xi_j) \tag{10}$$

忽略高阶项和涨落力可得

$$\xi_k\approx\xi_k(0)e^{\lambda_k t} \tag{11}$$

由上式可以看出注意参数 λ_k 控制模式变化的速度，如果对某模式加以“注意”，则可对它赋以较大的 λ_k，这样即使它所对应的序参量没有最大初始模值，依然有可能获得竞争的胜利，这也符合生物感知的特点。它的取值分为平衡注意参数和不平衡注意参数。

由于在平衡注意参数的情形下，协同神经网络的最后输出模式可以通过直接比较个体原型模式序参量初始模值大小得到，最大序参量初始模值所对应的原型模式最终将获胜，这样无需经过序参量的演化过程[8]。该情况下，协同神经网络无法通过网络的学习来正确判断，即网络失去了自学习能力。

而非平衡注意参数更有利于选择性的模式识别，特别是序参量之间相差不大时，注意参数在决定序参量的动力学演化中占了很大的比重。本文根据身份识别的实际应用提出一种注意参数动态调整的方法，根据原型模式向量与待测模式向量间的相似度选取合适的不平衡注意参数：

$$\lambda_k=\frac{\sum_{i=1}^{N}(v_{ki}-\bar{v}_k)(q_i-\bar{q}_i)}{\sqrt{\sum_{i=1}^{N}(v_{ki}-\bar{v}_k)^2\sum_{i=1}^{N}(q_i-\bar{q}_i)^2}} \tag{12}$$

我们选取待识别模式在原型模式之间的相关系数来作为不平衡参数的取值，它反映了原型模式向量与实验模式向量之间的相似处，原型模式越接近实验模式，则相应的注意参数越大，这是一种根据输入模式的特征自适应地确定主意参数的方法。

对于迭代速度，采取自适应动态的方式来调整，在实验中我们选取 $\gamma=0.5/D$，而常数 B 和 C 根据最佳经验值设为 $B=C=1$，序参量的演化式变为：

$$\begin{aligned}&\xi_k(n+1)-\xi(n)=\frac{0.5}{D(n)}(\lambda_k-D(n)+\xi^2(n))\xi(n)\\&D(n)=2\sum_{k'}\xi_{k'}^2(n)\end{aligned} \tag{13}$$

4 实验

4.1 数据集

实验数据集取自于自亚马逊购物网站上的用户评论[11]，包括 50 个作者，每个作者 50 篇评论。在预处理过程中我们提取样本中丰富的语言风格特征，如字符和单词的 n-grams 组合、单词长度、句子长度以及特殊符号出现次数等。

4.2 实验设置

将所有特征进行组合以构成全局特征，这里的全局特征是指提取的特征与所有样本群相关的特征集。由于初始全局特征向量维数较高，超过 10000 维。我们采取 χ^2 统计法来进行特征选择，将初始特征维数降为 2000。

另外，在测试阶段时，我们采取 5 折交叉法，这有利于算法鲁棒性的检测，在协同模式识别模块中，分别采用平衡注意参数(Balanced Attention Parameter，BAP)和自适应动态注意参数(Adaptive Attention Parameter，AAP)进行对比实验，并加入支持向量机(Support Vector Machine，SVM)、BP 神经网

络(BP)、径向基神经网络(RBF)以及朴素贝叶斯(Naïve Bayes,NB)算法进行比较。BAP情况下设置$\lambda_k=B=C=1$,AAP情况下,迭代次数为50次。

4.3 实验结果及分析

图2描述了在使用10个作者训练集时,BAP和AAP情况下协同神经网络对某测试样本的识别过程。基于BAP协同神经网络的最终识别结果为具有最大初始模值的序参量对应的模式。由于试验模式向量初始序参量中$\xi(10)$模值最大,大于$\xi(1)$的初值,从而使得测试样本被识别为第10个作者的文章,但该样本实际归属于第1个作者,因此网络识别失败;而采用注意参数动态调整时,注意参数被自适应改变,使模式类1在竞争中获胜,从而网络识别成功。此外,从演化过程可观察到,基于AAP的协同神经网络收敛速度更快,节省了序参量演化的时间。

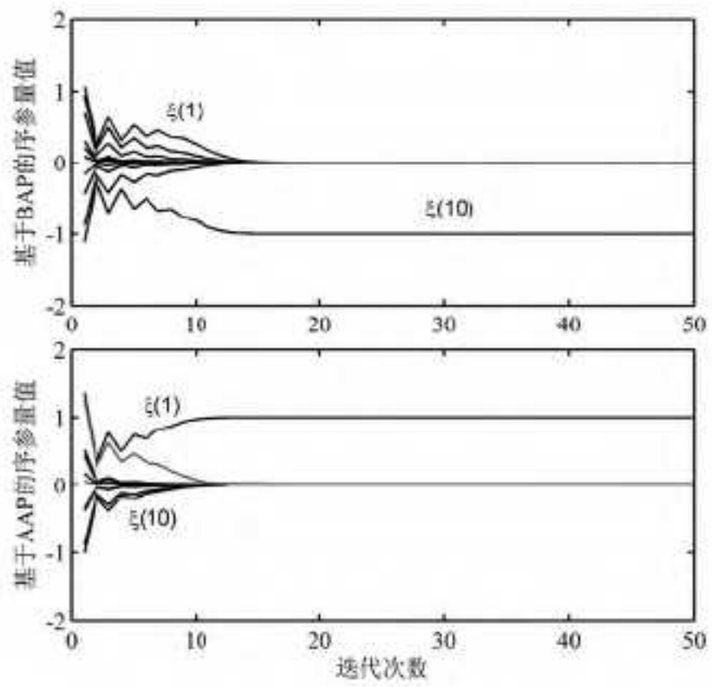

图2 在BAP与AAP调整策略下序参量演化效果图

为了验证改进的协同神经网络分类的效果,我们使用同一种数据集测试6种识别方法:BAP、AAP下的协同识别算法、SVM、BP、RBF以及NB分类算法。并且引入不同作者数目(50,40,30,20,10,5),每个作者包含50篇评论的数据集以验证6种算法的鲁棒性和识别精度。识别结果如表1所列。

表1 不同算法的分类准确率比较结果(%)

算法	作者数量					
	50	40	30	20	10	5
AAP	80.4	82.5	84.2	87.3	89.3	95.7
BAP	68.3	70.2	74.1	80.4	87.5	92.8
SVM	78.6	80.2	86.1	88.2	92.3	96.1
BP	77.8	78.6	84.5	87.5	90.8	95.2
RBF	77.2	77.9	82.6	86.8	86.7	93.6
NB	76.6	78.2	83.4	87.3	90.2	94.4

从实验结果可以观察,基于AAP协同神经网络由于具有自学习能力,能对初始状态中的错误识别进行纠正,在识别效果和鲁棒性方面已完全超过BAP条件下的识别方法;BP神经网络在作者数较少情况下表现较为稳定,但当作者数超过40时,识别精度均低于80%;而作为文本分类中常用的SVM与NB分类器在对5至30个作者时分类效果达到最佳,但在区分40和50个作者时效果比改进后的协同神经网络差,该现象表明协同神经网络与原型模式向量的选取方法有着重要的关系,伪逆法在区分类间相关性方面有很好的效果,不同作者类别的特征向量间相似度得到减小,在50个作者的识别中也能达到80%的精度。结果证实了改进后协同神经网络的分类效果更好。

结束语 本文提出了一种动态协同神经网络算法,该算法引入了一种注意参数动态调整策略,并将其应用于文本作者身份分类中。分类中,模型根据原型模式与待测样本模式的相关性来动态改变注意参数的取值,使得初始状态下被误识别的模式能在演化过程中得到纠正,最后被正确识别,网络具备了更好的自适应能力。在实验中,与传统分类算法的对比结果表明,基于动态协同神经网络的分类模型具有更好的识别准确性和鲁棒性。由于λ_k、B和C的取值对识别结果有着重要的影响,下一步将继续深入研究协同神经网络中参数的优化方法,以提高动态协同神经网络在大规模作者身份分类中的鲁棒性。

参 考 文 献

[1] Narayanan A, Paskov H S, Gong N Z, et al., On the feasibility of internet-scale author identification[C]// Proc. of the IEEE Symposium on Security & Privacy (IEEE S&P), California, USA, 2012: 300-314

[2] Iqbal F, Binsalleeh H, Fung B C M, et al. A Unified Data Mining Solution for Authorship Analysis in Anonymous Textua Communications[J]. Information Sciences, 2011, 231: 98-112

[3] Houvardas J, Stamatatos E. N-gram Feature Selection for Authorship Identification[C]// Proc. of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications. [S. l.]: Springer, 2006

[4] Burrows J F. Word patterns and story shapes: The statistical analysis of narrative style[J]. Literary and Linguistic Computing, 1992, 2: 6-67

[5] Li Jie-xun, Zheng Rong, Chen Hsin-Chun. From Fingerprint to Writeprint[J]. Communications of the ACM, 2006, 49(4): 76-82

[6] 孙建文,刘三妍,杨宗凯,等. 采用集成特征选择的网络书写纹识别研究[J]. 小型微型计算机系统, 2012, 33(5): 1108-1112

[7] 刘三妍,铁璐,刘智,等. 基于多元概率推理模型的中文书写纹识别[J]. 计算机工程, 2013, 39(11): 158-162

[8] Haken H. Synergetics[M]. Springer-Verlag Press, 1977

[9] Liu M, Fu Y. On Synergetic Theory and Application in Sports Curriculum Reform[C]// Proc. of the 2009 First International Workshop on Education Technology and Computer Science. 2009: 557-562

[10] Wagner T, Boebel F G. Testing Synergetic Algorithms with Industrial Classification Problems[J]. Neural Networks. 1994, 7(8): 1313-1321

[11] http://www.amazon.com, 2014