基于两阶段特征选择的医疗敏感文本分类

陈春玲*[1],姜慧敏[1],郭永安[2]
(1. 南京邮电大学 计算机学院、软件学院,江苏 南京 210023;
2. 南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:为完成对医疗数据的敏感性分类,采用文本分类技术从对医疗敏感数据的分类的角度对医疗信息隐私保护进行了研究。在传统的医疗文本分类基础上,提出基于LSI-TF-IDF两阶段特征选择的文本分类方法对医疗文本数据进行敏感性分类。分别采用基于TF-IDF的传统文本分类方法和基于LSI-TF-IDF的两阶段特征选择的文本分类方法对糖尿病文本数据进行敏感性分类,利用朴素贝叶斯、KNN、SVM三个分类器进行实验比较,采用准确率、召回率和F_1值作为评价标准。实验结果表明,基于LSI-TF-IDF两阶段特征选择的文本分类方法较之基于TF-IDF的传统文本分类方法在准确率、召回率和F_1值上均有所提升。证明了该方法在医疗文本数据的敏感性分类上具有更好的分类效果。

关键词:医疗数据;隐私保护;特征选择;敏感数据;文本分类
中图法分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2020)08-0129-05
doi:10.3969/j.issn.1673-629X.2020.08.022

Medical Sensitive Text Classification Based on Two-stage Feature Selection

CHEN Chun-ling*[1], JIANG Hui-min[1], GUO Yong-an[2]
(1. School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2. School of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In order to complete the sensitive classification of medical data, text classification technology is used to study the privacy protection of medical information from the perspective of classification of medical sensitive data. Based on the traditional medical text classification, a text classification method based on LSI-TF-IDF two-stage feature selection is proposed to classify medical text data. The experiment uses the traditional text classification method based on TF-IDF and the text classification method based on LSI-TF-IDF two-stage feature selection to classify the sensitivity of diabetes text data, with three types of naive Bayes, KNN and SVM. For the comparison of experiments, precision ratio, recall ratio and F_1 value are used as evaluation criteria. The experiment shows that the text classification method based on LSI-TF-IDF two-stage feature selection has improved in precision ratio, recall ratio and F_1 value compared with traditional text classification method based on TF-IDF. It is proved that the proposed method has better classification effect on the sensitivity classification of medical text data.

Key words: medical big data; privacy protection; feature selection; sensitive data; text classification

0 引 言

随着医疗信息化,对医疗大数据的深入分析研究,大量的医疗数据不仅是对医疗过程的记录,还被用于深入地进行数据挖掘和分析,从中总结出这些数据的隐藏价值,促进医学技术的发展,提高医疗检测系统的有效性[1]。但是,医疗大数据使用过程中出现的问题也不容小觑。对于医疗信息,一方面通过对医疗信息数据的挖掘,对病症的诊断、治疗、药物开发、临床试验、发现疾病等提供科学决策具有重要意义。但在另一方面,在数据对外发布使用之前,如果病人的数据被完全泄露,可能会侵犯到病人的隐私,甚至由于数据的泄露,个人数据信息被随意交易,给病人带来更加严重的影响。冯登国等人[2]对于大数据的隐私保护提出了六种方法,其中包括匿名、数据溯源、访问控制等方法

收稿日期:2019-10-15 **修回日期**:2020-02-26
基金项目:国家重点研发计划(2018YFC1314903)
作者简介:陈春玲(1961-),男,教授,研究生导师,通讯作者,研究方向为软件工程、分布式组件技术、网络信息安全及其应用;姜慧敏(1995-),女,硕士研究生,研究方向为信息安全与隐私保护。

类别。但是这些方法实际上都是对数据进行处理,对数据的路径追踪以及数据权限的访问,而缺乏对源数据的分类的考虑。国内现有的对医疗领域的隐私保护技术[3-4]主要是基于匿名化的医疗数据隐私保护、基于医疗数据加密的隐私保护、基于访问控制的医疗数据隐私保护以及医疗数据的分级保护。文中从对医疗敏感数据的分类入手,以提高后续敏感数据的处理效率为目标进行医疗数据的隐私保护。

传统的医疗文本分类方法,侧重于进行医疗数据的分类管理[5],对如何快速地从医疗数据中分类敏感数据与非敏感数据的研究较少,考虑选择合适的方法对医疗信息进行敏感性分类,对于医疗信息的隐私保护技术的提高具有重要意义[6]。在传统的文本分类方法中,通常只有一个阶段的特征选择,其中使用较为广泛的是基于 TF-IDF 的特征选择方法。TF-IDF 方法是一种利用文档中的术语频率赋予权重进行特征排名的特征选择方法,对于医疗敏感文本,由于其数据结构涵盖了结构化、半结构化和非结构化数据,仅仅考虑术语频率对于特征选择的效率和分类的效果是不够的。考虑到医疗敏感文本的特殊性,在特征选择的过程中,还要考虑特征的降维,以提高分类的准确率,因此需要对传统的医疗文本分类方法进行改进。文中提出的基于 LSI-TF-IDF 两阶段特征选择的医疗文本分类方法,在特征选择阶段通过连续的两个阶段的特征选取和特征降维,提高分类的准确率,从而达到对医疗敏感文本分类的高效性。

1 传统医疗文本分类

传统的医疗文本分类主要侧重于自由文本分类。文本分类的过程,一般包括四个步骤:文本预处理、特征选择、分类、评估[7]。

第一步中,在读取输入文本文档之后进行文本预处理。此时,文本文档被划分为特征,数据表示中的文本文档则被表示为向量空间,而其组件是该特征及其特征在该文本文档中每个特征的频率所占据的权重。而后对其进行删除非信息特征操作,包括消除停用词、处理分词、文本标记、词干还原等步骤。

第二步中,特征选择的主要作用是减小数据大小,提高预测精度,提取重要特征,轻松理解属性或变量,最终减少执行时间。特征选择的过程可以概括为:对预处理过的文本进行搜索生成特征子集,通过评估产生最好的子集,通过验证方法验证产生的子集是否是最佳子集,若是最佳子集,则达到停止标准;若不是最佳子集,则当达到最大迭代次数时会停止循环。根据特征选择的评估任务,可以将特征选择方法分为两类:基于过滤的方法和基于包装的方法,分类依据按照是否依赖分类器进行划分。具体的特征选择方法包括文档频率(DF)、Pearson 相关标准、相关系数、信息增益(IG)、互信息(MI)、$\chi 2$ 统计、期望交叉熵(CE)、文本证据权重(WET)、遗传算法(GA)等[8-10]。

在第三步中,分类器的功能是根据文本文档的内容将其合并为一个或多个预定义的类别[11]。传统文本分类方法来源于模式分类,可以分为三类[12]:第一类是基于统计方法,如朴素贝叶斯、支持向量机(support vector machine, SVM)、K-近邻(K-nearest neighbor, KNN)、Rocchio 等算法;第二类是基于连接的方法,如人工神经网络;第三类是基于规则的方法,如决策树、关联规则、粗糙集等。有很多算法用作分类器,但广泛使用的算法(分类器)[13-14]是决策树分类器、SVM 分类器、朴素贝叶斯分类器和 K-最近邻分类器。

第四步中,由于文本数据被分为测试集和数据集,通过测试集对训练集训练得出的分类器模型进行评估。对训练集和测试集的划分可以通过保持法和 K 折交叉验证法实现。目前对于文本的分类处理来说,评价的方法和指标包括召回率(recall ratio)、精确率(precision ratio)和 F_1 度量[15]。

传统医疗文本分类是基于 TF-IDF 特征选择方式仅考虑用文本中的词频进行特征选择,而未考虑到医疗敏感本文数据结构的复杂性,单一阶段的特征选择之后,分类效果并不够理想。不同于传统医疗文本分类的单一阶段特征选择方式,下一节提出的基于 LSI-TF-IDF 两阶段特征方法,通过连续利用现有特征降维和特征选择方法对文本进行特征选择,对医疗文本进行敏感性分类。

2 LSI-TF-IDF 两阶段特征选择的文本分类

基于 LSI-TF-IDF 两阶段特征选择方法包括两个阶段:第一阶段采用 LSI 方法,对原始文档进行特征降维;第二阶段采用 TF-IDF 方法从原始文档中获得权重数字表示,进行特征提取。

潜在语义索引(latent semantic indexing, LSI)[16]是一种流行的线性代数索引方法,通过词同现产生低维表示。LSI 的目的是在最小化全局重建误差的基础上,找到原始文档空间的最近似子空间。它基于奇异值分解(singular value decomposition, SVD)并将文档向量投影到近似子空间中,用余弦相似性准确地表示语义相似性[17]。给定一个术语文档矩阵 $\boldsymbol{X} = [x_1, x_2, \cdots, x_3] \in R^m$,假设 $\boldsymbol{X}$ 的等级为 R,对 $\boldsymbol{X}$ 进行奇异值分解,如式(1)所示:

$$\boldsymbol{X} = \boldsymbol{U} \Sigma \boldsymbol{V}^T \tag{1}$$

其中, $\Sigma = \mathrm{diag}(\delta_1, \delta_2, \cdots, \delta_r)$,且 $\partial_1 \geqslant \partial_2 \geqslant \cdots \geqslant \partial_r$ 是 X 的奇异值。$\boldsymbol{U} = [u_1, u_2, \cdots, u_r]$, $\boldsymbol{u}_i$ 是左奇异向量, $\boldsymbol{V} = [v_1, v_2, \cdots, v_r]$,其中 v_i 是右奇异向量。

词频-逆文本术语频率(term frequency - inverse document frequency,TF-IDF)是由 Sparck Jones[18] 提出的 IDF 演化而来的,TF-IDF 认为比起在少数文档中出现的术语,在许多文档中出现的术语,应该赋予更少的权重[19]。TF-IDF 术语加权的公式如下:

$$w_{i,j} = \mathrm{tf}_{i,j} * \log\left(\frac{N}{\mathrm{df}_i}\right) \tag{2}$$

其中, $w_{i,j}$ 表示文档 j 中术语 i 的权重, N 是集合中文档的数量, $\mathrm{tf}_{i,j}$ 是文档 j 中术语 i 的术语频率,由式(3)定义, df_i 是集合中术语 i 的文档频率,由式(4)定义。

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

其中, $n_{i,j}$ 表示术语 i 在文档 j 中出现的次数, $\sum_k n_{k,j}$ 表示文件 j 中所有词汇出现的次数总和。

$$\mathrm{df}_i = \log \frac{|D|}{|\{j: t_i \in d_i\}|} \tag{4}$$

其中, $|D|$ 表示文件总数, $|\{j: t_i \in d_i\}|$ 表示包含词语 t_i 的文件数目(即 $n_{i,j} \neq 0$ 的文件数目)。

LSI-TF-IDF 两阶段特征选择算法如下所示:

Algorithm:LSI-TF-IDF 算法。

1. 输入:经过文本预处理后的 dataset,存放于文档列表 doc_list 中

2. 输出:返回特征选择词典 dict_feature_select

3. 根据式(1)计算文档向量矩阵 $\boldsymbol{X}$ 和查询矩阵 $\boldsymbol{Q}$

4. 对矩阵 $\boldsymbol{X}$ 进行奇异值分解,得到左奇异值 $\boldsymbol{U}$ 和右奇异值 $\boldsymbol{V}$

5. 计算 query-document 的余弦相似度

6. 返回相似度最高的语句 sims

7. sims 切片存放于词列表 list_words 中

8. 进行词频统计

9. 根据式(3)计算每个词的 TF 值

10. 根据式(4)计算每个词的 IDF 值

11. 根据式(2)计算每个词的 TF * IDF 术语加权

12. 对字典值由大到小排序

在两阶段的特征选择中,首先采用 LSI 方法形成文档矩阵,采用余弦相似性评估文档矩阵,进行数据降维。而后利用 TF-IDF 方法,对降维后的文档数据中术语频率进行排序,从而完成特征提取。完成特征选择之后,紧接着就是对特征样本进行分类,最后在分类出的敏感样本和普通样本结果上,对其进行评估。因此,LSI-TF-IDF 两阶段特征选择算法用于医疗敏感文本分类的流程包括以下五个步骤:文本预处理、LSI 特征降维、TF-IDF 特征选择、分类、评估。

3　仿真与实验

3.1　数据集和实验方法

实验数据集使用 736 份糖尿病文本病历样本,对其进行处理,将它分为敏感样本和普通样本。在 Pycharm 平台上使用,并使用 python 语言进行设计,中文分词工具是 Jieba。在分类器训练部分,分别选择了朴素贝叶斯、KNN 和 SVM 作为文中的分类算法。在特征选择阶段,对基于 TF-IDF 方法和基于 LSI-TF-IDF 两阶段特征选择方法进行比较。对于分类结果评估,采用 10 倍交叉验证法对数据集进行评估。

3.2　评价标准

对于分类结果的评价标准可以通过精确率、召回率和 F_1 值进行评估。

召回率也称为查全率。在文本分类中,正确识别出属于 C 类的文本数与测试集中实际存在的属于 C 类的文本总数的比值为分类召回率,公式如下:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

其中,TP 表示由分类器正确计算将属于 C 类的文本判定属于 C 类;FN 表示由分类器错误地将属于 C 类的文本判定为属于其他类。

精确率也称为查准率。正确识别出的属于 C 类的文本数与识别出的属于 C 类的文本数二的比值,为分类精确率,公式如下:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

其中,TP 表示由分类器正确计算将属于 C 类的文本判定属于 C 类;FP 表示由分类器错误地将应属于其他类的文本判定为属于 C 类。

F_1 度量是基于精确率和召回率的调和平均,定义如下:

$$\mathrm{F}_1 = \frac{2 * P * R}{P + R} \tag{7}$$

其中, P 为精确率, R 为召回率。

3.3　实验结果

实验通过对糖尿病文本数据进行敏感样本分类,基于 TF-IDF 的传统文本分类方法的实验结果以及评价标准比较分别如表 1 和图 1 所示,基于 LSI-TF-IDF 的两阶段文本分类方法的实验结果以及评价标准比较分别如表 2 和图 2 所示。

表 1　基于 TF-IDF 的分类结果

类别	样本总数	朴素贝叶斯	KNN	SVM
敏感样本	370	313	296	327
普通样本	366	328	304	331

表 2 基于 LSI-TF-IDF 的分类结果

类别	样本总数	朴素贝叶斯	KNN	SVM
敏感样本	370	331	324	345
普通样本	366	330	318	339

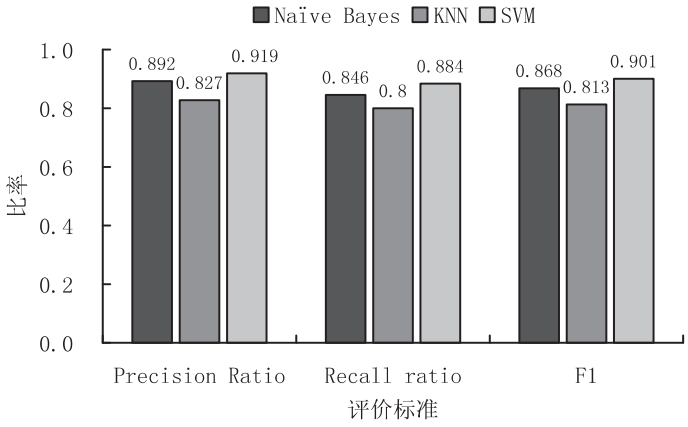

图 1 基于 TF-IDF 的医疗敏感文本分类

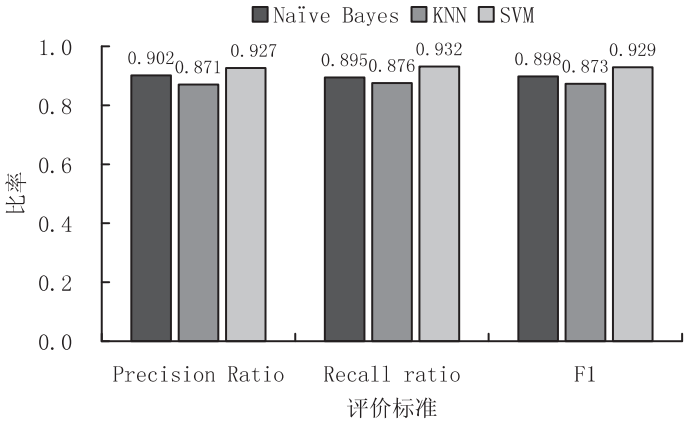

图 2 基于 LSI-TF-IDF 的医疗敏感文本分类

3.4 实验分析

由表 1 和表 2 对比显示,在经过连续两阶段的特征降维和特征选择后,正确分类出的敏感样本数目有所增加,对于朴素贝叶斯分类器,正确分类的敏感样本数由 313 份增加到 331 份;对于 KNN 分类器,正确分类的敏感样本数由 296 份增加到 324 份;对于 SVM 分类器,正确分类的敏感样本数由 327 份增加到 345 份。

由图 1 和图 2 对比显示,对于朴素贝叶斯分类,基于 LSI-TF-IDF 两阶段特征选择算法的分类精确率相比于单一的基于 TF-IDF 特征选择的分类精确率由 89.2% 上升到 90.2%,召回率也由 84.6% 上升到 89.5%;对于 KNN 分类,基于 LSI-TF-IDF 两阶段特征选择算法的分类精确率相比于单一的基于 TF-IDF 特征选择的分类精确率由 82.7% 上升到 87.1%,召回率也由 80% 上升到 87.6%;对于 SVM 分类,基于 LSI-TF-IDF 两阶段特征选择算法的分类精确率相比于单一的基于 TF-IDF 特征选择的分类精确率由 91.9% 上升至 92.7%,召回率由 88.4% 上升至 93.2%。可以看出无论是精确率、召回率还是 F_1 值,基于两阶段特征选择的医疗敏感文本分类都比传统医疗敏感文本分类有所提高。

因此,根据对实验结果的对比分析,对于医疗文本的敏感性分类而言,基于 LSI-TF-IDF 两阶段特征选择的文本分类方法比传统的基于 TF-IDF 文本分类方法取得了更好的分类效果。

4 结束语

从对敏感数据的分类角度切入隐私保护,针对传统的文本分类方法用于医疗敏感数据的分类准确性不足的问题,提出了基于 LSI-TF-IDF 两阶段特征选择的医疗敏感文本分类方法,通过连续两阶段的特征降维和特征提取,提高分类的准确性,解决了传统分类方法的不足。以 TF-IDF 特征选择为例,通过实验对传统的基于 TF-IDF 医疗文本分类方法和基于两阶段文本分类的医疗敏感文本分类方法进行比较。实验证明,基于 LSI-TF-IDF 两阶段特征选择的文本分类方法对于医疗敏感文本分类具有更好的效果。但是,由

于医疗数据量大,文本数据包括半结构化数据和非结构化数据,分类器的选择不同会导致分类速度和分类结果的准确性有差别,如何针对医疗数据的特点选择性能最好的分类器,需要具体的进一步研究。

参考文献:

[1] 许培海,黄匡时. 我国健康医疗大数据的现状、问题及对策[J]. 中国数字医学,2017,12(5):24-26.

[2] 冯登国,张　敏,李　昊. 大数据安全与隐私保护[J]. 计算机学报,2014,37(1):246-258.

[3] 史婷瑶,马金刚,曹　慧,等. 医疗大数据隐私保护技术的研究进展[J]. 中国医疗设备,2019,34(5):163-166.

[4] 王天屹,刘爱萍. 大数据环境下医疗数据隐私保护对策研究[J]. 信息技术与网络安全,2019,38(8):28-32.

[5] 许　杰. 基于机器学习的医疗健康分类方法研究[D]. 郑州:郑州大学,2018.

[6] RAJPUT K,CHETTY G,DAVEY R. PHIs (protected health information) identification from free text clinical records based on machine learning[C]//2017 IEEE symposium series on computational intelligence (SSCI). Honolulu:IEEE,2017:1-9.

[7] MANANA K,MAGDA T,MAIA A. Natural language processing based instrument for classification of free text medical records[J]. Biomed Research International, 2016, 2016: 8313454.

[8] NAIDU K,DHENGE A,WANKHADE K. Feature selection algorithm for improving the performance of classification:a survey[C]//2014 fourth international conference on communication systems and network technologies. Bhopal:IEEE,2014:468-471.

[9] ONAN A. Ensemble learning based feature selection with an application to text classification[C]//2018 26th signal processing and communications applications conference (SIU). Izmir:[s. n.],2018:1-4.

[10] XU G X,YU Z H,QI Q. Efficient sensitive information classification and topic tracking based on tibetan web pages[J]. IEEE Access,2018,6:55643-55652.

[11] BARAA S,NAZLIA O,ZEYAD S. An automated arabic text categorization based on the frequency ratio accumulation[J]. International Arab Journal of Information Technology,2014,11(2):213-221.

[12] MIAO F,ZHANG P,JIN L,et al. Chinese news text classification based on machine learning algorithm[C]//2018 10th international conference on intelligent human-machine systems and cybernetics (IHMSC). Hangzhou:[s. n.],2018:48-51.

[13] DU L,XIA C,DENG Z,et al. A machine learning based approach to identify protected health information in Chinese clinical text[J]. International Journal of Medical Informatics, 2018,116:24-32.

[14] 凤丽洲. 文本分类关键技术及应用研究[D]. 长春:吉林大学,2015.

[15] BRITTON K E,BRITTON-COLONNESE J D. Privacy and security issues surrounding the protection of data generated by continuous glucose monitors[J]. Journal of Diabetes Science and Technology,2017,11(2):216-219.

[16] STORMS S,SPEELMAN D,GEERAERTS D,et al. Within-concept similarities in a taxonomy:a corpus linguistic approach[J]. Language and Cognition,2015,7(2):194-218.

[17] ZHANG W,YOSHIDA T,TANG X. A comparative study of TF-IDF,LSI and multi-words for text classification[J]. Expert Systems with Applications,2011,38(3):2758-2765.

[18] JONES K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1972,28(1):11-21.

[19] TRSTENJAK B,MIKAC S,DONKO D. KNN with TF-IDF based framework for text categorization[J]. Procedia Engineering,2014,69:1356-1364.