• 医学信息研究 •

基于临床信息系统的数据集市构建及挖掘应用*

张 睿 杨晓妍 王觅也 李 楠 师庆科 黄 勇

(四川大学华西医院 成都 610041)

〔摘要〕 基于临床信息系统(Clinical Information System, CIS)构建临床数据集市,介绍临床数据的整合、数据集市结构设计及数据预处理,构建二维数据集并基于 Weka 软件进行特征选择,最后给出应用实例。

〔关键词〕 临床数据集市;数据挖掘;特征选择;临床信息系统

〔中图分类号〕R-056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2015.12.011

Construction of CIS-based Data Mart and Mining Applications *ZHANG Rui, YANG Xiao-yan, WANG Mi-ye, LI Nan, SHI Qing-ke, HUANG Yong, West China Hospital of Sichuan University, Chengdu 610041, China*

〔**Abstract**〕 To construct the clinical data mart based on Clinical Information System (CIS), the paper presents the integration of clinical data, architecture design of data mart and data preprocessing, construction of two-dimensional dataset, feature selection based on the software Weka, and finally application examples are given.

〔**Keywords**〕 Clinical data mart; Data mining; Feature selection; Clinical Information System (CIS)

1 引言

数据集市(Data Mart)也称数据市场。近年来,医疗市场竞争日趋激烈,医院要在市场竞争中取得竞争的优势,就必须考虑利用已经积累的诊断治疗等历史数据,通过深层挖掘、分析,快速获取有价值的信息,为医院提供准确、方便的决策支持。临床信息系统(Clinical Information System, CIS)的广泛应用,使更多的日常医疗业务数据以信息化方式存储下来。而依附于 CIS 构建的临床数据集市(Clinical Data Mart)可持续地为数据分析及挖掘提供数据基础[1-2]。但随着临床信息的进一步丰富,数据的实例数与维数(即特征变量或研究变量)急剧增加,由此带来两方面问题:一是“维数灾难”,维数膨胀给高维数据中模式识别及知识发现带来挑战,许多经典的低维数据处理方法在处理高维数据时存在困难;二是“维数福音”,高维数据中蕴藏着丰富的信息,为问题解决带来了新的可能性。因此,如何将高维数据在低维空间中表示,由此发现其可能的内在关联是高维数据处理的一个关键问题。未来基因芯片数据的加入,特征数目将继续膨胀,使大多数机器学习算法所需的训练样本数量也将急剧增加[3]。而医院拥有的病例样本始终有限,因此寻

〔修回日期〕 2015-05-06

〔作者简介〕 张睿,博士,发表论文 7 篇;通讯作者:黄勇。

〔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)。

找好的特征集以代表原始数据集，不仅可以降低计算复杂度、提高预测精度，更有助于寻找精简的、泛化能力更强的模型。本研究将基于 CIS 构建临床数据集市（涉及的 CIS 及临床数据集市均基于 Caché 数据库构建），依据研究目的对其整合后的数据集（Data Set）进行特征选择等数据挖掘分析（在医学领域也称为“变量筛选”）。

2 基于 CIS 的临床数据集市建立

2.1 相关临床数据的整合

临床数据集市构建的难点在于不同操作类型信息系统中的信息整合[4]。本研究将整合下列信息：（1）病案首页等基础信息，包含年龄、性别、民族、入（出）院日期、科室、住院日等。为保护患者隐私，所有个人隐私信息均被排除。（2）临床发现类术语（Clinical Finding）信息，其是 SNOMED CT 中最重要的顶层概念之一，包含症状、体征、既往患病等。依托既往研究成果，本研究从全院主诉及现病史中抽取症状、体征及疾病、病征等共 61 861 个临床发现类术语信息，其中部分术语已与 SNOMED CT 成功映射。（3）实验室检验信息，包含各类检验医嘱（如血细胞分析）及此医嘱下的各检验项名称（如血红蛋白）、结果、单位等，共 1 312 项。（4）病案相关信息整合，包括标准的入、出院诊断编码（ICD－10）、手术编码（ICD－9－CM）、肿瘤形态学编码（ICD－O－3）等。（5）费用类相关信息，以 3 种粒度存储于数据集市中：核算分类粒度如治疗费、西药费等，可进行费别分析；医嘱项粒度包含患者的医嘱明细清单；收费项粒度包含患者所有使用的收费项目明细。（6）其余电子病历相关信息，如身高、体重以及病理及影像学中部分可结构化存储的信息。

2.2 数据集市结构设计

数据集市以“住院就诊表”为核心，其余事实表通过“病案号”字段与其进行关联。研究共包含 14 个事实表及 30 余个维度表，见图 1。

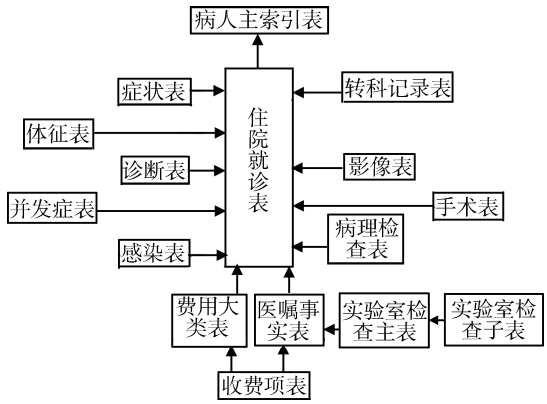

图 1 数据集市事实表关系

2.3 数据预处理

数据集市原始数据来源于 CIS 中诊疗业务数据，由于存在噪声、不完整及不一致等问题，原始数据不能直接使用。理论上所有的数据挖掘算法都是数据驱动，所以数据挖掘的结果极度依赖于数据集的质量[5]。本研究依据 CIS 中数据实际情况，进行了如下预处理工作：（1）错误数据处理。CIS 中包含多种类型的数据异常或错误，如所填数值与量纲明显不符，非法字符造成数据异常分隔、截取，非肿瘤病人出现肿瘤形态学编码或肿瘤分期等。针对这类错误，需核查原始数据及取值来源，找出问题根源，或与相关业务人员或工程师沟通后调整。（2）计量与等级资料的统一。医院不同时期的数据可能因各种原因（如检验方法或设备更替）造成存储格式前后不一致，如某些检验类指标前期是定性数据（－、1＋、2＋、3＋），其后因检验技术升级变为定量数据。这类数据需进行统一，向检验科求证以及查证相关专业文献后进行转换。（3）依据专业知识生产新的变量。如吸烟指数（包＊年）及身高体重指数（BMI）。

3 二维数据集构建与应用

3.1 构建

3.1.1 *需考虑的问题* 临床科研目的多种多样，但较普遍的是临床诊断、疾病预警以及病因及预后分析，如探索病例组与对照组间是否存在较好的疾

病鉴别特征，或不同类型的预后究竟可能与哪些因素相关，其本质是两组样本间的分析。针对这类普遍的组间分析需求，二维数据集的构建需考虑如下几点：（1）低粒度数据的汇聚。一次就诊，同一药物可能在住院期间多次使用，但应以患者就诊粒度进行组织，将同一药物多次用药信息汇聚后存储。（2）多时间点取值问题。考虑到患者在一次就诊中可能多次进行相同的检查项目，而二维表数据无法将所有同类项目完全纳入。结合多数研究目的，选取患者入院后该项目的首次检查结果以构成此数据项，以代表原始病情。（3）连续型变量是否进行离散化。如对数值型的实验室检查结果是否需离散化为“正常”、“过高”、“过低”等结果。从计算机角度，离散化后的数据在降维后可约减更多的属性，但离散化后的数据会损失部分信息量，因此本研究仅将临床上有明确等级划分的指标进行离散化处理。（4）缺失值的处理。临床业务数据中缺失值极其普遍，因为临床往往依据患者病情选择检查项目，而未进行的检查占绝大多数。但这些缺失数据其本身蕴含信息，而且部分缺失值较多的数据项（如 EB 病毒检测）还可能是构建医学分类器的关键指标，不可轻易忽略。Little 等[6]研究也表明，医学数据集的这类缺失是不可忽略、非随机缺失的，不可进行数据补齐。

3.1.2 *构建结果* 最终本研究设计了如下二维科研数据集，见表 1。此二维数据集拥有 8 万余个特征变量，存储于 Caché 数据库的 Global 中，较难直接应用。而且，在分析具体临床问题时，并非所有特征变量在此研究目的上均有体现，呈现出数据稀疏（Data Sparsity）问题。为此，研究设计如下方案对科研数据集进行动态优化，以减少特征变量的输出：针对欲研究的样本数据，遍历数据集中每个特征变量的取值情况，如其只出现过 n 次以下的非空值，则移除此特征变量（n 值可自行设定，有文献报道 n 约为总实例数 ×2% 为佳，本文为避免将潜在有意义的特征变量移除，将 n 值保守设置为 3）。于是在输出的数据中，那些无取值或取值极其稀少的特征变量被迅速移除，从而达到降低数据集维度的目的。

表 1 二维科研数据集

信息大类	属性名	英文名称	变量类型
基本信息	年龄	BaseAge	数值型
	性别	BaseGender	数值型
	体重指数	BaseBMI	数值型
医嘱信息（统计使用相应医嘱的数量，共 21 065 项医嘱信息）	某检查类医嘱	Arcim00001	数值型
	某手术类医嘱	Arcim00002	数值型
	某操作类医嘱	Arcim00003	数值型
	某药品类医嘱	Arcim00004	数值型
	某材料类医嘱	Arcim00005	数值型
	…	…	…
临床发现类术语信息（从主诉及现病史及体格检查中提取的症状、体征、疾病及病征等，共61 861 个术语信息）	T	Sign00001	数值型
	BP	Sign00002	数值型
	声音嘶哑	Symp00001	布尔型
	耳闷胀感	Symp00002	布尔型
	发热	Symp00003	布尔型
	颈部淋巴结肿大	Symp00004	布尔型
	…	…	…
实验室检查（从 LIS 系统中获取共 1 312 个实验室检查子项目，是否离散化如前所述）	凝血酶原时间	LIS0001	数值型
	癌胚抗原	LIS0002	数值型
	电解质 - 血清氯离子	LIS0003	数值型
	血肌酐	LIS0004	数值型
	…	…	…
决策属性（由具体研究目选择、决定，但必须是二分类）	是否为某类型疾病 喉癌与声带息肉的鉴别 是否有 31 天重返 是否住院期间死亡	Lable …	布尔型 …

3.2 基于 Weka 软件的特征选择

在进行数据挖掘之前，人们总希望选择有代表性的特征，但却并不知道哪些特征更富含信息量，而特征选择可很好地解决此类问题。特征选择[7]是模式识别及机器学习领域的重要研究方向，通过删除无关及冗余的特征变量，为特定的应用在不失去数据原有价值的基础上选择尽可能小的特征子集。临床上应用特征选择算法处理高维数据集，可避免无关及冗余特征对预测性能的影响，从而提高机器学习效率，增强学习模型的泛化能力，更可通过此过程发现富含信息的、潜在的、与研究病种高度相关的特征。特征选择主要分过滤式（Filter）及封装式（Wrapper）方法[8]。与

Wrapper 方法不同，Filter 方法不依赖后续具体的机器学习方法来进行特征评价，而是根据数据集内在性质评价每个特征对分类的预测能力，其通用性强、选择速度快，适合较大规模的数据集。Filter 方法进一步可分为单因素及多因素方法[9]，前者忽略特征间的相互作用，独立评估每个特征，按特征与类别的相关程度进行量化；后者则考虑多个特征间的相互作用，形成相应的特征子集。本研究主要以基于单因素的 Filter 方法进行特征选择。为方便应用数据挖掘平台 Weka 进行特征选择，通过程序实现将 Caché 中数据直接转换生成 ARFF 格式的文本文件。Weka 集成多种特征选择方法，其中，基于卡方统计量（X^2 Statistic）的特征选择方法[10]依据研究分类对每个特征计算卡方值后进行评估，对分类资料进行卡方检验量计算，而对于连续型变量一般是先将其离散化后再进行计算。卡方统计中使用特征与类别间的卡方值作为量化标准，卡方值越高，该特征相应就越重要，越应该保留供后续分析使用。

3.3 应用案例

本研究以鼻咽癌与耳鼻喉科良性疾病对比为例，选择不含医嘱信息的二维数据集进行研究。依据研究病种分类对数据集进行动态优化后，数据维度由原来的 6 万余维减少到 1 617 维，数据降维效果明显。将优化后的数据集导入 Weka 行特征选择，应用基于卡方统计量的特征选择后，不仅可以明确哪些特征与研究分类高度相关，还能给出量化结果。Weka 软件通过“特征权重算法 + 排序”方式，将相关特征按权重由高到低进行排列。经特征选择后，患者年龄、淋巴细胞绝对值、血清氯离子、回吸性涕血、鼻咽部新生物等在两组中分布差异有统计学意义，提示以上特征有助于两组疾病的鉴别。其中大部分指标符合临床预期及经验，但部分特征如血清氯离子等尚不符合临床预期。对这些不符合临床预期的指标应进行数据核查，当数据核查无误而临床仍较难理解时应查阅相关文献。如文献报道较少但数据分析组间确有统计学差异时，那么其很可能导致新的见解产生，这也是对临床数据集进行特征选择的目的。

4 结语

特征选择方法对机器学习准确率的影响比具体选择哪种机器学习算法更重要，而且特征选择算法可极大地提升医学诊断分类的准确性[11-12]。基于 CIS 构建临床数据集市，可使研究人员更便捷地获取完整的科研数据；而系统只需依据研究目的简单设置目标变量及相关纳入、排除条件，即可灵活、定制化地从数据集市中获取相应整合、降维后的数据，可通过 Weka 软件筛选富含信息量的重要特征变量，从而帮助临床医生更有效地利用 CIS 中的信息资源。就方法学而言，这类组间分析适用于临床诊断、疾病预警、病因及预后分析等多类型研究场景，具有较好的通用性。进一步而言，本研究基于临床实际数据得到的“知识”不仅具有定性特征，而且具有重要性排序的量化特征，且适用性更好，是应用信息技术辅助临床决策的有益尝试。

参考文献

1 石晓敬. 数据挖掘及其在医学信息中的应用 [J]. 医学信息学杂志, 2013, 34 (5): 2-6.

2 孔琳. 数据挖掘在医院信息系统中的应用 [J]. 医学信息学杂志, 2011, 32 (10): 37-39.

3 Jain A, Zongker D. Feature Selection: evaluation, application, and small sample performance [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19 (2): 153-158.

4 Sheta O E, Eldeen A N. Building a Health Care Data Warehouse for Cancer Diseases [J]. International Journal of Database Management Systems, 2012, 4 (5): 39-46.

5 Ting S L, Shum C C, Kwok S K, et al. Data Mining in Biomedicine: current applications and further directions for research [J]. Journal of Software Engineering, 2009, 2 (3): 150-159.

6 Little R J A, Rubin D B. The Analysis of Social Science Data with Missing Values [J]. Sociological Methods & Research, 1989, 18 (2/3): 292-326.

（下转第 60 页）

在以下几方面进一步探讨：（1）构建规模更大、质量更高的语料库。机器学习方法主要是通过统计训练语料来得到相关参数并建立模型，因此语料库所含基因实体越多、语料库质量越高，建立的模型识别效果越好。（2）提取深层次的实体特征，研究高效的特征表示方法。目前选取的单词特征、词性特征等只是对命名实体名称或语法成分的一种匹配，只用到了表层的文本信息，无法有效地识别句子中隐含的实体信息。在后续研究中，应更注重利用文本中的句法知识等深层次的信息，提取文本中命名实体的共指特征，从而提高系统识别命名实体的能力。（3）研究词典与机器学习方法更优的结合机制。基于词典是命名实体识别的一种比较简单的方式，完备的词典可提高系统识别已知命名实体的能力。因此，一方面可以通过词语原型化工具改进词典匹配算法，以降低英文单词的词形变化对词典特征构建的影响；另一方面还需要基于词典构建更多的特征加入到机器学习方法中，以减少机器学习模型对语料库的依赖，从而为基因命名实体识别系统从理论探索走向实际应用提供条件。

参考文献

1 Hatzivassiloglou V, Duboue′ PA, Rzhetsky A. Disambiguating Proteins, Genes and RNA in text: a machine learning approach [J]. Bioinformatics, 2001, 1 (1): 1 - 10.

2 National Center for Biotechnology Information, U. S. National Library of Medicine. Semantic Network - UMLS® Reference Manual [EB/OL]. [2015 - 02 - 10]. http://www.ncbi.nlm.nih.gov/books/NBk9679/.

3 王琦. 词典和机器学习相结合的生物命名实体识别 [D]. 大连：大连理工大学，2009.

4 郑强. 生物医学命名实体识别研究 [D]. 长沙：国防科学技术大学，2009.

5 黄浩炜. SVM 与基于转换的错误驱动学习方法相结合的生物实体识别 [D]. 长沙：国防科学技术大学，2007.

6 周荣鹏. 生物医学文献中命名实体的识别 [D]. 大连：大连理工大学，2009.

7 The Stanford Natural Language Processing Group. Stanford Log - linear Part - of - Speech Tagger [EB/OL]. [2015 - 02 - 15]. http://nlp.stanford.edu/software/tagger.shtml.

8 Smith L, Rindflesch T, Wilbur W J. MedPost: a part - of - speech tagger for bioMedical text [J]. Bioinformatics, 2004, 20 (14): 2320 - 2321.

9 Tsuruoka Y, Tateisi Y, Kim J D, et al. Developing a Robust Part - of - Speech Tagger for Biomedical Text [J]. Advances in Informatics Lecture Notes in Computer Science, 2005, (374): 382 - 392.

10 Department of Information Science, Faculty of Science, University of Tokyo. GENIA Tagger: part - of - speech tagging, shallow parsing, and named entity recognition for bio medical text [EB/OL]. [2015 - 02 - 15]. http://www.nactem.ac.uk/GENIA/tagger.

（上接第 53 页）

7 Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection [J]. Journal of Machine Learning Research, 2003, (3): 1157 - 1182.

8 Sun Z, Bebis G, Miller R. Object Detection Using Feature Subset Selection [J]. Pattern Recognition, 2004, 37 (11): 2165 - 2176.

9 Saeys Y, Inza I, Larrañaga P. A Review of Feature Selection Techniques in Bioinformatics [J]. Bioinformatics, 2007, 23 (19): 2507 - 2517.

10 Jin X, Xu A, Bie R, et al. Machine Learning Techniques and Chi - square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles [J]. Data Mining for Biomedical Applications, 2006, (3916): 106 - 115.

11 Deisy C, Subbulakshmi B, Baskar S, et al. Efficient Dimensionality Reduction Approaches for Feature Selection [C]. Conference on Computational Intelligence and Multimedia Applications, 2007.

12 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.