文章编号:1008 - 1402(2018)03 - 0354 - 04

基于卷积神经网络的文本分类算法①

王美荣

(安徽新华学院信息工程学院,安徽 合肥 230088)

摘 要: 为了解决分类算法在文本分类时出现特征维度过高和数据稀疏的问题,提出了一种基于卷积神经网络(convolutional neural network, CNN)的文本分类算法,该算法结合卷积神经网络论中的邻接矩阵对文本分类进行动态建模。对文本的词向量进行训练,并且通过分类邻接矩阵得到群的结构和个数分类。在提取出文本抽象特征的基础上用 CNN 分类器来进行分类。仿真分析表明:该算法在在进行文本分类效果显著。

关键词: 文本分类;卷积神经网络;动态建模;词向量

中图分类号: TP18 **文献标识码:** A

0 引 言

随着网络的飞速发展,社交网络逐渐成熟[1]。国外出现了 Facebook, Twitter 等社交平台,而国内则是新浪微博、博客等社交平台。越来越多的人喜欢在这些平台上发表自己的言论或者是点评被人所发表的[2]。这些评论通常都是比较简短的,字数受到一定的限制而且用语也有些不规范。内容涉及十分的广阔,比如教育、政治、经济、文化、医疗卫生等等[3]。这些文本里面包含了许多有用的信息,但是因为网络的更新速度太快,这些文本内容又没文本分类由多个相互协作的文本构成[4]。传统文本分类算法一般假设文本种类固定且不受分类干扰,基于 RFS 的分类算法可以有效解决上述问题并且还能够避免数据关联过程[5]。从这个角度来讲,RFS 更适合于解决文本的分类问题。为了得到各文本的轨迹分类本文在标签 RFS 框架下,采用 CNN 算法。在目前已有的文本分类算法中,都没有描述到文本的结构信息[6]。

采用了卷积神经网络理论对文本进行分类计算。首先,通过借助卷积神经网络理论对文本进行动态建模。在此基础上,再针对文本进行分类分类。因为不知道最初始的文本的协作关系,所以可以先假设文本之间没有关系,是独立的。采用 CNN 获得各文本的词向量分类和轨迹分类以及文本的个数分类。在获得文本中各成员的词向量分类基础上,通过计算每时刻的偏差矩阵分类获得邻接矩阵分类。

1 卷积神经网络

开始先定义卷积神经网络,如下内容:

定义 V_e 和 E_d 这两个集合组成了卷积神经网络,将这个记为 $G = (V_e, E_d)$。其中 E_d 表示边的有限集合,V_e 则表示节点的非空有限集合,当这些边有方向时则称为有向卷积神经网络,反之,称为无向卷积神经网络。

借助卷积神经网络结构和群结构的相似性。使用的邻接矩阵不需要知道文本之间的距离,这样可以大大的降低难度,便于得到一个文本邻接的矩阵,如下所示:

$$\boldsymbol{A}_{\mathrm{d}} = \begin{bmatrix} 0 & a(1,2) & \cdots & a(1,n) \\ a(2,1) & 0 & \cdots & a(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ a(n,1) & a(n,2) & \cdots & 0 \end{bmatrix} \quad (1)$$

其中,当第 i 个文本是第 j 个文本父节点时则 $a(i,j)$ 等于 1;其他情况下,$a(i,j)$ 等于 0。

若文本存在单个父节点时,该文本分类模型如下:

① 收稿日期:2018 - 04 - 05
基金项目:2016 年安徽省高等学校省级质量工程项目(2016mooc198);2018 年安徽高校自然科学研究项目重点项目(KJ2018A0587)。
作者简介:王美荣(1978 -),女,皖蚌埠人,讲师,硕士,研究方向:数据挖掘及软件测试. 。

$$x_{k+1,i} = F_{k,l}x_{k,l} + b_k(l,i) + B_{k,i}w_{k,i} \quad (2)$$

$$z_{k+1,i} = C_{k+1}x_{k+1,i} + v_{k+1,i} \quad (3)$$

其中，$x_{k,i} = [p_{k,x_{k,i}}, \dot{p}_{k,x_{k,i}}, p_{k,y_{k,i}}, \dot{p}_{k,y_{k,i}}]^T$，$p_{k,x_{k,i}}$，$p_{k,y_{k,i}}$ 和 $\dot{p}_{k,x_{k,i}}$，$\dot{p}_{k,y_{k,i}}$ 分别表示文本 i 在 x 和 y 轴上的位置和速度; $x_{k,i} \in X_k$. l 表示 i 文本的父节点. $b_k(l,i)$ 为一个补偿向量，表示文本 i 与其父节点之间的位置关系; F 和 C 分别表示词向量转移矩阵和观测矩阵; B 为词向量噪声系数矩阵; w 和 v 分别为卷积神经网络噪声和观测噪声且都服从正态分布。

通过研究邻接矩阵就可以轻松的判断出群中各个文本之间的关系和连接，比如父子关系。若没有父亲节点，那么就将这个文本称呼为头节点。头节点的分类会影响到其文本，而头节点自身分类不受其他文本影响。因此，头节点分类模型中补偿向量 b 为 0，并且 $x_{k,l}$ 为它自身在 k 时刻的词向量。否则，该文本存在着父节点并且该文本的分类受其父节点影响，所以通过这个文本分类模型，我们可以发现补偿向量 b 包含该节点与其父亲节点之间的方向和距离信息，当文本存在多个父节点时，线性条件下 $x_{k+1,i}$ 如下式表示:

$$x_{k+1,i} = \sum_{l \in P(i)} w_k(l,i)[F_{k,l}x_{k,l} + b_k(l,i)] + B_{k,i}w_{k,i} \quad (4)$$

$$x_{k,i} \in X_k, \sum_{l \in P(i)} w_k(l,i) = 1, w_k(l,i) \in [0,1) \quad (5)$$

其中，$P(i)$ 等价于 $\{P_1(i), P_2(i), \cdots, P_{jn}(i)\}$，$P_j(i)$ 表示 i 文本的第 j 个父节点。

根据该卷积神经网络算法获得文本的邻接矩阵，文本 1 是头节点，文本 2 和文本 3 是文本 1 的子节点，因此该群的分类模型如下:

$$\begin{cases} x_{k+1,1} = F_{k,1}x_{k,1} + B_{k,1}w_{k,1} \\ x_{k+1,2} = F_{k,2}x_{k,1} + b_k(1,2) + B_{k,2}w_{k,2} \\ x_{k+1,3} = F_{k,3}x_{k,1} + b_k(1,3) + B_{k,3}w_{k,3} \end{cases} \quad (6)$$

表 1　文本分类模型建立步骤

建立篮球迷分类模型的主要步骤
1. 遍历群中所有 1: n 个节点
2. 利用邻接矩阵找到该节点的父节点
3. 若在该节点上存在父节点
$x_k = \frac{1}{jn}\sum_{l \in p(i)}[x_{k,l} + b_{k-1}(l,i)] + B_{k-1}w_{k-1}$
4. 如果该节点不存在父节点
$x_k = F_{k-1}x_{k-1} + B_{k-1}w_{k-1}$

表 1 描述了在本文中如何建立一个群的分类模型的主要步骤，简单起见，假设权重 $w_{k-1}(l,i)$ 为等权重。

2　文本分类算法

为了获得各文本的轨迹分类，选择了 CNN 对文本进行分类分类。文本之间的词向量是有关系，非独立的。但是因为不知道起始阶段的文本之间的协作关系，目前可当作群结构和词向量之间是耦合在一起的，所以用一种两阶段的算法。在第一阶段，首先把文本看作为独立分类。

使用 CNN 分类文本的词向量和个数。其中，标准 *GLMB* 的算法定义如下:

$$\pi(X) = \triangle(X)\sum_{c \in C} w^{(c)}(L(X))[p^{(c)}]^X \quad (7)$$

其中，C 表示离散变量; $p^{(c)}(\cdot,l)$ 表示概率密度; $w^{(c)}(I)$ 为权重并且满足 $\sum_{(I,c) \in F(L) \times C} w^{(c)}(I) = 1$; $F(L)$ 为 L 上所有有限子集的集合。该标准 CNN 在贝叶斯递推下封闭。

为了便于计算，将上述表达式变形为如下所示的表达式，称作为 δ - *GLMB*:

$$\pi(X) = \triangle(X)\sum_{(I,\xi) \in F(L) \times \Xi} w^{(I,\xi)}\delta_I(L(X))[p^{(\xi)}]^X \quad (8)$$

比如，在 k 时刻，让 Ξ 为空集，假设有两种可能，如下表示:

1) 有 0.2 的概率存在 1 个文本，标签为(0，2)，即在 k 时刻存在文本(0，2)(即 0 时刻产生的文本 2)，并且该文本的概率密度为 $p(\cdot,(1,1)) = N(\cdot,m,P_2)$。

2) 有 0.8 的概率存在 2 个文本，标签分别为(1，1)和(0，2)(即 1 时刻产生的文本 1，0 时刻产生的文本 2)，概率密度分别为 $p(\cdot,(1,1)) = N(\cdot,0,P_1)$ 和 $p(\cdot,(0,2)) = N(\cdot,m,P_2)$。则，0 时刻的 δ - *GLMB* 表达式如下所示

$$\pi_0(X) = 0.2\delta_{\{(0,2)\}}L(X)p_0^X + 0.8\delta_{\{(1,1),(0,2)\}}L(X)p_0^X \quad (9)$$

预测步: 当多文本的先验概率密度形式如(8)式所示时，δ - *GLMB* 的预测步如下所示

$$\pi_+(X_+) = \Delta(X_+)\sum_{(I,\xi) \in F(L) \times \Xi} w_+^{(I,\xi)} \times \delta_{I_+}(L(X_+))[p_+^{(\xi)}]^{X_+} \quad (10)$$

其中

$$w_+^{(I,\xi)} = w_B(I_+ \cap B)w_S^{(\xi)}(I_+ \cap L) \quad (11)$$

$$p_+^{(\xi)}(x,\ell) = 1_L(\ell)p_S^{(\xi)}(x,\ell) + (1-1_L(\ell))p_B(x,\ell) \quad (12)$$

$$p_S^{(\xi)}(x,\ell) = \frac{(p_S(\cdot,\ell)f(x|\cdot,\ell),p^{(\xi)}(\cdot,\ell))}{\eta_S^{(\xi)}(\ell)} \quad (13)$$

$$\eta_S^{(\xi)}(\ell) = \int (p_S(\cdot,\ell)f(x|\cdot,\ell)),p^{(\xi)}(\cdot,\ell))\mathrm{d}x \quad (14)$$

$$w_S^{(\xi)}(L) = [\eta_S^{(\xi)}]^L \sum_{I\subseteq L} 1_I(L)[q_S^{(\xi)}]^{I-L} w^{(I,\xi)} \quad (15)$$

$$q_S^{(\xi)}(\ell) = (q_S(x|\cdot,\ell)),p^{(\xi)}(\cdot,\ell)) \quad (16)$$

其中，$w_B(I_+\cap B)$ 是新加入标签 $(I_+\cap B)$ 的权重；$w_S^{\xi}(I_+\cap L)$ 是保持标签 $(I_+\cap L)$ 的权重；$p_B(\cdot,\ell)$ 是新加入文本的概率密度；$p_S^{\xi}(x,\ell)$ 是由先验密度 $p^{(\xi)}(\cdot,\ell)$ 得到的保持文本的密度；$f(\cdot|\cdot,\ell)$ 表达了关于文本的概率密度。

更新步：若发现多个本文的预测密度非之前所预测，并且如(8)所示的那样，那么更新步如下

$$\pi(X|Z)\approx \triangle(X)\sum_{(I,\xi)\in F(L)\times\Xi}\sum_{\theta\in\Theta^{(M)}}\tilde{w}^{(I,\xi,\theta)}\times \delta_I(L(X))[p^{(\xi,\theta)}]^X \quad (17)$$

其中，在一个固定的 (I,ξ) 中，$\Theta^{(M)}=\{\xi^{(1)},\cdots,\xi^{(M)}\}$ 集合表示为在最大权重 $w^{(I,\xi,\theta^{(i)})}$ 时的 Θ 的 M 个元素。$\tilde{w}^{(I,\xi,\theta)}$ 为截断后的归一化权重。在获得文本词向量分类基础上，进一步分类群结构，获取群协作分类关系。

3 仿真实现

考虑线性和非卷积神经网络两个实验来验证文中所给算法。在实验中使用 CNN 和 CBMeMBer 进行比较。为评估文中所给算法的性能，采用最优子模型分配距离(Optimal sub pattern assignment，OSPA)：

$$\bar{d}_p^{(c)}(X,\hat{X}) = \left\{\frac{1}{n}\left[\min_{\pi\in\prod_\pi}\sum_{i=1}^{m} d^{(c)}(x_i,\hat{x}_{\pi(i)})^p + c^p(n-m)\right]\right\}^{\frac{1}{p}} \quad (18)$$

其中，X 和 $\hat{X}$ 分别为真实词向量集和分类词向量集，种类分别为 m 和 n，且 $m\leqslant n$，$1<p<\infty$，$d^{(c)}(x,\hat{x})=\min\{c,d(x,\hat{x})\}$，$c>0$，$\prod_k$ 表示 $1,2,\cdots,k$ 所有各种排列组成的集合。

从图1和图2中的 OSPA Card 可知，当真实文本种类发生变化时，CNN 对文本的个数分类出现了一个延迟过程。例如：在第 15s，文本种类发生变化，CNN 经历 6s 后跟上个数变化，而 CBMeMBer 滤波算法只需经历 1s 后并能跟上个数变化，然而在这过程后，CNN 并能够较稳定的分类出文本种类而 CBMeMBer 滤波算法在文本种类的分类过程中出现较多的波动。

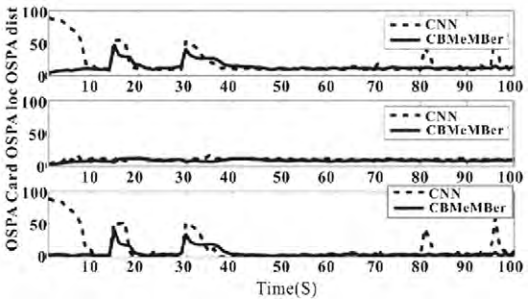

图1 OSPA 距离对比卷积神经网络(经 50 次 MC 平均)

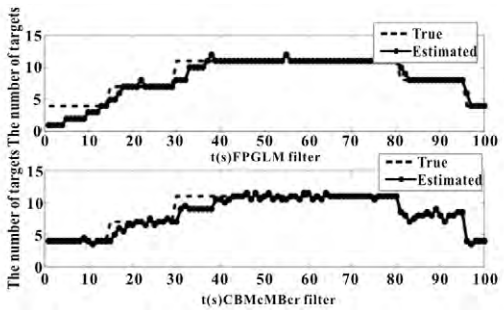

图2 文本种类分类

所使用的分析方法，是采用平均每步所消耗的 CPU 时间对 CNN 和 CBMeMBer 滤波算法。利用上面的方法进行仿真，平均每步所消耗的 CPU 时间实验结果如下表格所示，利用这个表格测试算法的 PC 机的 CPU 为 Intel(R) Core(TM) i5 - 4460M3.20GHz，RAM 为 4GB，32 位 Win7 卷积神经网络。

表2 算法性能分析

算法	CNN 算法		CBMeMBer 算法	
	线性	非线性	线性	非线性
时间(秒/步)	1.35	2	0.044	0.52

从表2中，可以发现 CBMeM - Ber 算法在进行本文分类下所消耗的时间都比 CNN 算法下消耗的时间要大，与此同时，因为需要预测和更新标签变量，增加文本词向量分布项数，这样会增加计算量，致使 CBMeMBer 算法消耗的时间要大于 CNN 算法所消耗的时间。

4 结 论

针对基于神经网络分类算法在文本分类中的不足，提出了一种基于卷积神经网络的文本分类算法。通过使用 CNN 获得各文本的词向量分类，然后，利用各文本每时刻的分类词向量可以得到每时每刻的邻接矩阵分类，利用邻接矩阵分类得到每时

每刻的子群个数分类。仿真实验表明: CNN 算法在文本中的分类效果更为显著。

参考文献:

[1] Moeskops, P., Viergever, M. A., Mendrik, A. M., Vries, L. S. D., Benders, M. J. N. L., & I? gum, I. (2016). Automatic segmentation of mr brain images with a convolutional neural network [J]. IEEE Transactions on Medical Imaging, 2016,35(5):1252 - 1261.

[2] 吴祥标. Kemeny 社会选择函数的 0 - 1 规划算法 [J]. 遵义师范学院学报, 2014, 16(1):81 - 83.

[3] Sijin, L. I., Liu, Z. Q., & Chan, A. B. Heterogeneous multi - task learning for human pose estimation with deep convolutional neural network [J]. International Journal of Computer Vision, 2015,113(1):19 - 36.

[4] Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., & Mougiakakou, S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Transactions on Medical Imaging, 2016,35(5):1207 - 1216.

[5] Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network [J]. Knowledge - Based Systems, 2016,108:42 - 49.

[6] 陈宁江. 浅析 XML 文档和关系数据库之间的信息交换 [J]. 遵义师范学院学报, 2009, 11(3):72 - 74.

Text Classification Algorithm Based on Convolution Neural Network

WANG Mei - rong

(College of Information Engineering, Anhui Xinhua University, Anhui Hefei 230088, China)

Abstract: In order to solve the problem of high feature dimension and sparse data, a kind of convolution neural network is proposed to solve the classification algorithm in text classification. (convolutional neural network, CNN) text classification algorithm, the algorithm combines adjacency matrix convolution neural network theory in the dynamic modeling of text classification. Secondly, the training of word vector the text, and the classification of the adjacency matrix to obtain the group structure and the number of classification. Finally, the extracted text based abstract features with CNN classifier. The simulation results show that the algorithm is effective in the text classification.

Key words: text categorization; convolutional neural network; dynamic modeling; word vector

(上接 353 页)

Technique and Management Solution for Smart Substation Configuration Files

SUN Chen[1], *TIAN Xiao - sheng*[2]

(1. Shanghai Jiaotong University, Shanghai 200240, China; 2. State Grid Shanghai Municipal Electric Power Company, Shanghai 200122, China)

Abstract: Smart substations are indispensable part of building a unified, strong smart grid. The smart substation configuration file system reflects configuration information of all intelligent electronic devices in the station which are an important part of the operation and maintenance of smart substations. Extensive management pattern of existing configuration file causes versions of chaos, which can not adapt to the needs of reformation and extension program of smart substations. This article describes the technique of smart substation configuration files. Management solution for configuration file is put forward based on virtual terminals which simplify management of smart substation configuration files.

Key words: smart substation; configuration files; virtual terminals; management solution