Journal of Computer Applications ISSN 1001-9081 2018-11-27
计算机应用 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(****)**-0000-00 doi:10.11772/j.issn.1001-9081.2018081757

结合改进主动学习的 SVD-CNN 弹幕文本分类算法

邱宁佳*，丛琳，周思丞，王鹏，李岩芳

(长春理工大学 计算机科学技术学院，长春市 130022)

(*通信作者电子邮箱 760560291@qq.com)

摘 要: 为解决传统卷积神经网络（CNN）模型使用池化层进行文本特征降维会损失较多文本语义信息的问题，提出一种基于奇异值分解（SVD）算法的卷积神经网络模型（SVD-CNN），首先采用改进的基于密度中心点采样的主动学习算法（DBC-AL）选择对分类模型贡献率较高的样本进行标注，以低标注代价获得高质量模型训练集，然后结合 SVD 算法建立 SVD-CNN 弹幕文本分类模型，使用奇异值分解的方法代替传统 CNN 模型池化层进行特征提取和降维，并在此基础上完成弹幕文本分类任务；最后使用改进的梯度下降算法（PSGD）对模型参数进行优化。为了验证改进算法的有效性，使用多种弹幕数据样本集，在本文提出的模型与常用的文本分类模型上进行对比实验。实验结果表明，改进的算法能够更好地保留文本语义特征，保证训练过程的稳定性并提高了模型的收敛速度，在不同的弹幕文本上较传统算法具有更好的分类性能。

关键词: 卷积神经网络；奇异值分解；主动学习；梯度下降；文本分类

中图分类号:***　　**文献标志码**：**A**

SVD-CNN barrage text classification algorithm combined with improved active learning

QIU Ningjia*, CONG Lin, ZHOU Sicheng, WANG Peng, LI Yanfang

(College of Computer Science and Technology, Changchun University of Science and Technology, Changchun Jilin 130022, China)

Abstract: In order to solve the problem that the traditional Convolutional Network (CNN) model uses the pooling layer to reduce the dimension of text features which could lose much text semantic information. This paper proposed a convolutional neural network model based on Singular Value Decomposition algorithm (SVD-CNN). Firstly, an improved Active Learning algorithm based on density center point sampling (DBC-AL) was used to obtain the high-quality model training set at a low tagging cost, Then the SVD-CNN barrage text classification model was established by combining SVD algorithm, and the singular value decomposition method was used to replace the traditional CNN model pooling layer for feature extraction and dimension reduction, and completing the barrage text classification task on this basis. Finally, the model parameters were optimized by using the improved gradient descent algorithm (PSGD). In order to verify the effectiveness of the improved algorithm, a variety of barrage data sample sets were used to compare experiments between the proposed model and the common text classification model. The experimental results show that the improved algorithm can better preserve the semantic features of the text, ensure the stability of the training process and improve the convergence speed of the model. The classification performance on different barrage texts is better than traditional algorithms.

Keywords: Convolutional Nerual Network (CNN);Singular Value Decompostion (SVD); active learning; gradient descent; text categorization

0 引言

国内外研究者使用有监督的深度学习神经网络进行文本分类，这种监督型检测方法需要大量已标记数据,人工标注大量数据耗时耗力，因而难以实施。针对已有方法存在的问题，谭侃等提出一种基于双层采样主动学习方法，用样本不确定性、代表性和多样性来评估未标记样本的价值，使用排序和聚类相结合的双层采样算法对未标记的样本进行筛选，使用少量有标签样本达到与有监督学习接近的检测效果[1]。徐海龙等提出一种基于 委员会投票选择算法（Query by Committee，QBC）的支持向量机（Support Vector Machine，SVM）主动学习算法，将改进的 QBC 与加权 SVM 有机结合并应用于 SVM 训练学习中，有效的减少了样本分布不均衡

收稿日期: 2018-08-23;　　**修回日期**:2018-10-29;　　**录用日期**: 2018-11-08。

基金项目:吉林省重大科技招标项目(20170203004GX)；吉林省省级产业创新专项资金项目(2017C051)

作者简介: 邱宁佳(1984—)，男，河南南阳人，讲师，博士，CCF 会员（65280M），主要研究方向：数据挖掘、算法分析；丛琳(1992—)，女，吉林吉林人，硕士研究生，主要研究方向：数据挖掘；周思丞(1994—)，男，吉林长春人，硕士研究生，主要研究方向：数据挖掘；王鹏(1973—)，男，内蒙古包头人，教授，博士，CCF 会员（20695M），主要研究方向：数据挖掘；李岩芳(1965—)，女，吉林长春人，教授，博士，主要研究方向：数据库与数据挖掘、软件工程与信息系统。

对主动学习性能的影响[2]。姚拓中等将 Boosting 思想应用到多视角主动学习框架中，通过将历史上各次查询得到的分类假设进行加权式投票来实现每次查询后分类假设的强化，相比于传统单视角主动学习算法能够更快地完成收敛并达到较高的场景分类准确性[3]。Li 等提出了结合半监督的主动学习方法，将主动学习过程产生的价值样本用来加速分类器的训练，和伪标签一起辅助分类器进行高效的分类[4]。Wan 等提出了基于主动学习的伪标签校验框架，极大地提高了半监督学习中伪标签的置信度[5]。Wang 提出了主动学习与聚类相结合的伪标签校验的方法，进一步提高了伪标签的置信度[6]。Samiappan 等提出了 Co-Training 与主动学习算法进行组合的半监督算法，缓解了 Self-Training 中容易产生的数据倾斜问题而导致的分类器持续恶化的情况[7]。上述主动学习采样方法普遍面临以下问题：1）基于概率型的采样算法不适用句子型文本。2）只考虑分类结果最明确的样本，这种样本对当前分类器影响较小，并不能提高模型的泛化能力。本文提出基于密度中心点采样的主动学习算法，根据样本间的可连接性不断扩展聚类簇，选择每个类别中与密度中心相似度最高与最低的样本进行标注，实现采样的多样性，从而适用于大规模未标注句子级弹幕样本，使用极少量的标签样本训练初始分类器，迭代选择信息量最大的未标记弹幕样本加入训练集，以此提高分类器的分类性能，完成弹幕文本分类任务。

随着深度学习的发展，越来越多的深度学习模型被应用于短文本分类任务中，魏超等提出基于自编码网络的短文本流形表示方法实现文本特征的非线性降维，可以更好的以非稀疏形式更准确地描述短文本特征信息，提升分类效率[8]。谢金宝等提出一种基于语义理解的多元特征融合中文文本分类模型，通过嵌入层的各个通路提取不同层次的文本特征，较神经网络模型（Conveolutional Neural Network，CNN）与长短期记忆网络模型（Long Short-Term Memory，LSTM）的文本分类精度提升了 8%[9]。孙松涛等使用 CNN 模型将句子中的词向量合成为句子向量，并作为特征训练多标签分类器完成分类任务，取得了较好的分类效果[10]。Kalchbrenner 等提出 DCNN 模型，在不依赖句法解析树的条件下，利用动态 k-max pooling 提取全局特征，取得了良好的分类效果[11]。Kim 等采用多通道卷积神经网络模型进行有监督学习，将词矢量作为输入特征，可以在不同代销的窗口内进行语义合成操作，完成文本分类任务[12]。郑啸等结合 CNN 和 LSTM 模型的特点，提出了卷积记忆神经网络模型（Convolutional Memory Neural Network，CMNN），相比传统方法，该模型避免了具体任务的特征工程设计[13]。ST Hsu 将 CNN 与循环神经网络（Recurrent Neural Network，RNN）有机结合，从语义层面对 sentense 进行分类，取得良好的分类效果[14]。Yin 提出一种基于注意力机制的卷积神经网络，并将该网络用在句子对建模任务中，证明了注意力机制和 CNN 结合的有效性[15]。上述方法使用传统 CNN 模型对文本进行特征提取和分类，但池化操作在进行特征提取和降维时会损失较多的文本语义信息，从而导致分类精度下降。本文使用奇异值分解算法代替池化层的特征提取与降维工作，将奇异值较高的特征作为主要特征来代替原有目标矩阵的表达，更好地保存句子原有的语义结构，提升分类模型的精度。

1 相关研究

1.1 主动学习算法概述

主动学习算法是为了解决现实中标签数据不足，标注数据耗时耗力的问题而提出的。该算法能够从未标记样例中挑选部分价值量高的样例，标注后补充到已标记样例集中来提高分类器和精度，降低领域专家的工作量。如何高效地选出具有高分类贡献度的未标记样本进行标注并补充到已有训练集中，逐步提高分类器精度与鲁棒性是主动学习亟待解决的问题。

主动学习根据选择未标记样本方式的不同，可以分为成员查询综合主动学习、基于流的主动学习和基于池的主动学习。其中，基于委员会的主动学习是当前应用最广泛的采样策略。根据选择未标记样例的标准不同，基于池的采样策略又可分为：不确定性的采样策略、基于版本空间缩减的采样策略、基于模型改变期望的采样策略以及基于误差缩减的采样策略。

1.2 卷积神经网络文本分类

近年来，CNN 模型在文本分类任务上取得了很好的实用效果。CNN 模型首先根据输入文本和词向量构建输入矩阵，然后通过卷积和池化操作，筛选和组合词的分布式信息。其模型结构如图 1 所示，在这样一个网络中，输入层表示的是由每个词的分布式向量组成的句子矩阵；卷积层使用若干个卷积核对于局部的词向量矩阵进行卷积运算；池化层使用最大池化策略把卷积的结果转换为一组特征向量；基于前两层运算得到的特征向量，使用 Softmax 函数进行分类。

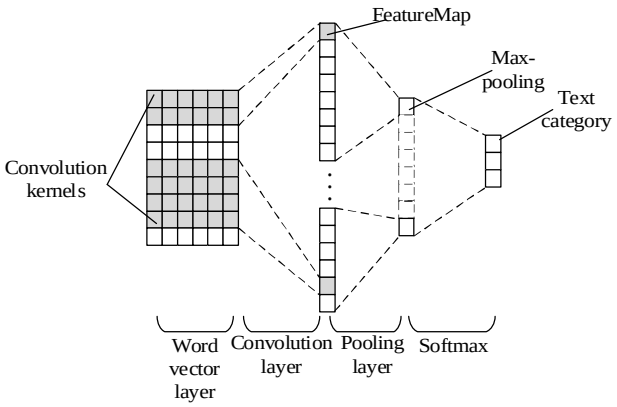

图1 CNN 模型结构

Fig. 1 CNN model structure

2 CNN 分类模型改进算法

2.1 主动学习算法改进

大多数传统的主动学习算法使用基于概率的启发式方法，这种方法建立在样例的后验概率分布基础之上，用信息熵较大的样本训练分类模型，这种基于概率的信息熵计算方法并不适用于句子级弹幕文本，所以本文在原有主动学习算法的思想基础上，提出一种基于密度中心点的主动学习采样算法，通过比对句向量间相似度与设定最小密度阈值对样本进行划分，根据阈值的约束条件来选择价值高的样本标注，提高了样本选择算法的鲁棒性。

2.1.1 基于相似度的密度聚类算法

传统的密度聚类算法是基于样本间距离的考察，本文针对句向量的空间分布提出向量间的相似度阈值来刻画样本类型的贴近程度，设计基于相似度的密度聚类算法，设置相似度与最小密度阈值，聚类核心步骤如下：

1）首先利用分词工具进行弹幕句子样本分词，将句子以词为单位形成一个词向量序列，如式（1）所示：

$$x_{1:n}=[x_1,x_2,...,x_n] \tag{1}$$

然后使用 Word2vec 模型将每一个词映射为一个多维的连续值词向量序列，最后利用 LSTM 算法，将词向量序列结合文本语序信息生成语义向量，公式表达如下：

$$\tilde{C}_t=\tanh\left(w_C\cdot[g_{t-1},x_t]+b_C\right) \tag{2}$$

$$C_t=f_t\times C_{t-1}+i_t\times\tilde{C}_t \tag{3}$$

$$g_t=o_t\cdot\tanh(C_t) \tag{4}$$

其中，$\tilde{C}_t$ 表示前一时刻的 cell 状态，C_t 表示当前时刻 cell 状态，f_t 表示遗忘门限，i_t 表示输入门限，O_t 表示输出门限，g_t 表示当前单元的输出，tanh 表示激活函数。

2）设置样本相似度阈值 g 与最小密度样本数 $MinPts$，如式（5）、（6）所示：

$$\frac{\sum_{i=1}^{n}\left(g_i\times g_j\right)}{\sqrt{\sum_{i=1}^{n}\left(g_i\right)^2}\times\sqrt{\sum_{i=1}^{n}\left(g_j\right)^2}}\geq g \tag{5}$$

$$\left|N_{\epsilon}\left(g_j\right)\right|\geq MinPts \tag{6}$$

其中，g_j 为核心密度点，$N_{\epsilon}\left(g_j\right)$ 表示以 g_j 为中心点的邻域。

3）从步骤（1）中筛选出符合步骤（2）中条件的点，加入到核心对象集合 Ω 中，如式（7）所示：

$$\Omega=\Omega\cup\{g_j\} \tag{7}$$

4）在核心对象集合中随机选取一个点 a，找出由它密度可达的所有样本，生成第一个聚类簇 B_1。

5）将 B_1 中包含的核心对象从 Ω 中去除，再从更新后的 Ω 中随机选取一个核心对象，作为种子来生成下一个聚类簇，反复迭代上述步骤，直至 Ω 为空。

2.1.2 主动学习采样策略

普通的主动学习采样策略，存在采样单一、采样偏置的问题。结合样本在特征空间中的分布结构，本文提出一种带约束条件的主动学习采样策略对未标记样本进行筛选，以聚类簇为单位，计算聚类中心点与其他样本间的相似度，其中相似度最高与最低的样本最能代表整个聚类簇的分布状态，依据上述方法可以在样本的信息性和预测标号的准确性两者之间获得较好的平衡，选出最有价值的弹幕样本给专家标注。核心步骤如下：

1）计算属于同一个密度中心点 g_j 阈值范围内的所有样本 g_i 到密度中心的相似度 $SimG$，如式（8）所示：

$$SimG=\frac{g_j\times g_i}{\sqrt{g_j^{\,2}}\times\sqrt{g_i^{\,2}}} \tag{8}$$

其中，g_i 为中心点 g_j 阈值范围内的所有样本。

2）选择相似度最高与最低的两个样本：$MaxSim$、$MinSim$，如式（9）、（10）所示：

$$MaxSim=\arg\max(SimG) \tag{9}$$

$$MinSim=\arg\min(SimG) \tag{10}$$

3）将 $MaxMin$、$MinSim$ 进行人工标注，加入到训练样本集 S 中。

结合上述两个算法将基于密度中心点采样的主动学习算法（Density-based Clustering of Active Learning，DBC-AL）归纳如下：

算法 1：DBC-AL 算法

输入：样本集 $D=\{g_1,g_2,...,g_m\}$，邻域参数 $(e,MinPts)$

输出：待标注样本 $MaxSim$、$MinSim$

1. 初始化核心对象集合 $\Omega=j$
2. $for\ j=1,2,...,m\ do$
3. $N_e\left(g_j\right)$ //确定 g_j 样本的 e-邻域

4. *if* $\left|N_e(g_j)\right| \geq MinPts$ *then*

5. $\Omega = \Omega \cup \{g_j\}$ //将样本 g_j 加入核心对象集合

6. *end if*

7. *end for*

8. 初始化聚类簇数：$l=0$

9. 初始化未访问样本集合：$z = D$

10. *while* $\Omega \neq j$ do

11. 记录当前未访问样本集合：$z_{old} = z$

12. $a = random\{\Omega\}, Q = \langle a \rangle, G = f$ //随机选取一个核心对象 $a \in \Omega$，初始化队列 $Q = \langle a \rangle$，初始化核心对象集合 $G = f$

13. $z = z - \{a\}$， $G = G \cup a$

14. *while* $Q \neq f$ *do*

15. 取出队列 Q 中的首个样本 q

16. *if* $\left|N_e(q)\right| \geq MinPts$ *then*

17. 令 $\Delta = N_e(q) \cap z$

18. 将 Δ 中的样本加入队列 Q

19. $z = z - \Delta$

20. *end if*

21. *end while*

22. $l = l+1$， $B_l = z_{old} - z$ //生成聚类簇

23. $\Omega = \Omega - B_l$

24. *end while*

25. *while* $B \neq \varnothing$ *do*

26. 取出 B 中一个簇 B_l

27. $SimG = \dfrac{g_j \times g_i}{\sqrt{g_j^2} \times \sqrt{g_i^2}}$ //由公式（8）计算 B_l 中每个元素到密度中心 O_l 的相似度 $SimG$ 并加入集合 S

28. $MaxSim = \arg\max(SimG)$
$MinSim = \arg\min(SimG)$ //取出集合 S 中最大元素 $MaxSim$ 和最小元素 $MinSim$ 加入待标注样本集 R

29. *end while*

2.2 SVD-CNN 模型

在自然语言领域，传统的 CNN 使用池化层对文本进行采样降维工作，该操作只是简单的从前一维 FeatureMap 中提取了最大值，并不关心特征的分布状态，从而导致特征的位置信息丢失，文本语义发生变化的问题。本文使用奇异值分解算法（Singular Value Decomposition，SVD）代替池化层的特征提取工作，根据奇异值的大小选取矩阵的主要特征。奇异值往往对应着矩阵中隐含的重要信息，每个目标矩阵都可以表示为一系列秩为 1 的特征矩阵之和，而奇异值则表征了这些特征矩阵对于目标矩阵的权重，因此奇异值较高的特征能够作为主要特征来代替原有目标矩阵的表达。如式（11）所示：

$$A = U_1 d_1 V_1^T + U_2 d_2 V_2^T + \ldots + U_n d_n V_n^T \tag{11}$$

其中，A 为目标矩阵，U_n 为左奇异值矩阵，d_n 为特征矩阵，V_n^T 为右奇异值矩阵。

本文在传统 CNN 分类模型基础上设计了基于奇异值分解算法的卷积神经网络模型（Convolutional Neural Network based on Singular Value Decomposition，SVD-CNN），利用 SVD 算法良好的数值稳定性和几何不变性完成对矩阵的主要特征提取和降维，较好的保留文本语义信息的完整性，整个模型体系如图 2 所示。

1）输入层。

模型的输入为一个 $n \times m$ 的句子矩阵，矩阵的每一行代表句子中每个词对应的向量，行数 n 代表句子的词数，列数 m 为向量的维数。

2）卷积层。

采用列数与行数相同的卷积矩阵窗口 $h \in R^{n \times m}$，为了获取不同类别的语义特征，采用多个不同尺寸（ h ）的卷积窗口与原矩阵进行卷积运算，得到卷积语义特征 F_i，如式（12）所示。

$$F_i = relu\left(W \cdot x_{i:i+h-1} + b\right) \tag{12}$$

3）奇异值分解层。

对特征矩阵 F_i 进行奇异值分解运算，降维后的特征矩阵记为 A，如式（13）所示：

$$A = f\left(\sum_{i=0}^{k} w_a U_i \frac{d_i}{\|d_i\|} V_i^T + b_a\right) \tag{13}$$

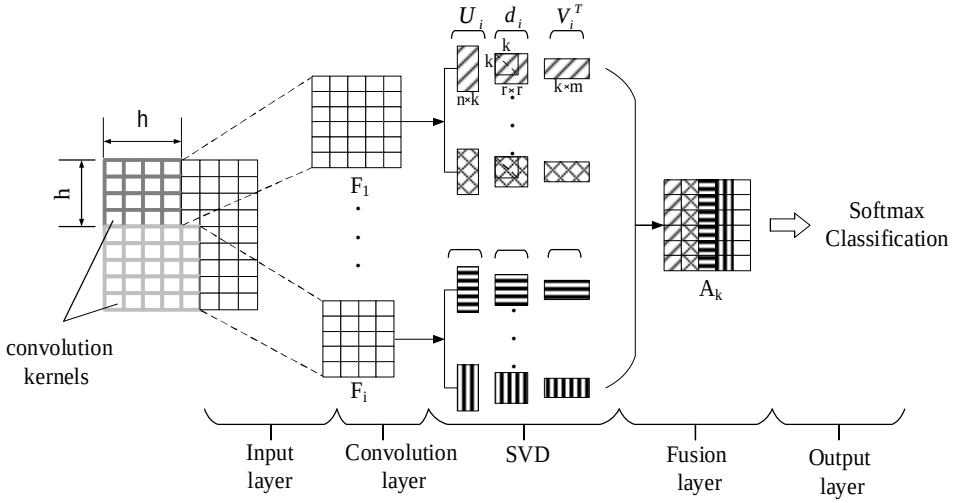

图2 SVD-CNN 模型

Fig. 2 SVD-CNN model

为了使各个特征处于同一数量级，在奇异值分解后，对各个特征进行归一化处理，$\frac{d_i}{\|d_i\|}$ 代表第 i 个特征的归一化奇异值，w_a 表示权重矩阵，其中 $U_i V_i^T$ 表示文本的第 i 个特征向量，b_a 表示偏置项。选择奇异值较大的特征向量组成原矩阵的主要特征向量，完成降维和特征提取工作，降维后的维数 k 可以通过设置阈值 t 来界定，如式（14）所示：

$$\frac{\sum_{i=1}^{k} d_i}{\sum_{i=1}^{n} d_i} \geq t \tag{14}$$

根据阈值选择前 k 个奇异值与其对应的标准正交基，构建原矩阵 A 的 k 秩近似矩阵 A_k 如式（15）所示：

$$A_k = \sum_{i=1}^{k} U_i d_i V_i^T \tag{15}$$

其中，U_i 为左奇异矩阵，d_i 为奇异值矩阵，V_i^T 为右奇异矩阵。

（4）输出层。

将得到的多个 A_k 矩阵融合，记为 S，如式（16）所示：

$$S = \{A_1^k, A_2^k, ..., A_l^k\} \tag{16}$$

其中，l 为语义特征 F_i 的个数，使用 Softmax 函数对 S 进行分类，最终输出弹幕样本在不同类别上的分布概率。

3 弹幕分类解决方案的构建

3.1 弹幕分类模型优化算法

考虑到深度学习模型是较复杂的非线性结构，在这种非凸问题上往往很难直接求解，所以本文采用梯度下降算法对模型参数进行优化以得到全局最优解。

兼顾随机梯度下降算法（Stochastic gradient Descent，SGD）的随机性，本文设计一种通过选取数据相关性较高的样本来形成批量数据训练集的梯度下降算法（Partial Sampling Gradient Descent，PSGD），该算法在保证训练过程稳定性的同时，提高模型的学习速度，使模型更快速地收敛，参数更新公式如式（17）所示：

$$q \leftarrow q - e\nabla_q J(q) \tag{17}$$

其中，q 为优化参数，e 为学习率，$\nabla_q J(q)$ 为参数梯度。

考虑到随机选取训练样本的不确定性可能会导致目标函数值出现震荡的现象，本文从模型正确预测出的数据集中随机抽取 10%样本，结合所有错误预测的样本，形成新的训练集来训练模型，具体算法描述如下：

算法 2：PSGD 梯度下降算法

输入：全样本训练集 U，误差函数 loss 和迭代终止阈值 p，学习率 e，初始参数 q

输出：更新后的参数 q

1. 初始化 SVD-CNN 模型
2. 使用全样本训练集 U 对模型进行训练
3. *while* $loss > p$ *do*
4. 从训练集 U 中的所有 n 个样本（m 个正确样本，$n-m$ 个错误样本）中，采包含 z 个正确样本 $\{x^{(1)}, ..., x^{(z)}\}$ 和 $n-m$ 个错误样本 $\{x^{(n-m+1)}, ..., x^{(n)}\}$

的小批量，其中 $\frac{z}{m}=0.1$， $x^{(i)}$ 对应目标为 $y^{(i)}$ 。

5. $J(q)\leftarrow+\frac{1}{z+n-m}\nabla_q\sum_i L\left(f\left(x^{(i)},y^{(i)}\right),y^{(i)}\right)$ // 计算梯度估计

6. $q\leftarrow q-eJ(q)$ //应用更新

7. *end while*

8. *return* q

3.2 模型描述

本文使用改进的主动学习采样策略,从未标注弹幕样本集中根据算法 1 设定的规则挑选少量弹幕样本，交由人工标注，使用标注好的弹幕样本训练 SVD-CNN 分类模型，为了能够较好保存句子的语义信息,模型使用 SVD 算法代替池化层来完成特征提取与特征降维，将得到的主要特征进行信息融合，并输入到 Softmax 函数中完成分类任务。整体结构如图 3 所示：

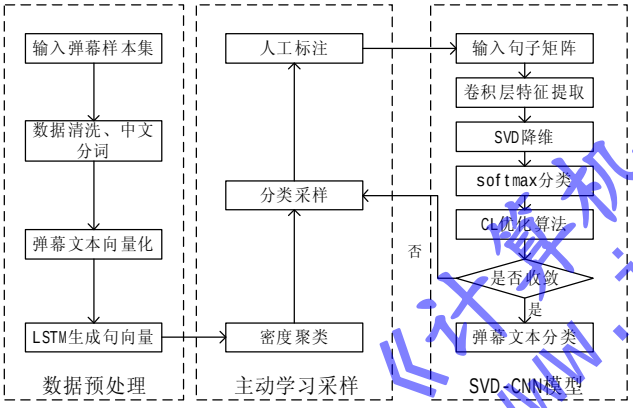

图3 SVD-CNN 算法解决方案

Fig. 3 SVD-CNN model solution

在数据预处理阶段，首先对弹幕样本进行数据清洗，然后利用分词工具与 Word2vec 模型将每一个词映射为一个多维的连续值词向量序列，最后利用 LSTM 模型将词向量序列结合文本语序信息生成语义向量。

使用算法 1 中的方法，将句向量样本根据相似度阈值和最小密度样本数进行聚类，对每个密度中心点相似度临界值进行采样，得到最能代表每个聚类簇的整体样本分布状态，将样本交由专家标注，以此提高训练样本集的代表性和广泛性。

为了获取不同的语义特征，采用不同尺寸的卷积窗口与原矩阵进行卷积运算，得到卷积语义特征 F_i，对特征矩阵 F_i 进行奇异值分解运算，根据设定的阈值选择前 k 个奇异值与其对应的标准正交基，构建原矩阵 A 的 k 秩近似矩阵，将多个 A_k 矩阵的融合 S ，通过 Softmax 函数计算得到样本属于各个类的概率分布，如式（17）所示：

$$p(y=i\mid Q)=\frac{\exp\left(q_i^{T}Q\right)}{\sum_{j=1}^{d}\exp\left(q_j^{T}Q\right)},i=1,2,...,d. \quad (18)$$

其中， Q 为一条弹幕样本， q_i^T 表示 S 中第 i 个特征的权重矩阵， $p(y=i\mid Q)$ 表示每个样本属于类别 i 的概率， d 为弹幕类别数。为了更好地训练模型，本文采用 loss 函数来衡量文本类别的真实概率分布 P_j 和预测的概率分布 h_j 之间的差距，如式（18）所示：

$$loss=-\sum_{Q\in R}\sum_{j=1}^{d}P_j\ \log_2 h_j \quad (19)$$

4 实验与结果分析

4.1 实验数据与参数设置

本文针对三个方面对改进算法的有效性进行验证。第一，通过模型不同的分类准确率，对比传统采样算法和 DBC-AL 算法的模型迭代次数，验证后者具有更高的效率；第二，使用本文提出的 SVD-CNN 模型对弹幕文本分类，同时考虑词向量维度和数据集泛化能力来验证其分类性能；第三，使用改进后得梯度下降算法对模型进行优化，通过收敛速度和模型训练速度来验证优化算法的有效性。

本文通过爬虫技术在不同视频网站分别爬取弹幕文本，根据视频类别形成不同的数据集进行对照试验，对本文提出的算法进行性能评估。详细的实验数据统计如表 1 所示：

表1 实验使用的数据集

Tab. 1 Experimental Data Statistics

数据集	积极倾向样本数量	消极倾向样本数量
Bilibili 弹幕训练集	1138	1362
Bilibili 弹幕测试集	1226	1274
爱奇艺弹幕训练集	1098	1402
爱奇艺弹幕测试集	1336	1164
优酷视频弹幕测试集	1256	1101
优酷视频弹幕测试集	1339	1206

本文在实验中选择不同尺寸的卷积核对输入的句子矩阵进行卷积操作，结合设定的阈值选取特征，使用奇异值分解算法完成矩阵的特征降维和特征提取，具体参数设置如表 2 所示：

表2 模型参数设置

Tab. 2 Model Parameter Settings

参数	参数描述	值
h	窗口大小	3,4,5
F	特征图数	32

P	Dropout 大小	0.5
e	学习率	10^{-3}
k	奇异值特征选取阈值	80%

4.2 算法性能验证

实验 1 主动学习采样算法性能验证

通过算法 1 中的方法对表 1 中的弹幕样本进行采样，为了验证该算法对短文本语句向量在减少人工标注上所起到的作用，比较在模型达到同一分类准确率时，不同采样算法所需的迭代次数。实验选择 QBC 算法，随机采样算法，基于最优标号和次优标号的（Best vs Second-best，BvSB）[16]算法作为对照实验，使用 CNN 分类模型，实验结果如图 4 所示：

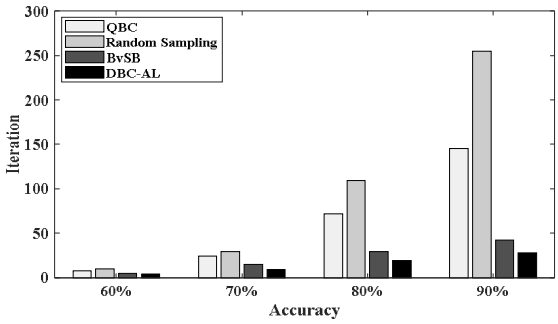

图4 不同模型精度下各个算法采样次数对比

Fig. 4 Comparison of sampling times of each algorithm under different model accuracy

从图 4 可以看出，在分类器识别精度为 60%时，模型的分类正确率较低，除随机采样算法外，其余 3 种分类算法使用采样所需的迭代次数没有明显差距，随着分类精度从 70%逐渐提升到 90%时，随机采样和信息熵采样算法所需迭代次数有着明显的升高，而 DBC-AL 算法和 BvSB 算法相对较为稳定。由于 BvSB 算法只考虑样本分类可能性最大的类别，因此相对前两种算法来说采样次数较少，但该算法忽略其他对样本的分类结果影响较小的类别，导致该算法采集的样本所含的信息量较少，相对于本文提出的 DBC-AL 算法需要更多的迭代次数，这说明了随着模型精度的提高，前三种传统的采样算法收集到的样本信息对于模型收敛提供的帮助越来越少，而 DBC-AL 算法根据样本间的相似度进行聚类，对每个聚类簇采集到对分类模型来说最有价值的样本，从而体现了 DBC-AL 算法在句向量中采样的优越性。

实验 2 模型分类性能对比

本文采用 SVM 算法，传统 CNN 模型，不加池化层的 CNN 模型，多通道卷积神经网络模型[17]（Multi-channels Convolution Neural Networks，MCCNN）和本文提出的 SVD-CNN 模型进行分类正确率对比实验验证 SVD-CNN 模型的有效性。考虑到不同数据集可能引起分类模型精度变化的问题，使用表 1 中 3 种数据集分别进行模型分类性能对比实验，实验结果如图 5 所示：

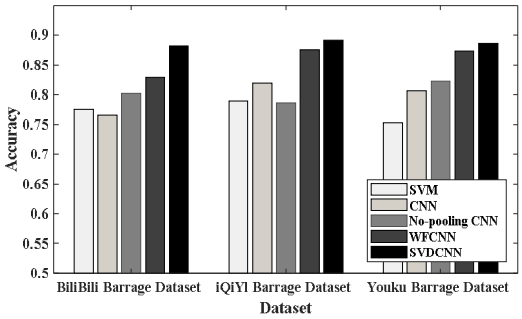

图5 不同数据集下各个分类模型的分类性能

Fig. 5 Classification Performance of Different Classification Models

从图 5 可以看出，SVM 模型最高取得了 78.9%的分类正确率，说明 SVM 模型在多分类问题上精度较低。CNN 的分类正确率受数据集影响波动较大，在 BiliBili 弹幕数据集上的分类精度降低到 76.6%，相比不加池化层的 CNN 模型分类正确率略有下降，这说明传统 CNN 模型的池化层并不能对于自然语言的文本特征进行有效提取。由于 MCCNN 模型采用多通道的特征提取方式，将不同的特征信息结合形成不同的通道作为卷积神经网络的输入，使得模型的分类效果优于前两种模型，最高分类精度达到了 87.6%，而本文提出的 SVD-CNN 模型相比前三个数据集上取得了最好的弹幕分类效果，其中在爱奇艺弹幕数据集上的分类精度最高达到了 89.2%，相对于传统使用池化层的 CNN 模型和 MCCNN 模型，分别提高了 7.3%和 1.6%，说明本文提出的对文本语义矩阵使用奇异值分解算法进行降维和特征提取的方法，更好地保留了文本语义特征，进而提高了模型分类精度，充分验证了 SVD-CNN 模型在处理文本语义分类上对特征信息选择的有效性。

实验 3 句向量维度实验

考虑到句向量维度会影响文本语义信息的表征，从而影响最终的分类结果，本文利用 CNN 模型，MCCNN 模型和 SVD-CNN 模型在 BiliBili 弹幕数据集上使用不同维度的句向量进行对比实验，分析句向量维度对分类结果的影响，实验结果如表 3 所示：

表3 不同句向量维度下模型分类正确率对比

Tab. 3 Accuracy of model classification under different sentence vector dimensions

分类模型 / 词向量维度	CNN	MCCNN	SVD-CNN
10	0.803	0.854	0.852
50	0.836	0.876	0.892
100	0.831	0.870	0.883
150	0.825	0.871	0.890
200	0.828	0.865	0.886

从表 3 中可以看出，当句向量的维度增加到 50，三种模型的分类精度都有着不同程度的提高，这说明随着句向量维

度的增加，文本语义的特征表达能力在逐渐提高。当维度继续增加时，语句特征分布会变得更为稀疏，CNN 模型与 MCCNN 模型使用池化层会忽略较多的文本语义特征，致使分类效果降低。本文使用 SVD 算法代替池化层进行特征提取，在语句特征分布较为稀疏的情况下仍然可以保留较多的文本语义特征，当词向量维度增加到 180 以上时，模型的分类精度仍处于平稳状态，充分验证了 SVD-CNN 在弹幕文本语义特征提取上的高效性。

实验 4 弹幕分类模型优化算法

为了验证本文 PSGD 算法的训练稳定性与训练速度，选择批量梯度下降算法（Batch Gradient Descent，BGD），SGD 算法，小批量梯度下降算法（Mini-Batch Gradient Descent，MBGD）和本文提出的 PSGD 算法设计两组对照实验来检验算法性能，使用表 1 中 BiliBili 弹幕数据集共 10000 条，设置迭代阈值为 2500 次，设计实验对比模型训练时误差变化率，验证 PSGD 算法的稳定性，如图 6 所示；设计实验对比模型分类精度随时间的变化率，验证 PSGD 算法具有较低的时间复杂度，如图 7 所示：

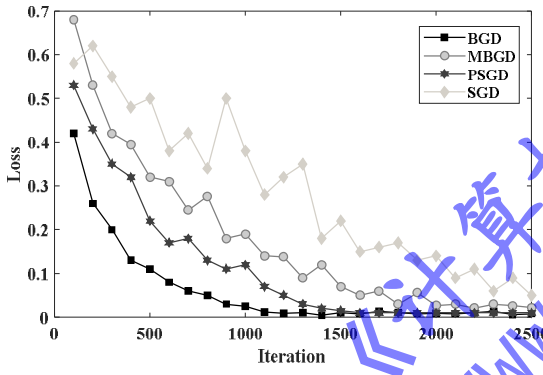

图6 不同的迭代次数对模型训练稳定性的影响

Fig. 6 Effect of different iterations on model training ability

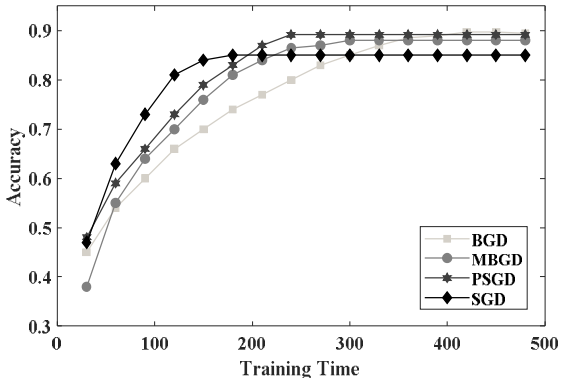

图7 不同训练时间下模型分类正确率对比

Fig. 7 Comparison of model's correct classification rate under different training time

由实验结果可以看出，随着迭代次数的增加，使用 BGD 算法进行优化的模型误差逐渐减小，训练过程比较平稳，模型分类精度较高，但由于该算法采用全样本训练的方式，导致模型训练时间长，模型训练速度慢；SGD 算法每次随机选取样本进行训练，训练时间较短，但相对于 BGD 算法存在的噪音较多，导致每次迭代并没有向着整体最优化方向进行，因此 SGD 的训练过程稳定性较差，模型易陷入局部最优点，致使分类精度降低；MBGD 算法每次迭代使用部分样本更新模型参数，相对于 SGD 算法训练过程比较稳定，训练时间较短，模型分类精度介于 BGD 算法与 SGD 算法之间；由于本文提出的 PSGD 算法将模型分类错误的样本引入到训练集中，相对于 MBGD 算法的随机性训练集包含更多的信息量，所以可以使模型训练时间更短，训练过程更稳定，模型分类精度更高，从而验证了 PSGD 算法的有效性。

5 结语

本文提出一种基于密度中心点采样的主动学习算法，利用样本间的相似度将样本进行聚类，并在每一个聚类簇中，按照设定的规则选择最具有价值的样本进行人工标注，减少人工标注的工作量；提出 SVD-CNN 模型，使用 SVD 算法代替传统 CNN 模型的池化层，更好的保留了文本语义特征，从而提高模型的分类精度；使用改进的 PSGD 算法选取信息量较大的训练样本对模型进行优化，保证了训练过程稳定性的同时提高了模型的训练速度。通过对比不同主动学习采样算法性能实验表明，DBC-AL 算法较传统的主动学习采样算法采集到的样本信息量更高，对模型的分类贡献更多；对比多种数据集和不同句向量维度下分类模型的分类精度，可以看出 SVD-CNN 模型能够提取到更多的文本语义特征，具有较高的分类准确率；对比不同模型优化算法的训练误差与训练时间，PSGD 算法具有良好的稳定性，模型收敛速度更快，总体训练效果优于其他算法。在主动学习采样的规则条件中，采样的阈值是通过经验选取，可能并不是最优的，如何根据数据集及当前分类模型来对该阈值进行自适应的调整是下一步工作中需要考虑的重要问题。

参考文献

[1] 谭佩, 高旻, 李文涛,等. 基于双层采样主动学习的社交网络虚假用户检测方法[J]. 自动化学报, 2017, 43(3):448-461. (TAN K, GAO M, LI W T, et al. Two-layer sampling active learning algorithm for social spammer detection[J]. Acta Automatica Sinica, 2017, 43(3): 448-461.)

[2] 徐海龙, 别晓峰, 冯卉,等. 一种基于 QBC 的 SVM 主动学习算法[J]. 系统工程与电子技术, 2015, 37(12):2865-2871.(XU H L, BIE X F, FENG H, et al. Active learning algorithm for SVM based on QBC,[J]. Journal of Systems Engineering and Electronics,2015, 37(12):2865-2871.)

[3] 姚拓中, 安鹏, 宋加涛. 基于历史分类加权和分级竞争采样的多视角主动学习[J]. 电子学报, 2017, 45(1):46-53.(YAO T Z, AN P, SONG J T. Multi-view active learning based on weighted hypothesis boosting and hierarchical competition sampling [J]. Atca Electronica Sinica, 2017, 45(1):46-53.)

[4] LI M, WANG R, TANG K. Combining semi-supervised and active learning for hyperspectral image classification[C]// Computational Intelligence and Data Mining. IEEE, 2013:89-94.

[5] WAN L, TANG K, LI M, et al. Collaborative active and semisupervised learning for hyperspectral remote sensing image classification[J]. IEEE Transactions on Geoscience & Remote Sensing, 2015, 53(5):2384-2396.
[6] WANG Z, DU B, ZHANG L, et al. A Novel semisupervised Active-Learning algorithm for hyperspectral image classification[J]. IEEE Transactions on Geoscience & Remote Sensing, 2017, 55(6):3071-3083.
[7] SAMIAPPAN S, MOORHEAD R J. Semi-supervised co-training and active learning framework for hyperspectral image classification[C]// Geoscience and Remote Sensing Symposium. IEEE, 2015:401-404.
[8] 魏超, 罗森林, 张竞,等. 自编码网络短文本流形表示方法[J]. 浙江大学学报:工学版, 2015, 49(8):1591-1599.(WEI C, LUO K L, ZHANG J, et al. Short text manifold representation based on autoencoder network[J]. Journal of Zhejiang University(Engineering Science), 2015, 49(8):1591-1599.)
[9] 谢金宝, 侯永进, 康守强,等. 基于语义理解注意力神经网络的多元特征融合中文文本分类[J]. 电子与信息学报, 2018(5).(XIE J B, HOU Y J, KANG S Q, et al. Multi-feature fusion based on semantic understanding attention neural network for Chinese text categorization. Journal of Electronics&Information Technology, 2018(5).)
[10] 孙松涛, 何炎祥. 基于 CNN 特征空间的微博多标签情感分类[J]. 四川大学学报(工程科学版), 2017, 49(3):162-169.(SUN S T, HE Y X. Multi-label Emotion classification for microblog based on CNN feature space[J]. Advanced Engineering Sciences, 2017, 49(3):162-169.)
[11] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A Convolutional neural network for modelling sentences[J]. Eprint Arxiv, 2014, 1.
[12] KIM Y. Convolutional neural networks for sentence classification[J]. Eprint Arxiv, 2014.
[13] 郑啸, 王义真, 袁志祥,等. 基于卷积记忆神经网络的微博短文本情感分析[J]. 电子测量与仪器学报, 2018(3).(ZHENG X, WANG Y Z, YUAN Z X, et al. Sentiment analysis of micro-blog short text based on convolutional memory neural network[J]. Journal of Electronic Meatruement and Instrumentation, 2018(3).)
[14] HSU S T, MOON C, JONES P, et al. A hybrid CNN-RNN alignment model for phrase-aware sentence classification[C]// Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017:443-449.
[15] YIN W, SCHUTZE H, XIANG B, et al. ABCNN: Attention-based convolutional neural network for modeling sentence pairs[J]. Computer Science, 2015.
[16] 陈荣, 曹永锋, 孙洪. 基于主动学习和半监督学习的多类图像分类[J]. 自动化学报, 2011, 37(8):954-962.(CHEN R, CAO Y F, SUN H. Multi-class image classification with active learning and semi-supervised learning [J]. Acta Automatica Sinica, 2011, 37(8):954-962.)
[17] 陈珂, 梁斌, 柯文德,等. 基于多通道卷积神经网络的中文微博情感分析[J]. 计算机研究与发展, 2018(5).(CHEN K, LIANG B, KE W D, et al. Chinese icro-Sentiment analysis based on Multi-Channels convolutional neural networks[J]. Journal of Computer Research and Development, 2018(5).)

This work is partially supported by the Industrial Technology Research and Development Special Project of Jilin Province(2016C090).

QIU Ningjia, born in 1984, Ph. D., lecturer. His research interests include data mining, analysis of algorithms.
CONG Lin, born in 1992, M. S. candidate. Her research interests include data mining.
ZHOU Sicheng, born in 1994, M. S. candidate. His research interests include data mining.
WANG Peng, born in 1973, Ph. D., professor. His research interests include data mining.
LI Yanfang, born in 1965, Ph. D., professor. Her research interests include database and data mining, software engineering and information system.