基于机器学习算法的虚假新闻检测研究

柳晓翠[1]
LIU Xiaocui

摘 要 针对社交媒体虚假新闻检测问题，为实现快速有效的检测，提出了一种利用机器学习算法进行虚假新闻检测研究的方法。首先通过词频-逆向文件频率（term frequency–inverse document frequency，TF-IDF）算法提取文本特征向量，然后使用 K 折交叉验证法寻找支持向量机 (support vector machines，SVM) 模型的最优参数，最后利用已寻找的最优参数训练 SVM 模型，并对新闻数据集进行真假新闻分类，从而识别出虚假新闻。实验结果表明，与朴素贝叶斯和决策树算法相比，提出的方法在虚假新闻检测问题上表现出较好的评价指标，其中正确率、召回率和 $F1$ 值高于其他两种方法，ROC 曲线也优于其他两种算法。

关键词 机器学习；虚假新闻检测；TF-IDF；支持向量机；K 折交叉验证法

doi：10.3969/j.issn.1672-9528.2021.09.072

0 引言

随着互联网时代社交媒体的迅速发展，虚假新闻毫无节制地网络上快速传播，已经渗入到包括政治、文化、经济、生活等多个领域。虚假新闻的传播不仅误导公众，扰乱社会秩序，甚至容易引起社会的动荡。鉴于虚假新闻的影响和危害，近几年，社交媒体中的虚假信息检测研究引发了工业界和学术界的广泛关注。

早期的假新闻检测方法主要依赖于人工核查，但这种方法耗时耗力，效率低，无法满足当前海量虚假信息的检测需求。人工智能和大数据技术的蓬勃发展为自动检测虚假新闻提供了技术支撑。2011 年，CASTILLO 等人[1]提取 Twitter 上信息的文本特征、用户特征、主题特征、传播特征等特征，并对 SVM、决策树等机器学习方法在虚假新闻检测上的性能做了实验对比。2015 年，LIANG 等人[2]提出了基于用户行为的谣言事件检测特征，并对 Logistics 回归、SVM、朴素贝叶斯、决策树和 K 近邻五种机器学习方法做了实验结果对比。

本文运用机器学习算法进行虚假新闻检测的研究。利用 TF-IDF 算法提取新闻文本内容的特征向量。使用 K 折交叉验证法选取 SVM 的最优参数，并利用获取的最后参数对 SVM 分类器进行模型训练，使用测试集验证训练好的模型，获取评价指标。实验结果表明，本方法可以有效地识别虚假新闻，且具有较高的识别准确率。

1. 山东大学新闻传播学院 山东济南 250100

1 算法实现

本文研究的虚假新闻检测算法流程如图 1 所示。首先，对原始数据进行数据预处理，并将数据集划分为训练集和测试集两个互斥的子集。针对各个子集分别提取其文本特征，并使用交叉验证法寻找 SVM 模型的最优参数，以此最优参数构建的 SVM 分类训练模型验证测试集，获取验证结果，并进行结果指标评估。

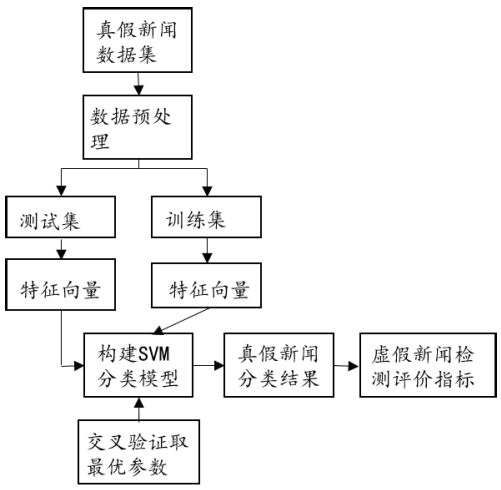

图 1 虚假新闻识别流程图

1.1 数据预处理

数据预处理是虚假新闻检测过程中的重要步骤之一，主要过程包括缺失数据补充、重复数据剔除、分词处理、去停用词等。

鉴于原始数据集中新闻文本内容是由中文完整的句子组成的，因此首先利用中文分词技术对新闻文本内容进行分词。

本文的研究使用了以 python 为基础的 jieba 分词工具，采用其“精确模式”完成分词过程。分词后，有些词汇对真假新闻文本不仅没有实际意义，反而因为其出现频率高，可能导致分类模型准确度降低。例如“的”“了”“一般”以及特殊标点符号等词汇，称之为停用词。因此，本文采用哈工大停用词对分词进行停用词处理。

1.2 提取文本特征向量

本文采用 TF-IDF 算法提取新闻文本的特征向量。TF-IDF 算法是提取特征词向量的重要算法之一，也是生成词向量化的主要技术之一。

TF-IDF 算法指的是如果某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或短语具有很好的分类区分能力，适合用来分类 [3]。简单地说，TF-IDF 可以反映出语料库中某篇文档中某个词的重要性。

TF-IDF 算法以统计的方式评估某个词语对某篇文档或语料库中其他文档的重要程度，以此来判断文档的特征词。其基本思想：如果某个词语在某篇文档中出现的频率很高，但是在语料库内的其他文档中出现的频率很低，则认定此词语在某种程度上可作为该文档的特征词，具备类别区分能力，可作为分类的依据。

TF-IDF 算法分为 TF（词频）算法和 IDF（逆文档频率）算法，其中 TF 算法表示特定词语的计数与文档中词语总数的比值，代表特定单词在文档中出现的频率。IDF 算法表示语料库中的文档总数与特定词语在语料库中出现的文档数比率的对数，即：

$$TF-IDF = TF \times IDF \tag{1}$$

某词语 i 在文档 j 中词频的计算公式为：

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{kj}} \tag{2}$$

式中：n_{ij} 为词语 i 在文档 j 中出现的次数，$\sum_k n_{kj}$ 是文档 j 中所有的词语出现次数的总和。

词语 i 在语料库中的逆文档频率 IDF 计算公式为：

$$IDF_i = \log \frac{|D|}{|D_i| + 1} \tag{3}$$

式中：$|D|$ 表示语料库中的文档总数，$|D_i|$ 表示包含词语 i 在语料库中出现的文档总数目，注意，这次采用 $|D_i|$+1 的目的是避免分母出现 0 而导致错误。

词语的 $TF\text{-}IDF$ 为：

$$TF-IDF = \frac{n_{i,j}}{\sum_k n_{kj}} \log \frac{|D|}{|D_i| + 1} \tag{4}$$

1.3 SVM 算法

本文提出虚假新闻检测视为二分类问题来处理，研究机器学习中的 SVM 方法对虚假新闻检测的效果。SVM 算法是由 Vapnik 等人于 20 世纪 90 年代初根据统计学理论创建的一种机器学习方法，具有坚实的数学理论基础，并遵循结构风险最小化原则，初始创建目的是解决二分类问题。SVM 算法如图 2 所示。

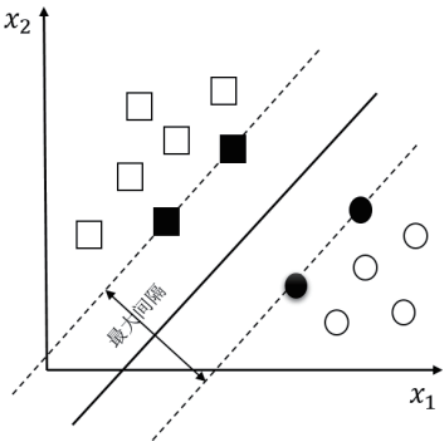

图 2 SVM 分类器示意图

SVM 算法的主要思想：寻找一个最优超平面作为分类的决策面，将两类数据进行区分，并满足两类数据之间的隔离边缘最大化。SVM 算法通过选择适当的核函数和惩罚系数来构造最优化求解问题，从而寻找到最优分类超平面 [4]。

假定训练数据集 D 为：

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \tag{5}$$

式中：x_i 为训练数据，y_i 为训练标签。

求解最优化超平面问题转换为通过选取核函数 $K(x_i, y_i)$ 和惩罚系数 C 进行最优解求解。

$$\left\{ \begin{array}{l} Q(a) = \max \sum_{j=1}^{n} a_j - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j a_i a_j K(x_i, x_j) \\ s.t. \sum_{i=1}^{n} y_i a_i = 0, \quad 0 \le a_i \le C, \ i = 1, 2, \ldots, n \end{array} \right\} \tag{6}$$

本文研究的虚假新闻检测方法采用高斯核函数来构建 SVM 分类模型，其表达式为：

$$K(x_j, x) = \exp\left(\frac{-\|x_i - x\|}{2\delta^2}\right) \tag{7}$$

1.4 K 折交叉验证

交叉验证法用于评估模型的预测性能，提高模型的学习能力，可以有效减少过拟合和欠拟合。交叉验证法的基本思想是将数据集划分为互斥的训练集和测试集，其中训练集用于训练分类模型，测试集用于测试训练得到的模型，并作为模型的评价指标。

K 折交叉验证法是交叉验证法常用的验证形式，基本思想：将数据集划分为 K 个互斥的子集，其中一个子集作为验证模型的数据集，剩余 K-1 个子集用来训练模型。重复交叉

验证 K 次，每个子集验证一次。以 K 次验证结果的平均值作为最终的评价指标。

K 值的选择根据实际情况进行调节。一般情况下，当原训练集较小时，K 值可以选择大一点，当原训练集较大时，K 值可以选择小一点。其中，10 折交叉验证是最常用的，图 3 展示了 10 折交叉验证示意图。

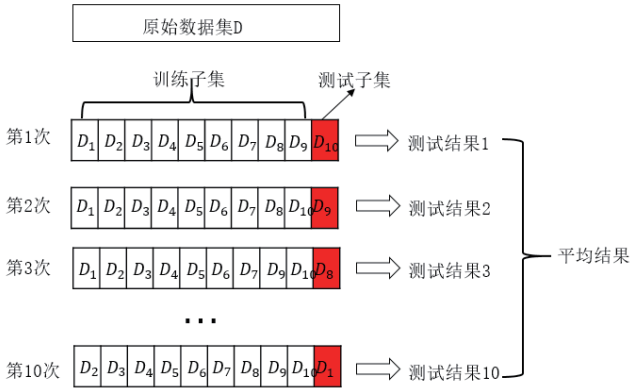

图 3　10 折交叉验证示意图

2　实验结果

本文实验的数据选取 [5] 论文提供的中文相关的真假新闻。为验证 SVM 分类模型在虚假新闻检测中的有效性，本文将原始新闻数据集随机划分为 75% 的训练集和 25% 的测试集。针对 75% 的训练集进行 10 折交叉验证寻找 SVM 模型中最优的惩罚系数 c 和高斯核函数参数。设置参数 c 和的取值范围为，按照图 3 的 10 折交叉验证方法在 2 的指数范围内进行离散化验证。以 10 次平均交叉验证的分类准确率为最终的评估结果，测试表明，当参数取值为：c=2，g=0.5 对应最优参数，其准确率为 0.904。

按照最优参数训练的 SVM 分类模型与其他分类算法在测试集上验证的结果如表 2 所示。

表 2　各分类算法评价指标

评价指标 / 模型	正确率	准确率	召回率	F1 值
SVM	0.92	0.87	0.97	0.92
朴素贝叶斯	0.88	0.89	0.83	0.86
决策树	0.78	0.79	0.68	0.73

表 2 结果表明，按最优参数训练的 SVM 分类模型在正确率、召回率和 $F1$ 值上的评价指标高于其他两种算法，准确略低于朴素贝叶斯算法。本文研究的机器学习算法对虚假新闻检测具有良好的检测效果。

图 4 展示了三种方法的 ROC 曲线，从图中可以看出 SVM 算法与 x 轴之间的面积最大，故性能最好。通过表 2 和图 4 以上评价指标综合比较，得出 SVM 算法的评价指标最优。

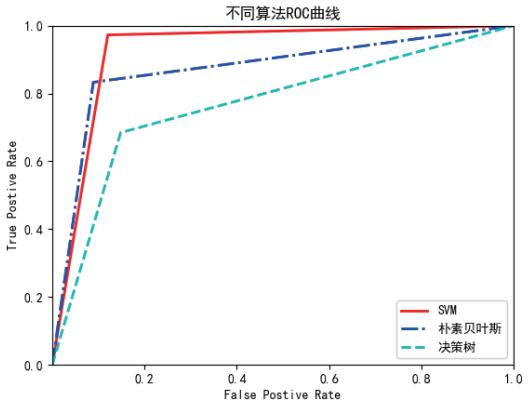

图 4　不同算法的 ROC 曲线

3　结束语

本文采用机器学习算法研究虚假新闻检测，并对比了 SVM、决策树、贝叶斯算法的检测结果。实验结果表明，SVM 在本研究中的性能指标优于其他算法，其准确率、精准率、召回率和 $F1$ 值都高于其他算法。在未来的工作中，将搜集更多的虚假新闻信息，构建虚假新闻数据库，并继续深入研究深度学习在虚假新闻检测中的应用。

参考文献：

[1] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter[C] // Proceedings of the 20th International Conference on World Wide Web. ACM, March 28 - April 1,2011,Hyderabad, India. New York:ACM, c2011: 675-684.

[2] LIANG G, HE W, XU C. Rumor identification Microblogging systems based on users' behavior[J].IEEE Transactions on Computational Social Systems, 2015, 2(3): 99-108

[3] 余本国 . 基于 Python 的大数据分析基础及实战 [M] . 北京：中国水利水电出版社 , 2018.

[4] 周志华 . 机器学习 [M] . 北京：清华大学出版社 , 2016.

[5] WANG Y, MA F, JIN Z, et al. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection[C]//The 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 19-23, 2018, London United Kingdom. New York: ACM,c2018: 849–857.

【作者简介】

柳晓翠（1987—），女，山东烟台人，山东大学新闻传播学院助理实验师，工学硕士，研究方向：机器学习、大数据分析。

（收稿日期：2021-07-26　修回日期：2021-08-07）