

数据挖掘技术及其在临床恶性肿瘤诊疗中的应用*

张 睿 王觅也 李 楠 杨晓妍 师庆科 黄 勇

(四川大学华西医院 成都 610041)

〔摘要〕 介绍数据挖掘相关技术,包括特征选择、离群值检测模型、聚类模型、关联规则模型、分类模型、集成学习算法等方面,对数据挖掘在临床恶性肿瘤诊断、预后及管理中的应用进行具体阐述。

〔关键词〕 数据挖掘技术;恶性肿瘤;诊断及预后研究

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2015.10.011

Data Mining Technology and Its Application in the Diagnosis and Treatment of Clinical Malignant Tumors ZHANG Rui, WANG Mi-ye, LI Nan, YANG Xiao-yan, SHI Qing-ke, HUANG Yong, West China Hospital of Sichuan University, Chengdu 610041, China

〔Abstract〕 The paper introduces technologies related to data mining, including the feature selection, outlier detection model, clustering model, association rule model, classification model, ensemble learning algorithm, etc. It makes detailed explanation of the application of data mining in the diagnosis, prognosis and management of clinical malignant tumors.

〔Keywords〕 Data mining technology; Malignant tumor; Research of diagnosis and prognosis

1 引言

肿瘤是一种严重危害人类健康和生命质量的疾病,其发病率和死亡率近年来逐年上升、居高不下。肿瘤早期无特异性临床症状,一般不会引起患者重视,且临床缺乏对于高危人群的有效早期诊断方法,因而当患者有典型临床表现再就诊时,大多已属于晚期,因此探讨及发展早期发现、早期诊断的有效方法,对于改善肿瘤患者的治疗和预后、患

者的健康及生命质量都有着重要的意义。数据挖掘的重点是发现知识,辨别冗余及无用信息并将其删除,强调以自动化的方式在海量数据中搜索潜在有用的模式^[1];而针对医疗卫生领域的知识发现是一个复杂而艰巨的工程。但不可否认,数据挖掘在医疗领域特别是在医学研究与临床实践中已获得较好的效果^[2],越来越多的临床诊疗数据以信息化方式保存及数据挖掘软件的应用普及,也使得临床数据挖掘较过去更容易与简便^[3]。本文将对数据挖掘相关技术及其在恶性肿瘤诊疗中的应用进行综述。

2 数据挖掘相关技术概述

2.1 特征选择

恶性肿瘤基于基因水平的数据集往往具有极高

〔修回日期〕 2015-05-06

〔作者简介〕 张睿,博士研究生,发表论文3篇;通讯作者:王觅也。

〔基金项目〕 863 国家科技计划项目“数字化医疗区域协同应用示范”(项目编号:2012AA02A615)。

的数据维度（维度可理解为需要研究的因素，或特征属性），应用特征选择能在不失去数据原有价值的基础上，有效地去除与研究目的无关及冗余的特征属性，以提高数据挖掘的效率，改善预测精度，主要分筛选器（Filter）及封装器（Wrapper）两类方法。Filter 类方法计算效率较高，其仅依据数据集内在性质来快速评价每个特征对分类的预测能力；而 Wrapper 类方法则需结合后续具体机器学习算法对特征子集进行评价，存在较大的计算开销以及对机器学习算法的依赖，因此在生物医学领域中，其研究关注度弱于 Filter 类方法^[4]。特征选择方法常用于数据挖掘前的特征子集选择。Gandhi 等^[5]在面乳腺癌数据库中众多属性信息时，应用特征选择方法构建特征子集以降低巨大的计算开销，特征子集也获得了较原始数据集准确率更高的模糊规则。因为便于处理高维数据，特征选择方法在恶性肿瘤基因领域的应用尤为普遍，Lee 等^[6]在处理卵巢癌的数万个基因数据时，在不降低卵巢癌分类精度前提下获得了较优的特征子集，其不仅去除了大量的无关基因，生成易理解的分类规则，而且可显著提高卵巢癌的分类精度。多项研究证实，特征选择算法结合决策树算法可极大地提升医学诊断的准确性^[7-8]。

2.2 离群值检测模型

离群值检测可发现异常值、噪声或有用的信息，例如异常检测提供了一类能够在大型数据集中识别稀有事件的技术^[9]。离群值可能由错误数据引起，但同时也可能导致新的见解产生^[10]。Wu 等^[11]通过异常检测将异于正常表达水平的基因样本进行标识，发现某些癌基因只在一小部分样本中激活。所以离群值不应被简单认为就是噪声数据而被剔除，而应核实、分析后妥善处理。

2.3 聚类模型

聚类分析在恶性肿瘤中多用于疾病危险因素探索研究或观察性学习，其一般在应用其他数据挖掘方法之前进行，以提高算法准确性，因此在统计、生物医学以及机器学习领域中有较高的应用需

求，其算法的选择有赖于数据集类型及特定的应用目的。Chen 等^[12]应用聚类方法对前列腺癌的预后因素进行分析，在对肿瘤患者基因数据聚类后，应用卡方检验计算各个聚类与最终临床预后结果的关联程度，获得与预后高度相关的基因。

2.4 关联规则模型

在医学领域，关联规则用来发现数据间隐含的关联关系^[13]。Agrawal 等^[14]对美国肿瘤研究所的 SEER 数据库中肺癌患者数据进行关联规则挖掘，生成了数百条关联规则，其中大部分规则符合目前医学认知。随后基于领域知识对大部分已知的关联规则手工删除后，发现了一些影响肺癌患者远期生存率的新颖见解。关联规则还被用于癌基因的寻找中，Lopez 等^[15]利用模糊关联规则得到部分疑似与乳腺癌发病有关的基因。

2.5 分类模型

该算法是有监督学习算法，通常用于恶性肿瘤的诊断及预后分析。分类模型较多，而综合多种方法的集成算法也在不断涌现，如 Fan 等^[16]提出的一种基于案例的数据聚类方法结合模糊决策树的混合模型，以完成针对乳腺癌的分类。决策树作为在恶性肿瘤中应用最多的分类模型，其不仅有较快的训练速度，并且可产生显式分类规则，在临床研究中广泛使用。部分文献^[17-18]认为决策树模型是最优、最实用的预测模型。目前文献已经报道数百种决策树模型，而分类回归树（CART）也被部分学者认为是最适合医学数据挖掘的分类模型^[19]。

2.6 集成学习算法

该算法是一类有效提高分类准确率的方法，其核心是应用多种分类器后，投票决定最终的分类。几乎所有分类算法均可采用多分类器集成（Bagging）方法进行集成式学习。Liu 等^[20]运用集成的 C5 决策树算法对乳腺癌生存率进行预测。Kaewchinporn 等^[21]将决策树、集成学习算法以及聚类方法综合应用，在多个医学数据集上取得较好的效果。

3 数据挖掘在临床恶性肿瘤诊疗中的应用

3.1 在恶性肿瘤诊断中的应用

对恶性肿瘤数据集进行数据挖掘,可预测个体是否罹患恶性肿瘤,为其早期诊断及预警提供指导^[22]。既往决策树方法在恶性肿瘤诊断中的应用较多,但其在众多数据挖掘方法中分类精度却并非最优。就分类精度而言,部分文献认为神经网络及支持向量机算法在众多机器学习算法中效果突出。Abdelaal等^[23]研究比较了支持向量机与决策树在DDSM数据集中应用乳房肿块影像特点及年龄来预测是否罹患乳腺癌,最终支持向量机获得了最大化ROC面积。Sawarkar等^[24]在Wisconsin乳腺癌数据集中应用支持向量机与神经网络算法对乳腺癌进行预测,结果显示在预测精度上两种方法均优于临床医生,高达97%的准确率可以帮助患者免受活检痛苦。Sarvestani等^[25]比较了一系列神经网络算法的能力,其中包括了多层感知器(Multilayer Perceptron, MLP)、自组织映射算法、径向基函数(Radial Basis Function, RBF)及概率神经网络,以用来比较对Wisconsin乳腺癌数据集及Shiraz Namazi医院乳腺癌数据集的诊断分类效果,结果显示概率神经网络分类精度最优。Padmavati等^[26]同样对Wisconsin乳腺癌数据集进行乳腺癌预测,其单独应用RBF与MLP对比Logistic回归。结果显示,两种神经网络模型在构建时虽比Logistic回归花费更多的时间,但其敏感度及特异度均优于Logistic回归。综上,针对恶性肿瘤诊断的数据挖掘研究中,决策树方法能输出显式的分类规则,因此在临床研究中应用较多;而神经网络及支持向量机等算法却拥有更高的分类精度。因此,临床研究中算法的选择应通过研究目的、预试验效果、是否需要输出易理解的显式规则等综合考虑。

3.2 在恶性肿瘤预后中的应用

临床医生需评估多种治疗方案的优劣及预估患者的预后。预后分析主要是生存时间分析,因为截尾数据的存在,肿瘤预后的预测相对肿瘤诊断来说

更为困难。因为只有一小部分患者能观察到肿瘤的复发,对于这类不知道何时、是否复发的截尾数据,随访只能得到最后一次的随访结果,定义这种情况为“无病生存期”。对此,Pantel^[27]总结数据挖掘在肿瘤预后预测中的3个研究热点:(1)预测肿瘤生存率(危险因素评估)。(2)预测肿瘤的复发。(3)预测肿瘤的生存概率。肿瘤预后的预测也可分成两类问题进行处理:一类是尚未复发的患者(截尾数据),另一类是某个时点已复发的患者^[28]。肿瘤预后研究多选择临床可获取的数据进行分析,但加入基因数据的联合研究也越来越多,Gevaert^[29]选择将临床数据与基因微阵列数据共同用于乳腺癌预后的预测,通过贝叶斯网络模型自动进行特征选择,识别出这些相关因素与乳腺癌的相关程度。在肿瘤预后研究中,决策树是应用最多的算法之一。Delen等^[30]就乳腺癌的生存率进行了研究,使用神经网络、决策树以及Logistic回归模型建立了基于20万份乳腺癌样本数据的预测模型,通过特征选择筛选出72个特征以构建预测模型,决策树C5.0获得最优精度。实际应用中发现,如能将多种机器学习模型联合应用,充分利用各模型的优点,可提高肿瘤预后的预测精度。Khan等^[31]在SEER数据集上基于模糊集-决策树方法建立了一种混合模型,以尝试不同的决策树规则与不同模糊集的组合情况,发现混合的模糊决策树较单个算法具有更好的鲁棒性及平衡性。Choi等^[32]比较了神经网络、贝叶斯网络模型以及包含二者的混合模型,利用SEER数据集中9个临床认可的数据变量以预测乳腺癌的预后,最终神经网络与混合模型的正确率均较高。

3.3 在恶性肿瘤疾病管理中的应用

疾病管理是一种对慢性疾病进行综合干预的模式,通过运用标准化的临床指南、循证实践以及强调对病人的教育来预防病情恶化,达到提升临床预后的目的^[33],而数据挖掘可协助评价疾病管理的效果。Labib等^[34]对埃及儿童急性淋巴细胞白血病管理项目的数据进行了挖掘研究,通过Clementine数据挖掘工具展现了人群地理、年龄分布,揭示了可能的疾病相关危险因素,形成有指导意义的公共卫

生决策。Kang 等^[35]对医院电子病历信息进行抽取后构建临床数据仓库,完成对乳腺癌患者为期 5 年的疾病管理。其后对临床数据仓库中不同乳腺癌术式的术后存活率进行了数据挖掘研究,客观地对疾病管理效果做出了合理评价。

4 结语

来自医学及信息科学的学者从多角度、多领域将数据挖掘相关技术引入到恶性肿瘤的研究中,取得了令人鼓舞的成果,这些工作为恶性肿瘤数据的进一步研究奠定了良好基础,也预示了医学与信息科学进行学科交叉、相互结合的美好前景。然而基于临床信息系统,应用数据挖掘后真正能持续性辅助临床决策的系统目前报道仍较少。这一方面是由于学科间的交叉、整合不够,另一方面是临床专业有其固有的复杂性。但可预见,随着医院信息化建设的进一步深入以及学科间的相互渗透,数据挖掘在临床中的应用会越来越多,最终进一步推动恶性肿瘤的深入研究。

参考文献

- Mishra D. Predictive Data Mining: promising future and applications [J]. *Int J of Computer and Communication Technology*, 2010, 2 (1): 20 - 28.
- Iavindrasana J, Cohen G, Depeursinge A, et al. Clinical Data Mining: a review. [J]. *Yearb Med Inform*, 2009, (4): 121 - 133.
- Roddick J, Fule P, Graco W. Exploratory Medical Knowledge Discovery: experiences and issues [J]. *ACM SIGKDD Explorations Newsletter*, 2003, 5 (1): 94 - 99.
- Saeys Y, Inza I, Larrañaga P. A Review of Feature Selection Techniques in Bioinformatics [J]. *Bioinformatics*, 2007, 23 (19): 2507 - 2517.
- Gandhi K, Karnan M, Kannan S. Classification Rule Construction Using Particle Swarm Optimization Algorithm for Breast Cancer Data Sets [C]. *Bangalore: Signal Acquisition and Processing*, 2010: 233 - 237.
- Lee Z J. An Improved Algorithm with Gene Selection and Decision Rules for Ovarian Cancer [J]. *Advances in Computer Science and Its Applications*, 2012, 1 (1): 26 - 31.
- Deisy C, Subbulakshmi B, Baskar S, et al. Efficient Dimensionality Reduction Approaches for Feature Selection [C]. *Siva Kasi: Conference on Computational Intelligence and Multimedia Applications*, 2007: 121 - 127.
- Karegowda A, Manjunath A, Jayaram M. Feature Subset Selection Problem Using Wrapper Approach in Supervised Learning [J]. *International Journal of Computer Applications*, 2010, 1 (7): 13 - 17.
- Chandola V, Banerjee A, Kumar V, et al. Anomaly Detection: a survey [J]. *ACM Computing Surveys*, 2009, 41 (3): 75 - 79.
- Jacob S, Ramani R. Mining of Classification Patterns in Clinical Data Through Data Mining Algorithms [C]. *Proceedings of the International Conference on Advances in Computing*, ACM, 2012: 997 - 1003.
- Wu B. Cancer Outlier Differential Gene Expression Detection [J]. *Biostatistics*, 2007, 8 (3): 566 - 575.
- Chen X, Xu S, Wang Y, et al. Identification of Biomarkers for Prostate Cancer Prognosis Using a Novel Two - Step Cluster Analysis [J]. *Lecture Notes in Computer Science*, 2011, (7036): 63 - 74.
- 牟冬梅, 冯超, 王萍. 数据挖掘方法在医学领域的应用及 SWOT 分析 [J]. *医学信息学杂志*, 2015, 36 (1): 53 - 57.
- Agrawal A, Choudhary A. Association Rule Mining Based HotSpot Analysis on SEER Lung Cancer Data [J]. *International Journal of Knowledge Discovery in Bioinformatics*, 2011, 2 (2): 34 - 54.
- Lopez F J, Cuadros M, Cano C, et al. Biomedical Application of Fuzzy Association Rules for Identifying Breast Cancer Biomarkers [J]. *Medical & Biological Engineering & Computing*, 2012, 50 (9): 981 - 990.
- Fan C, Chang P, Lin J, et al. A Hybrid Model Combining Case - based Reasoning and Fuzzy Decision Tree for Medical Data Classification [J]. *Applied Soft Computing*, 2011, 11 (1): 632 - 644.
- Aruna S, Rajagopalan S P, Nandakishore L V. An Empirical Comparison of Supervised learning algorithms in Disease Detection [J]. *International Journal of Information Technology Convergence and Services*, 2011, 1 (4): 81 - 92.
- 李怀庆. 决策树算法在医院数据挖掘中的应用探索 [J]. *医学信息学杂志*, 2009, 30 (8): 11 - 13.
- Lavanya D, Usha R. Performance Evaluation of Decision

- Tree Classifiers on Medical Datasets [J]. International Journal of Computer Applications, 2011, 26 (4): 1-4.
- 20 Liu Y, Wang C, Zhang L. Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data [C]. Beijing: Bioinformatics & Biomedical Engineering . International Conference on ICBBE, 2009: 1-4.
- 21 Kaewchinporn C, Vongsuchoto N, Srisawat A. A Combination of Decision Tree Learning and Clustering for Data Classification [C]. Nakhon Pathom: Computer Science and Software Engineering, 2011 Eighth International Joint Conference on IEEE, 2011: 363-367.
- 22 武会苹, 李莉. 基于 CBR 的原发性心脏恶性肿瘤诊断系统设计 [J]. 医学信息学杂志, 2011, 32 (1): 41-43.
- 23 Abdelaal M, Sena H, Farouq M, et al. Using Data Mining for Assessing Diagnosis of Breast Cancer [C]. Wisla: Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on IEEE, 2010: 11-17.
- 24 Sawarkar S, Ghatol A, Pande A. Neural Network Aided Breast Cancer Detection and Diagnosis Using Support Vector Machine [C]. Cevtat: Proceedings of the 7th WSEAS International Conference on Neural Networks, Cavtat, Croatia. 2006: 158-163.
- 25 Sarvestani A, Safavi A, Parandeh N, et al. Predicting Breast Cancer Survivability Using Data Mining Techniques [J]. Lap Lambert Academic Publishing, 2010, (2): 227-231.
- 26 Padmavati J. A Comparative Study on Breast Cancer Prediction Using RBF and MLP [J]. International Journal of Scientific & Engineering Research, 2011, 2 (1): 1-5.
- 27 Pantel P. Breast Cancer Diagnosis and Prognosis [D]. Winnipeg University of Manitoba, 1998.
- 28 Chi C, Street W, Wolberg W. Application of Artificial Neural Network - based Survival Analysis on two Breast Cancer Datasets [C]. Palo Alto: AMIA Annual Symposium Proceedings, 2007: 130-134.
- 29 Gevaert O, De Smet F, Timmerman D, et al. Predicting the Prognosis of Breast Cancer by Integrating Clinical and Microarray Data with Bayesian Networks [J]. Bioinformatics, 2006, 22 (14): 184-190.
- 30 Delen D, Walker G, Kadam A. Predicting Breast Cancer Survivability: a comparison of three data mining methods [J]. Artificial Intelligence in Medicine, 2005, 34 (2): 113-128.
- 31 Khan M, Choi J, Shin H, et al. Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare [C]. Vancouver: Engineering in Medicine and Biology Society. 30th Annual International Conference of the IEEE. 2008: 5148-5151.
- 32 Choi J, Han T, Park R, et al. A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis [J]. Healthcare Informatics Research, 2009, (1): 49-57
- 33 Krumholz H, Currie P, Riegel B, et al. A Taxonomy for Disease Management: a scientific statement from the American heart association disease management taxonomy writing group [J]. Circulation, 2006, 114 (13): 1432-1445.
- 34 Labib N, Malek M. Data Mining for Cancer Management in Egypt Case Study: childhood acute lymphoblastic leukemia [J]. Transactions on Engineering, Computing & Technology, 2005, (8): 309-314.
- 35 Kang E, Han S, Kim S, et al. Five - years of Breast Cancer Management in A New Hospital: analysis using clinical data warehouse [J]. Journal of Breast Cancer, 2010, 13 (1): 96-103.

敬告作者

《医学信息学杂志》网站现已开通, 投稿作者请登录期刊网站: <http://www.yxxxx.ac.cn>, 在线注册并投稿。

《医学信息学杂志》编辑部