基于随机森林的儿童心理健康信息快速分类

艾映彤,唐 敏,李 艳,张大卫*
(昆明学院 学前与特殊教育学院,云南 昆明 650214)

摘 要:儿童心理健康信息档案管理混乱,导致档案调取困难,为此,提出基于随机森林算法的儿童心理健康信息快速分类方法。构建儿童心理健康信息表示模型,利用特定向量表示法叙述特定数据信息;结合词语和短语构建特征向量,确定选取特征;采用 TF-IDF 法检索健康信息,保留重要特征词;通过 TextRank 方法提取理健康信息关键点,生成有向有权图;计算相邻词语间余弦类似性生成数据信息集合;采用余弦类似性判断所有决策树之间类似性,通过随机森林法实现儿童心理健康信息快速分类。实验结果证明:所提方法可有效提升儿童心理健康信息分类的精度。

关键词:随机森林算法;儿童心理健康信息;非二进制方法;余弦类似性;加权投票方法;快速分类
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2022)0144-04

Rapid Classification of Children's Mental Health Information Based on Random Forest

AI Ying-tong, TANG Min, LI Yan, ZHANG Da-wei*
(School of Preschool and Special Education, Kunming College, Kunming 650214, China)

Abstract: The management of children's mental health information file is chaotic, which leads to the difficulty of file retrieval. Therefore, a fast classification method of children's mental health information based on random forest algorithm is proposed. The information representation model of children's mental health is constructed, and the specific vector representation is used to describe the specific data information. The feature vectors are constructed by combining words and phrases to determine the selected features. TF-IDF method was used to retrieve health information and retain important feature words. The key points of physical and health information are extracted by textrank method, and the directed weighted graph is generated. The cosine similarity between adjacent words is calculated to generate data information set; The cosine similarity is used to judge the similarity between all decision trees, and the random forest method is used to realize the rapid classification of children's mental health information. The experimental results show that the proposed method can effectively improve the accuracy of children's mental health information classification.

Key words: random forest algorithm; children's mental health information; non binary method; cosine similarity; weighted voting method; fast classification

0 引 言

社会快速发展,生活节奏加快,令成年人以及儿童或多或少产生各种压力以及负面情绪,在该背景下,压抑等不良情绪若不能获得有效地缓解,不断的累积会对身心健康、工作学习等造成极大影响。通常成年人会利用自己的方式发泄,以此获得缓解,以输入和输入情感来维持情绪平衡。社会快节奏环境下,自然输出负面情绪比较困难,尤其是儿童,正处于语言行为以及自主行为的发展期,比成人更加需要情感的表达,以及情感的发散。一旦发现儿童心理健康出现异常,这时需要通过就医方式解决。在就医过程中,会留存儿童心理相关数据信息,对这些信息进行分类,有助于儿童心理健康的治疗效果。为此,该领域研究者对其进行了很多研究,并取得了一定成果。

文献[1]提出使用机器学习方法对信息进行分类研究,针对人工训练特征提取操作难度会较高问题,为避免关键特征丢失情况,通过网络层次结合法,设计 CRNN 且添加 attention 的机制,以此建立 Text-CRNN +attention 模型来对文本进行分类。该方法虽然分类精度高,不过需要大量样本进行训练,分类速度达不到实际应用要求;文献[2]提出利用参数共享方式增强事件之间信息关联,先利用门控卷积神经网络来对语

收稿日期:2021-08-04
基金项目:云南省哲社基金项目(QN202015)
作者简介:艾映彤(1989-),女,博士,副教授,研究方向为儿童早期发展、心理健康;通讯作者:张大卫(1981-),男,硕士,副教授,研究方向为信息工程、数据驱动。

义信息进行学习,凭借记忆网络来编码输入获取语义信息,实现语义输入全连接层,利用 Softmax 函数完成分类。但该方法过于重视语音信息关联性,容易受参数影响,降低分类结果精度。

为此,该文提出一种基于随机森林算法的儿童心理健康信息快速分类方法,以词语和短语相结合作为特征向量,再利用 K 均值聚类算法,对信息聚类分析,通过随机森林算法的余弦类似性以及投票原理,实现快速分类。

1　儿童心理健康信息表示模型

为了令随机森林算法能够实现儿童心理健康信息分类,需要先设计信息表示模型。利用特定向量表示法,叙述某一个特定数据信息,同时只考虑心理健康的语义信息。将待处理信息设置为 d_j,文本特征合集利用 T 表示。实际分类过程中,首先选择心理健康的信息特征;然后利用向量叙述每个特征权重,具体公式为:

$$\vec{d}_j = < w_{1j}, w_{2j}, \cdots, w_{Tj} > \tag{1}$$

式中,w 代表权重,此向量内所有权重都代表相应特征所贡献心理健康语义信息含量。构建特征向量的过程中存在两个很重要的因素。

(1)特征选取。一般是以词语作为单位,实现特征选择,利用短语实现特征选取效果同样良好,为了达到更高标准,将两种方法相结合后再选择特征[3]。

(2)计算权重。现阶段计算方式包括两种,分别为二进制权重以及非二进制权重,其中非二进制法主要是利用 0 ~ 1 内 1 个小数代表权重,而二进制通过离散数字 1 以及 0 代表权重。即非二进制方法利用特征频率-反转文档的频率函数,可以得到具体公式为:

$$\mathrm{fidf}(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|\mathrm{Tr}|}{\#\mathrm{Tr}(t_k)} \tag{2}$$

式中,$\#(t_k, d_j)$ 代表特征项 t_k 处于信息 d_j 内出现的数目,$\#\mathrm{Tr}(t_k)$ 代表全部信息内所有特征项 t_k 出现过的信息个数和。当某一个特征出现在儿童心理健康信息内的个数越多,那么这些特征就可以更好表达此信息内容。若某一个特征在儿童心理健康信息内出现的数目越多,那么该特征的分辨能力越差[4]。

为了令权重值处于[0,1]之间,将全部信息权重向量具有同样的长度,此时余弦规范化处理,得到:

$$w_{kj} = \frac{\mathrm{tfidf}(t_s, d_j)}{\sqrt{\sum_{s=1}^{|T|} (\mathrm{tfidf}(t_s, d_j))^2}} \tag{3}$$

式中,t_s 代表余弦规范化处理时长。

由于儿童心理健康信息特征的维度较大,无法直接利用分类器完成。因此本文使用降维方式选取有效特征子集 T',其中原始特征集要大于特征子集维度,通过降维的方法对分类器过度拟合问题进行改善[5]。

降维过程中也可能会删除一部分有用的特征信息,其特征函数的计算过程如下:

设定儿童心理健康信息的信息增益 $\mathrm{IG}(t_k, c_i)$,即:

$$\mathrm{IG}(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log \frac{P(t, c)}{P(t) P(c)} \tag{4}$$

式中,信息增益率 $\mathrm{IG}(t_k, c_i)$ 代表在特征项 t_k 处于训练集的样本内所出现概率。IG 数值越高,那么对于分类预测所提供信息会越多,在设置一个阈值,如果 IG 值比阈值小,那么将其删除,从而降低特征空间的维数。

特征词条以及种类之间的互信息存在不同计算方法,通常会选择互信息量比较大的名称作为特征词。这是由于此类词语处于某一个种类内所出现的概率比较大,不过在其它种类内所出现概率比较小,即互信息量越大,种类以及名词同时出现的概率越大。此时,儿童心理健康信息的互信息 $\mathrm{MI}(t_k, c_i)$ 公式为:

$$\mathrm{MI}(t_k, c_i) = \log \frac{P(t_k, c_i)}{P(t_k) P(c_i)} \tag{5}$$

当计算得到的卡方信息数值越大,代表类别与特征词关联越大,即训练信息集内包括特征词的信息隶属于某一个种类概率就会越大,相反则越小[6]。

2　基于随机森林算法的信息快速分类

2.1　儿童心理健康信息特征最优选取

儿童心理健康信息文本快速分类中,其文本信息特征质量越高、个数越多越好,从而最后的分类结果会更好,因此,需要提取儿童心理健康信息特征的最优值。

(1)首先,利用 TF-IDF 法对儿童心理健康信息检索和挖掘数据。其中,TF 值代表处于文件 j 内第 i 个词语的重要性,具体公式为:

$$\mathrm{TF}_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{6}$$

式中,$n_{i,j}$ 代表第 i 词语处于文件 j 内所出现的频数,$\sum_k n_{k,j}$ 代表文件 j 内包括 k 个单纯所出现总频数。

$$\mathrm{IDF}_i = \log \frac{|D|}{|\{j: t_i \in d_j\}| + 1} \tag{7}$$

式中,$|D|$ 代表语料库内文件的总数量,t_i 代表需要检验第 i 个词,d_j 代表文件 j 所包括词语集合,$|\{j: t_i \in d_j\}|$ 代表包括 t_i 全部文件的频数。

根据上述计算,可对儿童心理健康词语出现的频率和处于文件集内出现文件频率低的词汇生成权重较

高的 TF-IDF。以此对常用词语进行过滤,从而留存重要词语[7]。

(2)再次,利用 PageRank 方法评判儿童心理健康信息的重要程度,然后利用有向无权图打分。在设置 V_j 代表心理健康信息的 j 节点,$\mathrm{In}(V_j)$ 代表 j 节点的集合,$\mathrm{Out}(V_j)$ 代表 j 所指向其余信息节点集合,$|\mathrm{In}(V_j)|$ 代表集合节点个数。得到:

$$S(V_i)=(1-d)+d\times\sum_{j\in \mathrm{In}(V_i)}\frac{S(V_j)}{|\mathrm{Out}(V_j)|} \tag{8}$$

式中,d 代表阻尼系数。

利用 TextRank 方法提取儿童心理健康信息的关键词语,会生成有向有权图。利用一个可变窗口描述文本信息清除停用词,从而认定在窗口中所有词语之间存在联系,且对相邻词语间余弦类似性计算,得到所有词语间权重:

$$\mathrm{WS}(V_i)=(1-d)+d\times\sum_{V_j\in \mathrm{In}(V_i)}\frac{W_{ji}}{\sum\limits_{V_k\in \mathrm{Out}(V_j)}W_{jk}}\times \mathrm{WS}(V_j) \tag{9}$$

式中,$\mathrm{WS}(V_i)$ 代表节点 V_i 获得的分数,W_{ji} 代表利用余弦类似度对节点 V_j 计算 V_i 的权重。

(3)借助 K 均值聚类算法将儿童心理健康信息进行聚类研究,完成最优值的选取。聚类前选择聚类中心,且利用起始聚类中心实现循环迭代,一直到聚类结果最后停止。不同起始聚类中心会致使聚类的结果发生不稳情况,生成多种局部最佳值。找到适当的 k 值,确定起始化质心。选择的数值良好,会提升方法的准确率。因此,利用 TF-IDF 法和 TextRank 方法所提取出的儿童心理健康信息集合,作为 K-mean 算法 k 值输入。定义各个样本与类和类间距离,将距离最近两类合并,在重建对新类计算和其它类之间距离,同时,按照最小距离方法归类。并且重复该过程,每一次减少一类。

此时儿童心理健康信息样本数据之间的距离利用欧式距离计算 d_{ij},得到:

$$d_{ij}=(x_i-x_j)^T(x_i-x_j)=\sqrt{\sum_{i,j=1}^{8}(x_i-x_j)^2} \tag{10}$$

式中,x_i 和 x_j 分别代表儿童心理健康信息样本数据。

利用 D_{KL} 代表第 K 个样本数据类 G_K 至第 L 个类 G_L 距离。计算 $G_1,G_2,\cdots$ 内所有儿童心理将康样本和类中心欧式距离,将其求平方和,获得结果被称之为离差的平方和。将距离最近两个类(G_K 以及 G_L)合并,令其变成新类 G_M,即新类 G_M 离差平方和具体公式为:

$$W_K=\sum_{c\in G_K}(X_C-\mu_K),(X_C-\mu_K) \tag{11}$$

$$W_L=\sum_{c\in G_L}(X_C-\mu_L),(X_C-\mu_L) \tag{12}$$

$$W_M=\sum_{c\in G_K}(X_C-\mu_M),(X_C-\mu_M) \tag{13}$$

式中,μ_K、μ_L、μ_M 分别代表类 G_K、G_L、G_M 中心,W_K、W_L、W_M 代表各自类内的样本分散程度度量。G_K 以及 G_L 间平方距离可表示为:

$$D_{KL}^2=W_M-W_K-W_L=\frac{\mathrm{num}_L\,\mathrm{num}_K}{\mathrm{num}_M}(\mu_K-\mu_L),(\mu_K-\mu_L) \tag{14}$$

依据聚类结果,计算出对应聚类中心,聚类中心会以某类内全部样本特征平均值代表,以此可以计算出最优特征值 Q,即:

$$Q=\frac{1}{\mathrm{num}_K}\sum_{X\in G_K,i=1}^{\mathrm{num}_K}X_i \tag{15}$$

2.2 基于随机森林的快速分类实现

根据上述确定儿童心理健康信息最优值,借助随机森林算法进行分类研究。将获得的决策树节点集合进行转换,变成词向量,采用余弦类似性判断所有决策树之间的类似性,通过随机森林法训练样本数据,利用决策树对节点信息集合进行精确率测试。儿童心理健康信息类似度计算公式为:

$$\mathrm{Similariy}=\frac{A\cdot B}{\|A\|\,\|B\|}=\frac{\sum\limits_{i=1}^{n}A_i\times B_i}{\sqrt{\sum\limits_{i=1}^{n}(A_i)^2}\times\sqrt{\sum\limits_{i=1}^{n}(B_i)^2}} \tag{16}$$

式中,A 和 B 代表两个节点的集合词向量,n 代表词向量维数。经过计算所有节点的集合类似性,选取最佳参数 d 以此比较,从而抽取类似性比较小节点集合,以此作为备选决策树集合[8]。

在利用训练集抽取子信息集合时,即可产生袋外数据,所以采用该部分的袋外数据,即可测试决策树的准确率,以此找出最佳决策树对儿童心理健康数据进行分类,具体公式为:

$$\mathrm{Weight}(i)=\frac{N^{\mathrm{cor}}(i)}{N} \tag{17}$$

式中,N 代表总体样本个数,即袋外样本个数,N^{cor} 代表正确分类个数,$\mathrm{Weight}(i)$ 代表决策树 i 正确率。

3 实验证明

3.1 实验数据

为了验证该方法对儿童心理健康信息是否能够快速精准分类,实验选取某医院儿童心理档案室内自闭症、口吃、恐惧症、过动症 4 种数据来作为输入信息,一共 12 475 篇文档,其中 6 233 篇数据信息作为训练集,6 242 篇数据信息作为测试集,将其作为基础实验数据。选取 1 000 个词语作为特征词。为了保证实验结

果的真实性,选择平均绝对误差 MAE 以及平方根误差 RMSE 作为验证指标,具体公式为:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{s=1}^{N}\sum_{i=1}^{10}\left|\vec{P}_r^{(i)}-\vec{r}_r^{(i)}\right|^2}{10*N}} \tag{18}$$

$$\mathrm{MAE}=\frac{\sum_{s=1}^{N}\sum_{i=1}^{10}\left|\vec{P}_r^{(i)}-\vec{r}_r^{(i)}\right|^2}{10*N} \tag{19}$$

式中,$\vec{P}_r$ 代表文中方法分类结果可靠的数据序列,$\vec{r}_r$ 代表实际情况中观测所获得的有效分类中可靠性数据序列。$\vec{P}_r^{(i)}$ 以及 $\vec{r}_r^{(i)}$ 分别代表处于第 i 时间点 $\vec{P}_r$ 以及 $\vec{r}_r$ 所表示的数值,N 代表分类的次数。MAE 以及 RMSE 结果是针对分类结果与理想值之间的误差,所以 MAE 与 RMSE 数值越小说明分类结果越精确。

3.2　分类结果精度对比

将文中的分类方法与 Text-CRNN+attention 架构下的信息分类方法、信息交互增强的分类方法进行对比,结果如图 1 所示。

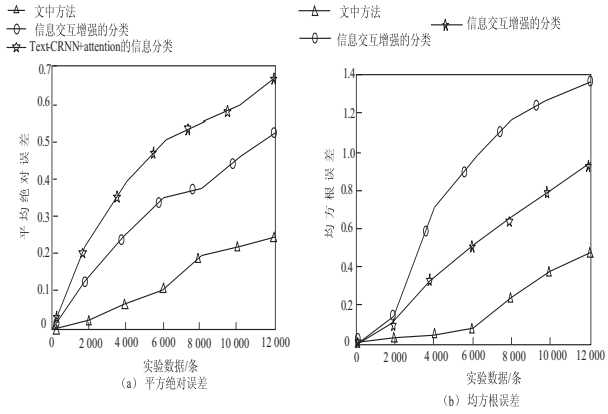

图 1　三种方法的分类误差对比结果

通过观察图 1 能够看出:在实验数据较少时,所有方法的分类结果准确率较高,随着实验数据的逐渐增加,所有方法分类准确率不断降低。其中,信息交互增强分类、Text-CRNN+attention 架构下的信息分类方法,误差曲线会随着数据量的增加呈现指数上升趋势,观察其分类记录,发现二者没有提取出有效信息结果,且容易在分类过程中受到二次噪声污染,降低使用者工作的效率。反观文中方法的分类结果准确度要高于对比方法,不过在达到 6 000 条数据以后,分类精度也会出现明显降低的趋势,但误差较小可忽略不计。

3.3　分类速度对比

在实际分类过程中,儿童心理健康的档案信息复杂,其中包括各种各样病情,并且数据量较多,所以需要设置时间阈值,在阈值之内说明分类速度满足实际需求,若超出阈值,则说明分类较慢不可使用。基于选择的实验数据,设置阈值为 20 分钟,总实验时间为 40 分钟。对比三种方法下分类速度,实验结果如图 2 所示。

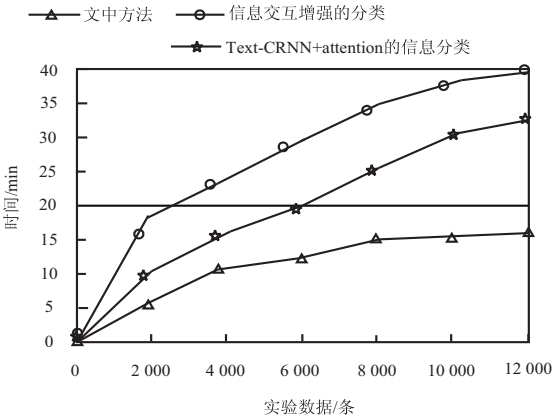

图 2　分类速度对比

通过观察图 2 能够看出:文中方法的分类速度最快,所用时间最少,且处于阈值以内。这是由于文中使用随机森林法,选择最佳决策树获取最能证明文本信息特征点,降低其他信息干扰,使多次训练后分类收敛速度保持平稳,直到实验结束仅出现非常小幅度的增长,满足儿童心理健康信息快速分类需求。

4　结束语

提出的基于随机森林算法的儿童心理健康信息快速分类方法,通过构建信息表示模型,提取信息特征,利用余弦类似性以及加权投票的方式,实现快速分类。该分类方法精度较高、速度较快,具有一定可行性。

参考文献:

[1] 卢　健,马成贤,杨腾飞,等. Text-CRNN+attention 架构下的多类别文本信息分类[J]. 计算机应用研究,2020,37(6):1693-1696.

[2] 周新宇,李培峰. 基于信息交互增强的事件时序关系分类方法[J]. 计算机科学,2020,47(11):244-249.

[3] 吴　超,李雅倩,张亚茹,等. 用于表示级特征融合与分类的相关熵融合极限学习机[J]. 电子与信息学报,2020,42(2):386-393.

[4] 高　楠,彭鼎原,傅俊英,等. 基于专利 IPC 分类与文本信息的前沿技术演进分析——以人工智能领域为例[J]. 情报理论与实践,2020,43(4):123-129.

[5] 杨　锐,陈　伟,何　涛,等. 融合主题信息的卷积神经网络文本分类方法研究[J]. 现代情报,2020,40(4):42-49.

[6] 杨春霞,吴佳君,李欣栩. 融合实体信息的循环神经网络文本分类模型[J]. 小型微型计算机系统,2020,41(12):2516-2521.

[7] 李　晟,周　超. 基于随机森林的钛加工表面质量评定研究[J]. 机械制造与自动化,2020,49(6):36-38.

[8] 段中兴,毕瀚元,张作伟. 基于 D-S 证据理论的不完整数据混合分类算法[J]. 信息与控制,2020,49(4):455-463.