

# 基于电子作业挖掘的学生学习预警模型研究

张笑非, 段先华, 刘 镇, 钱 萍

(江苏科技大学计算机学院, 江苏 镇江 212003)

**摘 要:** 混合式教学的普及使得电子作业成为一种评价学生学习效果的重要数据来源, 利用机器学习对电子作业进行建模是对学生学习预警的一种有益探索。本文在对电子作业进行分词和向量化基础上, 通过k-means聚类 and 轮廓系数来判断其语义的多样性, 通过计算文档向量相似性矩阵的网络效率来评价电子作业的中心性。实验结果显示, 该方法可以有效寻找电子作业聚类效果最优时的簇类多样性, 也可以有效评价电子作业相似度的网络中心性。因此, 该方法作为一种学生学习预警模型, 可以对电子作业文档的多样性和中心性给出客观的总体评价。

**关键词:** 文档向量; k-means聚类; 轮廓系数; 文档相似度; 图论效率

**中图分类号:** TP181 **文献标识码:** A

## Research on School Precaution Model Based on Electronic Assignments Mining

ZHANG Xiaofei, DUAN Xianhua, LIU Zhen, QIAN Ping

(School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

**Abstract:** The popularity of hybrid teaching makes electronic assignment an important data source for evaluating students' learning effects. Modeling electronic assignment with machine learning is a useful exploration for school precaution. Based on the word segmentation and vectorization of electronic assignments, this paper determines the semantic diversity by k-means clustering and silhouette coefficient, and evaluates the centrality of electronic assignment by calculating the network efficiency of document vector similarity matrix. The experimental results show that the method can effectively find the cluster diversity when the clustering effect of electronic assignments is optimal, and can also effectively evaluate the network centrality of the similarities of electronic assignments. Therefore, as a school precaution model, this method can give an objective overall evaluation of the diversity and centrality of electronic assignments.

**Keywords:** document vector; k-means clustering; Silhouette coefficient; documents' similarity; graph theoretic efficiency

## 1 引言(Introduction)

随着慕课(Massive Online Open Courses, MOOC)和翻转课堂(Small Private Online Course, SPOC)提供了在线远

程学习的条件<sup>[1]</sup>, 对学生的自主学习态度和学习能力的要求也越来越高, 如何同样利用网络平台及人工智能等手段检验和评价学生的学习效果也是一个亟待解决的问题。作业电子化

和网络化的尝试,一方面减少了对纸张的使用,符合各高校类似办公自动化的绿色理念;另一方面,这意味着可以利用各种先进的信息技术对电子作业进行收集、分类、管理、挖掘等操作,从中可以挖掘电子作业文档的特征并构建模型,充当对慕课和翻转课堂等“教”的环节补充,从中挖掘学生“学”的情况。

电子作业抄袭检测也一直是教育信息技术应用的一个研究热点。文献[2]提出了基于网络邮件的作业自动收集技术,并通过句子相似度的方法来发现抄袭的方法。文献[3]针对机房环境和网络环境,分别采用了信息隐藏技术和向量空间距离模型进行电子作业反抄袭。文献[4]利用空间向量模型及相似度分布图辅助教师进行作业抄袭甄别工作。文献[5]采用了直觉模糊聚类的方法来进行电子作业抄袭检测。本文通过将电子作业文档向量化,采用k-means聚类和相似性矩阵进行建模,基于轮廓系数和网络效率对电子作业文档样本整体质量进行评价,以此作为学生学习预警依据。

## 2 电子作业文档预处理(Preprocessing of Electronic Assignments)

### 2.1 电子作业文档分词

学生学习预警模型的建立首先需要能够对电子作业文档进行分词,由于学生的电子作业是以中文为主,所以这里采用了“结巴”中文分词组件。该组件是基于前缀词典实现高效的词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图,并采用了动态规划查找最大概率路径,能够找出基于词频的最大切分组合。而对于未登录词,则采用了汉字成词能力的隐含马尔可夫模型,并通过Viterbi算法进行分词。

### 2.2 电子作业文档向量化

文档向量化是对词向量化算法的一种扩充,通过对句子、段落、甚至整个文档这样的大块文本进行无监督学习,从而得到文档的定长特征向量。传统的文本定长特征向量使用词袋(bag-of-words)时存在两个不足,即词袋内词是无序的、且不考虑词的语义信息。文献[6]提出了名为“段落向量”的非监督学习算法,可以从变长的文本单元(如语句、段落、文档)中学习得到定长的特征向量。该算法将文档表达为密集向量用于训练模型,并以此预测文档中的词。

Doc2vec(<https://radimrehurek.com/gensim/models/doc2vec.html>)是一个有效的能够为分词后的文档创建向量的工具,图1中是通过Doc2vec对一门课程61名学生的电子作业文档进行向量化后得到的结果,每份电子作业文档被表示为具有100个特征的特征向量。

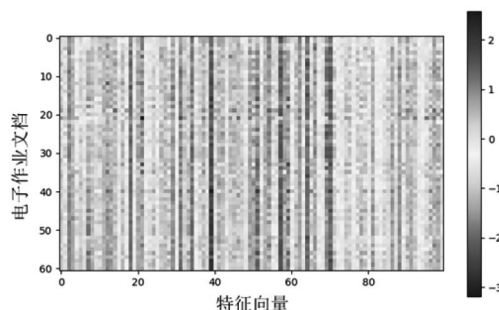


图1 电子作业文档的特征向量

Fig.1 Feature vectors of electronic assignments

## 3 学习预警建模(Modeling of school precaution)

### 3.1 聚类分析

K-means聚类作为一种常用的非监督学习算法,将一组观测向量 $(x_1, x_2, \dots, x_n)$ 分配到 $k$ 个簇类中,其中 $2 \leq k \leq n$ 。在 $k$ 选定后,算法要使得簇类方差和最小,其目标函数如公式(1):

$$\arg \min_S \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 = \arg \min_S \sum_{i=1}^k |S_i| \text{Var} S_i \quad (1)$$

其中, $\mu_i$ 是簇类 $S_i$ 的中心点,即 $S_i$ 内所有向量的均向量。本文将电子作业文档向量通过k-means算法聚类,分别选择 $k=2, 3, 4, \dots, 49$ 进行计算。图2中分别给出了当 $k=4, 5, 6, 7$ 时的聚类结果,为了更好地对聚类结果进行观察,可视化结果通过PCA将电子作业文档的特征向量转换成两个主成分进行显示。

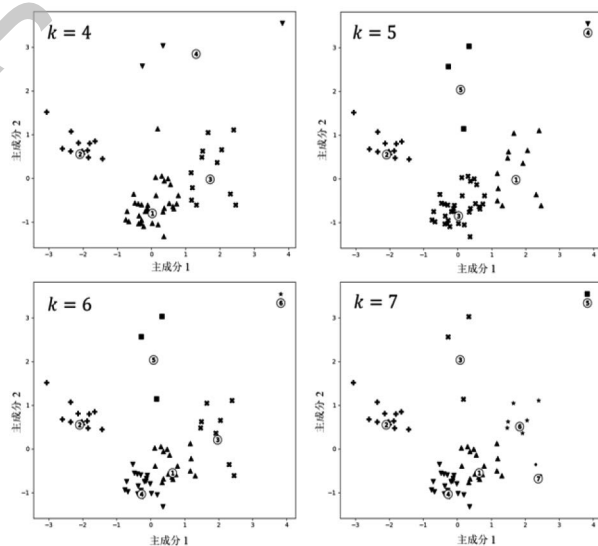


图2 电子作业文档的聚类结果

Fig.2 Clustering results of electronic assignments

轮廓分析(silhouette analysis)是对簇类数据一致性进行解释和验证的一种方法,为每个样本与所在簇类拟合性能提供了简洁的图形表示<sup>[7]</sup>。轮廓系数作为表明一个簇中每个点与相邻簇类内点的接近程度的度量,其值域为 $[-1, 1]$ ,该系数接近+1时表示当前样本远离相邻聚类,该系数为0则表示样本在两个相邻簇之间的决策边界上或非常接近,该系数为负数

时则表示当前样本可能被分配到错误的簇类。聚类中的每个样本 $i$ 都会有一个轮廓系数 $S_i$ ，定义如公式(2)：

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (2)$$

其中， $a_i$ 是当前样本 $i$ 与所在簇类中的其他节点之间距离的平均值，它反映了当前样本 $i$ 与所在簇类的拟合程度。 $b_i$ 则需要首先计算当前样本与其他每个簇类中所有节点的平均距离，然后再从这些平均距离中取最小值，对应最小平均距离的簇类成为样本 $i$ 的相邻簇类(neighboring cluster)，相邻簇类是除了样本 $i$ 所在簇类之外与该样本拟合得最好的簇类。由公式(2)可以推导出公式(3)：

$$S_i = \begin{cases} 1 - a_i/b_i, & a_i < b_i \\ 0, & a_i = b_i \\ b_i/a_i - 1, & a_i > b_i \end{cases} \quad (3)$$

由公式(3)可以看出，当 $S_i$ 接近1时，说明样本很适合分配到当前簇类。而当 $S_i$ 接近-1时，说明样本 $i$ 更应该被分配到它的相邻簇类。当 $S_i$ 等于0时，说明当前样本 $i$ 分配到当前簇类和相邻簇类是等价的，处在两个簇类的边界上。所有样本轮廓系数的平均值 $S$ 则作为当前聚类性能的评价指标，定义如公式(4)：

$$S = \frac{1}{n} \sum_{i=1}^n S_i \quad (4)$$

平均轮廓系数 $S$ 可以用于诸如k-means等聚类算法中簇类数的选择，通过寻找不同 $k$ 值下聚类结果的最大平均聚类系数，确定聚类的最优簇类数<sup>[8]</sup>。图3中给出了 $k=4, 5, 6, 7$ 时的每个电子作业文档与所在簇类的其他文档之间距离的平均值分布，以及当前 $k$ 值下的轮廓系数。

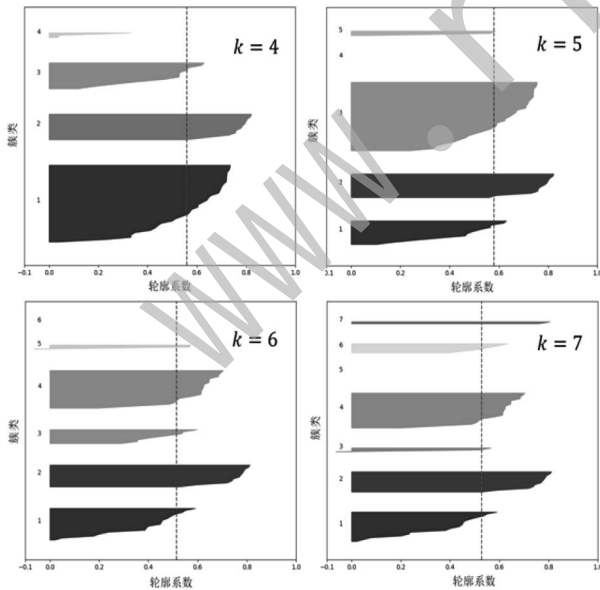


图3 电子作业文档聚类的轮廓系数

Fig.3 Clustering silhouette coefficients of electronic assignments

### 3.2 图论分析

文档相似度需要首先计算每份电子作业文档中每个词的

$tfidf$ 值，其定义如公式(5)所示。

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \cdot \log_2 \frac{N}{|\{d \in D: t \in d\}|} \quad (5)$$

其中， $f_{t,d}$ 是电子作业文档 $d$ 中词 $t$ 出现的次数， $\sum_{t' \in d} f_{t',d}$ 是电子作业文档 $d$ 中所有词出现的次数之和； $N$ 是所有电子作业文档集合 $D$ 中文档个数的总数，这里为61； $|\{d \in D: t \in d\}|$ 是电子作业文档库 $D$ 中包含词 $t$ 的文档个数的总数， $\log$ 在这里取底数2。

两份电子作业文档的相似度的计算基于余弦相似度，其定义如公式(6)所示。

$$similarity(d_1, d_2) = \frac{f(d_1) \cdot f(d_2)}{\|f(d_1)\| \|f(d_2)\|} = \frac{\sum_{i=1}^m f(d_1)_i f(d_2)_i}{\sqrt{\sum_{i=1}^m f(d_1)_i^2} \sqrt{\sum_{i=1}^m f(d_2)_i^2}} \quad (6)$$

其中， $f(d)$ 是向量 $[tfidf(t_1, d, D), tfidf(t_2, d, D), \dots, tfidf(t_m, d, D)]$ ， $t_i \in D$ ， $m$ 是 $D$ 的词典中所有词的数量总和。

图4是电子作业文档计算得到的相似度矩阵，其中越接近0代表电子作业文档差异性较大，1则代表两份文档完全相似，因此矩阵的对角线元素都为1。

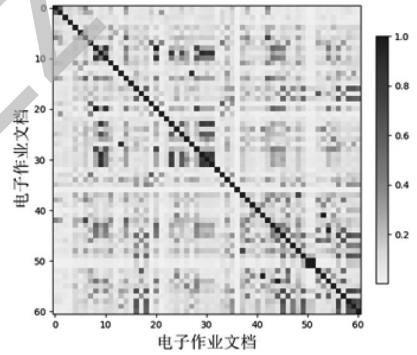


图4 电子作业文档的相似度矩阵

Fig.4 Similarity matrix of electronic assignments

在计算得到文档相似性矩阵后，可以通过基于图论的度量来找到该矩阵的特点，这里采用的是网络效率<sup>[9]</sup>。其中网络效率又分为全局效率和局部效率<sup>[10]</sup>，全局效率能够反应在整个矩阵中存在中心性较高的节点，其中全局效率的定义如公式(7)：

$$E_{global} = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{i,j}} \quad (7)$$

其中， $N$ 是整体网络 $G$ 中节点的总数， $d_{i,j}$ 是节点 $i$ 与 $j$ 之间的最短距离。局部效率首先需要计算每个节点的邻接节点所构成的子图 $G_i$ ，并计算子图 $G_i$ 的全局效率，然后求所有子图全局效率的平均值，其定义如公式(8)所示。

$$E_{local} = \frac{1}{N} \sum_{i \in G} E_{global}(G_i) \quad (8)$$

由于网络效率计算的对象必须是二值矩阵，而之前得到的文档相似矩阵不是二值矩阵。因此，将通过设置稀疏度的方式分别选取文档相似度前1%、2%、3%、...、100%的边，共产生100个二值矩阵并计算相应的网络效率。

## 4 实验结果(Experiment results)

如图5所示为电作业文档进行k-means聚类后,在不同k值上的轮廓系数。可以看出,当 $k=5$ 时轮廓系数最高,也就是说本次用于实验的电子作业文档被分为五个簇类时k-means聚类结果最优。 $k=3,4,5,7,8$ 是轮廓系数最高时的五个k值,这反映出电子作业文档在整体多样性上还在一个可接受范围内。

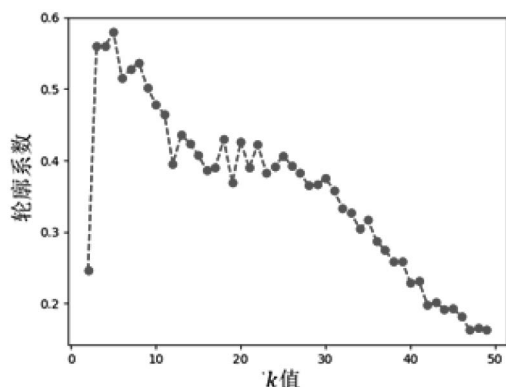


图5 聚类轮廓系数在k值上的分布

Fig.5 Distribution of clustering silhouette coefficients over k values

如图6所示为根据电子作业文档相似性矩阵在100个稀疏度上,分别计算全局效率和局部效率的结果。可以看出,与通过与随机网络产生的仿真数据相比,实际数据的全局效率和局部效率在主要稀疏度区间上都比仿真数据要低。这说明实验用的电子作业文档无论是整体的存在高中心度节点的情况,还是节点邻居网络存在高中心度节点的情况都不算严重,说明电子作业文档具有较高的多样性,质量较高。

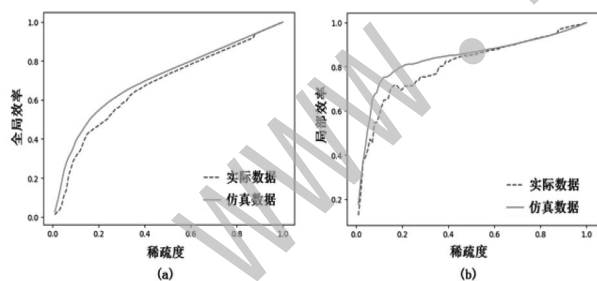


图6 电子作业文档相似矩阵的网络效率

Fig.6 Network efficiency of similarity matrix of electronic assignments

## 5 结论(Conclusion)

信息技术的发展使得教学在形式上发生了变化,机器学习的应用可以更好地对教学数据进行建模和利用。同论文撰写、项目申请等文档的电子化一样,作业的电子化也成为一种趋势,将电子作业作为机器学习的样本,不仅能够得到作业质量本身的信息,还能够挖掘出学生的学习行为信息。本

文正是从这点出发,通过对学生电子作业的挖掘,从中建立学习预警模型,通过聚类效果的评价及网络分析的度量对电子作业文档的整体质量作评估。

## 参考文献(References)

- [1] Kaplan A M,Haenlein M.Higher education and the digital revolution:About MOOCs,SPOCs,social media,and the Cookie Monster[J].Business Horizons,2016,59(4):441-450.
- [2] 秦新国.电子作业管理和作业抄袭检测技术研究[D].南京师范大学,2007:12-26.
- [3] 付兵,谢本贵.网络环境与机房环境下电子作业反抄袭策略[J].实验室研究与探索,2013,32(04):79-82,91.
- [4] 张星,刘帅.基于空间向量的电子作业相似度检测的设计与实现[J].福建电脑,2014,30(02):125-126.
- [5] 张洁,鱼先锋.基于直觉模糊聚类的电子作业抄袭检测研究[J].计算机与现代化,2014,(06):106-110.
- [6] Le,Quoc V.,Mikolov,Tomas.Distributed Representations of Sentences and Documents[C].International Conference on Machine Learning,2014,(4):1188-1196.
- [7] Rousseeuw P J.Silhouettes:A graphical aid to the interpretation and validation of cluster analysis[J].Journal of Computational & Applied Mathematics,1999,20(20):53-65.
- [8] Amorim R C D,Hennig C.Recovering the number of clusters in data sets with noise features using feature rescaling factors[M].Elsevier Science Inc,2015:10-13.
- [9] Latora,V.,M.Marchiori.Efficient Behavior of Small-World Networks[J].Physical Review Letters,2001,87(19):1-4.
- [10] Latora,V.,M.Marchiori.Economic small-world behavior in weighted networks[J].The European Physical Journal B-Condensed Matter and Complex Systems,2003,32(2):249-263.

## 作者简介:

张笑非(1980-),男,博士生,讲师.研究领域:物联网技术,人工智能.

段先华(1965-),男,博士,教授.研究领域:模式识别.

刘 镇(1961-),男,硕士,教授.研究领域:计算机系统结构,软件定义网络.

钱 萍(1978-),女,博士,讲师.研究领域:信息安全.