·论 著·

文本分析联合支持向量机的肿瘤ICD-O-3病理形态学自动分类效果评价

潘劲，龚巍巍，费方荣，王蒙，周晓燕，胡如英，钟节鸣

浙江省疾病预防控制中心慢性非传染性疾病防制所，浙江 杭州 310051

摘要：目的 评价文本分析联合支持向量机（SVM）对肿瘤ICD-O-3病理形态学自动分类的准确性，为汉语环境的肿瘤分类编码研究提供参考。**方法** 通过浙江省慢性病监测信息管理系统收集2017—2019年浙江省户籍居民肿瘤报告卡，根据ICD-O-3编码，对病理学文本提取关键词，采用SVM进行自动化分类；并与16名有2年以上肿瘤编码经验的专业技术人员分类结果比较，计算准确率、召回率及两者的调和平均数（*F*值）评估分类效果。**结果** 纳入2017—2019年浙江省肿瘤报告卡83 082例，17个形态学分类，以腺癌、鳞状和移行细胞癌为主，52 877例占63.65%。通过文本分析筛选出1 090个关键词，准确率为77.20%，召回率为96.27%，*F*值为85.69。**结论** 采用文本分析联合SVM可提高肿瘤ICD-O-3病理形态学自动分类效率，但准确性有待进一步提升。

关键词：肿瘤；病理学；文本分析；支持向量机；自动分类

中图分类号：R181.2 **文献标识码：**A **文章编号：**2096-5087（2021）03-0255-05

Automated classification of ICD-O-3 morphology code from pathology reports using text-mining and support vector machine

PAN Jin, GONG Weiwei, FEI Fangrong, WANG Meng, ZHOU Xiaoyan, HU Ruying, ZHONG Jieming

Department of Non-communicable Disease Control and Prevention, Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, Zhejiang 310051, China

Abstract: Objective To evaluate the accuracy of automated classification of ICD-O-3 morphology code from pathology reports by text-mining and support vector machine (SVM), in order to provide basis for automated tumor coding in Chinese. **Methods** The tumor report cards of Zhejiang residents from 2017 to 2019 were collected from Chronic Disease Surveillance Information Management System of Zhejiang Province. According to ICD-O-3, the keywords of the pathology reports were extracted, and SVM was used for automatic classification. The classification results were compared with those of 16 professionals with more than two years of experience in tumor coding, and the accuracy rate, recall rate and *F*-score were calculated for effect evaluation. **Results** Totally 83 082 cases from 2017 to 2019 were included and were categorized into 17 morphological classifications, with 52 877 (63.65%) cases of adenocarcinoma, squamous carcinoma and transitional cell carcinoma. A total of 1 090 keywords were enrolled into main corpus. The total *F*-score, accuracy rate and recall rate are 85.69, 77.20% and 96.27%, respectively. **Conclusion** Text-mining combined with SVM can improve the efficiency of ICD-O-3 morphology coding; however, the accuracy needs to be further improved.

Keywords: neoplasm; pathology; text-mining; support vector machine; automated classification

肿瘤是影响居民健康的重要公共卫生问题。2015年全球肿瘤新发1 750万，死亡870万，我国肿瘤发病和死亡分别占21.8%和26.9%[1]。肿瘤登记有助于了解肿瘤的流行特点、危险因素及生存情况，为有效防治肿瘤提供依据。肿瘤登记包括资料收集、编码和利用3个主要环节，其中肿瘤编码最关键且技术难度大。同时国际疾病分类编码不断更新，如果缺乏对编码人员的及时培训，可能影响编码的准确性[2]。采用文本分析自动识别的方式实现肿瘤病理学编码，可以大大提高肿瘤登记的准确性和工作效率。文本分析技术是自然语言处理（nature language process，NLP）

DOI：10.19485/j.cnki.issn2096-5087.2021.03.009

基金项目：浙江省医药卫生科技计划（2018PY007，2019KY355）

作者简介：潘劲，硕士，主管医师，主要从事慢性病流行病学与监测信息化工作

通信作者：钟节鸣，E-mail：jmzhong@cdc.zj.cn

的一种分支技术，应用于医疗信息挖掘、疾病风险预测、疾病筛查与诊断、心理治疗研究、医疗卫生政策和医疗卫生资源评价等领域[3]。国外学者采用 NLP 对肿瘤登记资料、电子病案资料等非结构化文本进行分词，并在此基础上通过朴素贝叶斯、支持向量机（the support vector machine，SVM）和卷积神经网络人工智能分类算法对自动化编码进行比较，SVM 分类效果优于朴素贝叶斯，操作难度低于卷积神经网络，效率高，结构简单，易于实现[4-6]。目前，肿瘤病理学文本分析与自动分类的研究主要针对英语、法语和葡萄牙语等印欧语系，尚未有汉语相关报道。本研究在汉语环境下，采用文本分析和 SVM 技术对肿瘤进行自动编码，参照《国际疾病分类肿瘤学分册第 3 版》（ICD-O-3）对编码质量进行评价，为人工智能技术在肿瘤病理学自动编码领域的应用提供依据。

1　资料与方法

1.1　资料来源　2017—2019 年浙江省户籍居民肿瘤报告卡来源于浙江省慢性病监测信息管理系统。纳入肿瘤病理学类型、ICD-M 编码、ICD-O 等字段的肿瘤报告卡，排除病理学类型“无”“不详”等不合格的报告卡、继发转移性肿瘤（淋巴结或其他器官转移）报告卡、回顾性报告卡（报告日期与确诊日期相差 30 d 以上）和重复报告卡。

1.2　方法

1.2.1　分类标签标注　邀请 16 名具有 2 年以上肿瘤编码经验的专业技术人员根据《中国肿瘤登记工作指导手册》、国际癌症研究机构（IARC）和国际癌症登记协会（IACR）要求，根据 ICD-O-3 标准对肿瘤分类编码进行审核与修订[7-9]。ICD-O-3 由解剖部位编码与形态学编码 2 部分组成：（1）解剖部位编码，多部位原发肿瘤编码分类（IARC），将 330 类完整编码分为 54 大类，其中特定形态学的肿瘤（Kaposi 肉瘤及血液系统肿瘤）被单独分为一类。例如 C50.2 的完整解剖部位编码为乳腺内上象限，IARC 分类为乳腺。（2）形态学编码，根据 BERG[10] 的形态学分类方法，将 553 类完整形态学编码分为 17 类。例如 8 500/31 完整形态学编码为腺癌。见图 1。

图 1　ICD-O-3 代码示例

1.2.2　训练集和测试集构建　构建训练集和测试集评价自动编码效果。按照 3∶1 的比例将肿瘤报告卡随机分为训练集和测试集，并按照 ICD-O-3 进行分类。其中肥大细胞肿瘤报告卡只有 2 例，无法满足分类要求，所以不参与构建。

1.2.3　文本分析　清除肿瘤报告卡文本中的空格、换行符，将所有标点符号替换为空格，对处理完的文本进行分词，按照预先设定的停用词表去除文本中的“可见”“检查”“手术”等无用词，形成病理学语料库。采用向量空间模型（vector space model，VSM）[11] 对分词后的文本进行转换与提取：（1）选定特征集与降维，将病理学语料库进行汇总归类，按照 16 名专业技术人员的意见进行筛选归纳，若有至少 12 人（3/4）意见达成一致，则将该关键词剔除，最终获得病理学特征集（关键词库）。（2）计算特征权重，采用 TF-IDF 法[12]，公式为 $TF\text{-}IDF(F_k, d_k) = (F_k, d_k) \log T/T(F_k) + 1$。公式前半部分 (F_k, d_k) 为词频，后半部分为逆词频的 log 值，词频即某个词在文本中出现的次数，T 表示语料库中的个案文本条数，$T(F_k)$ 表示语料库中包含特定词的个案文本条数。

1.2.4　分类算法　采用 SVM[13] 进行分类计算。SVM 是通过某种事先选择的非线性映射将输入向量 x 映射到一个高维特征空间 Z，并在这个空间中构建最优超平面，属于有监督的机器学习算法。

1.2.5　自动分类效果评价　计算准确率、召回率（查全率）和 F 值，评价肿瘤 ICD-O-3 自动分类效果，准确率、召回率和 F 值越大表示效果越好。以 16 名专业技术人员的判断结果为金标准，准确率指该分类方法相对于金标准的准确水平，即特定分类中，分类方法与专业技术人员评判结果一致的数量占该分类总数的比例。召回率，即特定分类中分类方法与专业技术人员评判结果一致的数量占专业技术人员判定分类总数的比例。F 值是准确率和召回率的调和平均数[14]，综合评判该方法的可靠性，F 值=（2×准确率×召回率）/（准确率+召回率）。

1.3　统计分析　采用 Python 2.7.18 软件建立数据库并统计分析，采用 Jieba Package 软件进行病理报告等文本的处理及分词，采用 Sci-kit-learn Package 软件处理分类算法。

2　结　果

2.1　肿瘤形态学分类及训练、测试集构建　共收集 2017—2019 年浙江省肿瘤报告卡 570 683 例，纳入符合要求的报告卡 83 082 例。以腺癌、鳞状和移行

细胞癌为主，52 877 例占 63.65%；基底细胞癌、间皮瘤、T 细胞和 NK 细胞肿瘤、霍奇金淋巴瘤、肥大细胞肿瘤、组织细胞和附属淋巴样细胞肿瘤以及 Kaposi 肉瘤构成比均不到 1%。见表 1。构建训练集 62 304 例，测试集 20 776 例。

表 1 肿瘤形态学分类及训练、测试集构建

分类	编码	报告卡例数	构成比（%）	训练集例数	测试集例数
鳞状和移行细胞癌	8051～8084, 8120～8131	14 527	17.49	10 895	3 632
基底细胞癌	8090～8110	693	0.83	519	174
腺癌	814～8149, 8160～8162, 8190～8221, 8260～8337, 8350～8551, 8570～8576, 8940～8941	38 350	46.16	28 762	9 588
其他特指类型癌	8030～8046, 8150～8157, 8170～8180, 8230～8255, 8340～8347, 8560～8562, 8580～8671	9 408	11.32	7 056	2 352
非特指类型癌NOS	8010～8015, 8020～8022, 8050	3 022	3.64	2 266	756
肉瘤和软组织肿瘤	8680～8713, 8800～8921, 8990～8991, 9040～9044, 9120～9125, 9130～9136, 9141～9252, 9370～9373, 9540～9582	1 989	2.39	1 491	498
间皮瘤	9050～9055	115	0.14	86	29
白血病	9840, 9861～9931, 9945～9946, 9950, 9961～9964, 9980～9987	935	1.13	701	234
B细胞肿瘤	9670～9699, 9728, 9731～9734, 9761～9767, 9769, 9823～9826, 9833, 9836, 9940	2 385	2.87	1 788	597
T细胞和NK细胞肿瘤	9700～9719, 9729, 9768, 9827～9831, 9834, 9837, 9948	585	0.70	438	147
霍奇金淋巴瘤	9650～9667	212	0.26	159	53
肥大细胞肿瘤	9740～9742	2	<0.01	0	0
组织细胞和附属淋巴样细胞肿瘤	9750～9758	50	0.06	37	13
非特指类型	9590～9591, 9596, 9727, 9760, 9800～9801, 9805, 9820, 9832, 9835, 9860, 9960, 9970, 9975, 9989	1 506	1.81	1 129	377
Kaposi肉瘤	9140	16	0.02	12	4
其他特指类型的肿瘤	8720～8790, 8930～8936, 8950～8983, 9000～9030, 9060～9110, 9260～9365, 9380～9539	3 839	4.62	2 879	960
非特指类型的肿瘤	8000～8005	5 448	6.56	4 086	1 362

注：肥大细胞肿瘤报告卡不满足分类要求，不参与构建训练和测试集。

2.2 肿瘤文本提取 按照预处理的方法整理病理语料库，通过分词提取 6 324 个词，16 名专家在审核编码的同时，结合编码规则与个人经验对语料库进行提取和筛选，其中 1 090 个关键词得到至少 12 名专家的认可，进入关键词库，17 个分类的关键词分布见表 2。

2.3 SVM 分类结果评价 采用 SVM 法的分类效果进行评价，参与测试 20 776 例，总体 *F* 值为 85.69，准确率为 77.20%，召回率为 96.27%。其中鳞状和移行细胞癌、基底细胞癌、腺癌的 *F* 值均高于 95，间皮瘤与白血病的 *F* 值均低于 50。鳞状和移行细胞癌、基底细胞癌、腺癌、T 细胞和 NK 细胞肿瘤、霍奇金淋巴瘤、Kaposi 肉瘤以及非特指类型的准确率均高于 90%，其中 Kaposi 肉瘤达到 100%，鳞状和移行细胞癌、基底细胞癌准确率均高于 95%，间皮瘤、白血病、其他特指类型的肿瘤、非特指类型的肿瘤准确率均低于 50%。各类肿瘤召回率均高于 70%，其中鳞状和移行细胞癌、基底细胞癌、腺癌、肉瘤和软组织肿瘤、白血病、B 细胞肿瘤、其他特指类型的肿瘤召回率均高于 95%。见表 2。

3 讨 论

研究结果显示，文本分析联合 SVM 对肿瘤 ICD-O-3 病理形态学自动分类准确率为 77.20%，召回率为 96.27%，总体 *F* 值为 85.69，提示分类效果较好。17 类肿瘤分类效果不同，可能由于样本量分布不均，鳞状和移行细胞癌、基底细胞癌、腺癌的 *F*

表 2　肿瘤形态学 SVM 分类结果

分类	参与测试例数	构成比（%）	准确率（%）	召回率（%）	*F* 值	关键词数
鳞状和移行细胞癌	3 632	17.48	98.19	95.76	96.96	91
基底细胞癌	174	0.84	96.67	100.00	98.31	33
腺癌	9 588	46.15	92.66	99.37	95.90	152
其他特指类型癌	2 352	11.32	73.52	94.30	82.62	124
非特指类型癌NOS	756	3.64	80.25	75.79	77.96	15
肉瘤和软组织肿瘤	498	2.40	60.23	95.78	73.95	154
间皮瘤	29	0.14	29.17	72.41	41.58	6
白血病	234	1.13	30.88	97.01	46.85	100
B细胞肿瘤	597	2.87	84.51	95.98	89.88	69
T细胞和NK细胞肿瘤	147	0.71	94.07	86.39	90.07	40
霍奇金淋巴瘤	53	0.26	94.12	90.57	92.31	8
肥大细胞肿瘤	0	0	—	—	—	3
组织细胞和附属淋巴样细胞肿瘤	13	0.06	48.00	92.31	63.16	30
非特指类型	377	1.81	94.10	88.86	91.41	28
Kaposi肉瘤	4	0.02	100.00	75.00	85.71	4
其他特指类型的肿瘤	960	4.62	41.34	96.67	57.91	222
非特指类型的肿瘤	1 362	6.56	41.54	93.91	57.60	11
合计	20 776	100.00	77.20	96.27	85.69	1 090

值高于 95，分类效果较好；间皮瘤与白血病的 *F* 值未超过 50，分类效果不佳。在中文新闻文本进行多种分类方法比较的研究中发现，支持向量机在分类效果上较优于朴素贝叶斯和 KNN 两种分类器，而且 SVM 召回率最高[15]。

本研究主要针对 ICD-O-3 中形态学进行自动化编码研究。筛选后的肿瘤数据中，由于各地区编码人员水平不一，总体编码准确率较低，除存在常见的 9 种不合逻辑的组合情况[16]，还有一些分类不正确，归类错误的情况。本研究通过 16 位专业技术人员审核和订正，训练的效率和效果得到较大提高，与国外类似研究达到同等水平甚至更好[17-20]。

综上所述，文本分析联合 SVM 对肿瘤 ICD-O-3 病理形态学自动分类效果较好，能极大提高专业人员的编码效率，避免不同编码人员主观上的判断偏差。现阶段训练中的总体准确率与实际工作尚存在一定的差距，但通过后期更大样本的训练，方法的准确性将得到进一步的提升，逐步满足实际工作的需要。

参考文献

[1] FITZMAURICE C, ALLEN C, BARBER R M, et al. Global, regional, and national cancer incidence, mortality, Years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study [J]. JAMA Oncol, 2017, 3 (4): 524-548.

[2] 魏矿荣，梁智恒，刘静 . 肿瘤登记软件和商业智能在肿瘤登记中的应用 [J]. 中国肿瘤，2012，21（7）：484-487.

[3] 秦瑞，方乐，俞敏 . 文本分析方法在医学研究中的应用进展 [J]. 浙江预防医学，2015，27（10）：1008-1011.

[4] JOUHET V, DEFOSSEZ G, BURGUN A, et al. Automated classification of free-text pathology reports for registration of incident cases of cancer [J]. Methods Inf Med, 2012, 51 (3): 242-251.

[5] ALAWAD M, GAO S, QIU J X, et al. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks [J]. JAMA, 2020, 27 (1): 89-98.

[6] OLEYNIK M, PATRAO D F C, Finger M. Automated classification of semi-structured pathology reports into ICD-O using SVM in Portuguese [J]. Stud Health Technol InForm, 2017, 235: 256-260.

[7] 潘劲，胡如英，俞敏，等 . 浙江省慢性病监测信息管理系统的架构及作用 [J]. 中国预防医学杂志，2010，11（11）：1156-1157.

[8] TARONE R E. Conflicts of interest, bias, and the IARC monographs program [J]. Regul Toxicol Pharmacol, 2018, 98: A1-A4.

[9] 杜灵彬，毛伟敏，李辉章，等 . 浙江省肿瘤登记膀胱癌发病及死亡特征分析 [J]. 浙江预防医学，2014，26（5）：473-476.

[10] BERG J W. Morphologic classification of human cancer [M] // SHOTTENFELd D F J, Jr. Cancer epidemiology and prevention. 2nd ed. New York: OxFord University Press, 1996.

[11] 王庆，陈泽亚，郭静，等 . 基于词共现矩阵的项目关键词词库和关键词语义网络 [J]. 计算机应用，2015，35（6）：1649-1653.

[12] KWON O S, KIM J, CHOI K H, et al. Trends in deqi research: a text mining and network analysis [J]. Integr Med Res, 2018, 7 (3): 231-237.

[13] HUANG S, CAI N, PACHECO P P, et al. Applications of support vector machine (SVM) learning in cancer genomics [J]. Cancer Genomics Proteomics, 2018, 15 (1): 41-51.

（下转第 263 页）