基于两步聚类算法的高血压电子病历数据挖掘研究*

杨美洁

（重庆医科大学医学信息学院 重庆 400016）

〔摘要〕 采用 SQL 技术对高血压患者电子病历的基本信息和病程记录进行数据预处理，利用 SPSS 19.0 软件中的两步聚类算法进行分析，挖掘出肺炎、脑梗塞、糖尿病等预测高血压的重要因素信息，为高血压的诊断和治疗提供参考依据。

〔关键词〕 SQL；高血压；电子病历；两步聚类算法

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2016.12.003

Research on Data Mining of Hypertension Electronic Medical Records（EMR）Based on Two－step Clustering Algorithm

YANG Mei－jie, Medical Informatics College, Chongqing Medical University, Chongqing 400016, China

〔**Abstract**〕 The paper preprocesses the data about basic information of Electronic Medical Records（EMR）and the progress note of hypertension patients through SQL technology, conducts analysis based on the two－step clustering algorithm of SPSS 19.0 software, and explores the important factors like pneumonia, brain infarction and diabetes and so on for predicting hypertension, in order to provide reference for diagnosis and treatment of hypertension.

〔**Keywords**〕 SQL; Hypertension; Electronic Medical Records（EMR）; Two－step clustering algorithm

1 引言

高血压是最常见的心血管疾病，危害巨大，是冠心病、心肌梗塞的主要诱因[1]，不仅具有较高致残率和致死率[2]，而且严重消耗了我国的医疗资源，位于引起死亡的 10 大危险因素之首[3]。近年来随着人们生活节奏的加快、生活压力的增大，高血压发病率也呈现出上升趋势，但高血压知晓率、治疗率和控制率却一直很低，因此如何提高高血压的知晓率、治疗率和控制率具有重要的研究意义。

近年来电子病历以其传递速度快、存储容量大、使用方便、共享性好和成本低等优点，在我国医疗系统中得到广泛应用。电子病历是以电子化方式管理的有关个人终生健康状态和医疗保健行为的信息[4]，在给患者带来方便的同时，也积累了海量数据，为数据挖掘提供了基础。本文通过研究高血压患者电子病历的基本信息和病程记录，选取基本信息中的性别、年龄和病程记录中与高血压相关的“冠状动脉粥样硬化性心脏病”[5]、“脑梗塞”[6]、“糖尿病”[7]、“肺炎”[8]等因素，采用两步聚类算法进行挖掘，以期为高血压的诊断和治疗提供参考依据。

〔修回日期〕 2016－09－09

〔基金项目〕 杨美洁，讲师，发表论文 7 篇。

〔基金项目〕 重庆市社会事业与民生保障科技创新专项（项目编号：cstc2015shms－ztzx10003）。

2 资料与方法

2.1 资料来源

收集重庆市某综合医院近年主诊断为“高血压”的电子病历 3 736 份，经过数据预处理后剩余 2 312 份。

2.2 数据收集

研究高血压患者电子病历的基本信息和病程记录，从基本信息中选取住院号、性别、年龄 3 个属性，从病程记录中选取住院号和记录内容两个属性，从记录内容中抽取“冠状动脉粥样硬化性心脏病”、“脑梗塞”、“糖尿病”、“肺炎”等与高血压相关的因素。

2.3 数据预处理

电子病历数据结构化处理是临床数据分析的前提[9-10]，包括数据清洗、集成和转换[11]。本文利用 SQL Server 2008 对数据进行预处理，得到符合研究需要的数据和模型。

2.3.1 基本信息数据预处理 对基本信息中的住院号、性别、年龄进行处理。性别按照男、女分别取值为 1、2。其中部分代码和结果如下：

```
update 基本信息 set 性别 = 1 where 性别 = ‘男’
update 基本信息 set 性别 = 2 where 性别 = ‘女’
```

基本信息预处理，见表 1。

表 1 基本信息预处理

住院号	性别	年龄
11002615	2	62
11015108	2	81
……	……	……

2.3.2 病程记录数据预处理 对病程记录中的住院号、记录内容进行处理。从记录内容中抽取与高血压相关的“冠状动脉粥样硬化性心脏病”（1 035 人）、“脑梗塞”（570 人）、“糖尿病”（863 人）、“肺炎”（555 人）等因素进行研究；若记录内容中出现“无脑梗塞”、“无糖尿病”、“无肺炎”、“无冠状动脉粥样硬化性心脏病”等内容时，相应的因素取值为 0，否则取值为 1。其中部分代码和结果如下：

```
update 糖尿病病程记录
set 记录内容 = 0
where 记录内容 like ‘% 无糖尿病% ’
update 糖尿病病程记录
set 记录内容 = 1
where 记录内容 ! = 0
update 病程记录 set 糖尿病 = 记录内容
from 糖尿病病程记录，病程记录
where 病程记录. 住院号 = 糖尿病病程记录. 住院号 and 记录内容 = 1
```

病程记录预处理，见表 2。

表 2 病程记录预处理

住院号	冠状动脉粥样硬化性心脏病	脑梗塞	糖尿病	肺炎
10002528	0	0	0	0
12002334	0	0	1	0
……	……	……	……	……

2.3.3 数据集成 以住院号为关联字段，将基本信息预处理表和病程记录预处理表集成为 1 个表。其中部分代码和结果如下：

```
select distinct 基本信息. 住院号，冠状动脉粥样硬化性心脏病，脑梗塞，糖尿病，肺炎，性别，年龄
into 数据集成
from 病程记录，基本信息
where 病程记录. 住院号 = 基本信息. 住院号
```

数据集成，见表 3。

表 3 数据集成

住院号	冠状动脉粥样硬化性心脏病	脑梗塞	糖尿病	肺炎	性别	年龄
10002528	0	0	0	0	2	75
12002334	0	0	1	0	2	77
……	……	……	……	……	……	……

3 两步聚类算法

3.1 算法估计

聚类是数据挖掘描述任务的一个重要组成部分，其目的是发现数据集中的模式。两步聚类是一种探索性的算法，旨在揭示数据集中不明显的自然集群[12]。两步聚类算法不需要提前预定聚类数目，能够诊断出离群点和噪声数据[13]，可用来处理海量数据，并且具有同时处理离散变量和连续变量的能力。两步聚类分为预聚类和正式聚类。第 1 步：对记录进行初始归类，用户可以自定义最大类别数。通过 BIRCH（Balanced Iterative Reducing and Clustering using Hierarchies）[14] 算法构建和修改聚类特征树（Cluster ing Feature Tree）完成初步归类。第 2 步：对预聚类的初步聚类结果进行再聚类，系统根据一定的统计标准得到最佳聚类结果。

3.2 聚类模型概要及聚类大小分布

本文采用 SPSS 19.0 软件的两步聚类算法对集成后的数据进行分析。以“冠状动脉粥样硬化性心脏病”、“脑梗塞”、“糖尿病”、“肺炎”等作为分类变量，以年龄作为连续变量进行输入，见图 1。由图 1 可知，输入 6 个变量（5 个分类变量、1 个连续变量），最终将数据聚成 5 类，聚类质量一般。5 类包含样本的数量和所占比例，见图 2。聚类 1 包含 492 个样本（21.3%），聚类 2 包含 555 个样本（24.0%），聚类 3 包含 425 个样本（18.4%），聚类 4 包含 332 个样本（14.4%），聚类 5 包含 508 个样本（22.0%）。

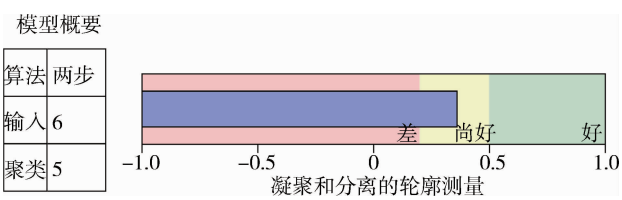

图 1 聚类质量

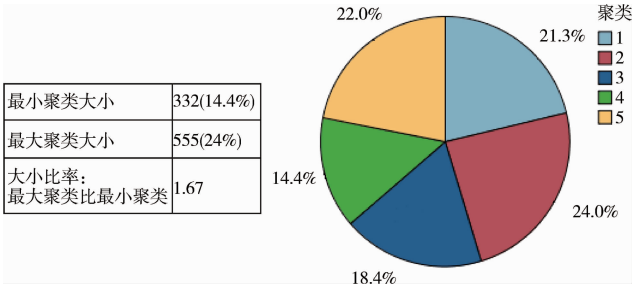

图 2 聚类大小分布

3.3 预测变量重要性

从图 3 可以看出，在高血压的预测变量重要性中，肺炎和脑梗塞是最重要的，值为 1.0，糖尿病值为 0.93，冠状动脉粥样硬化性心脏病值为 0.59，年龄值为 0.2，性别值为 0.02。

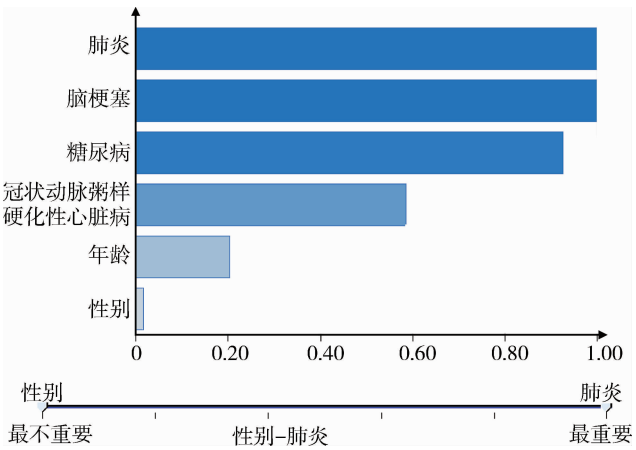

图 3 预测变量重要性排序

3.4 聚类总体分布

从图 4 可以看出，第 1 类特征——均无肺炎和脑梗塞，均有糖尿病，冠状动脉粥样硬化性心脏病患者占 44.3%，年龄均值为 68.63 岁，女性患者占 63.8%；第 2 类特征——均有肺炎，脑梗塞患者占 26.1%，糖尿病患者占 37.3%，冠状动脉粥样硬化性心脏病患者占 45.6%，年龄均值为 75.34 岁，女性患者占 52.6%；第 3 类特征——均无肺炎，均有脑梗塞，糖尿病患者占 38.6%，冠状动脉粥样硬化性心脏病的患者占 43.1%，年龄均值为 72.75 岁，女性患者占 50.4%；第 4 类特征——均无肺炎、无脑梗塞、无糖尿病，均有冠状动脉粥样硬化性心脏

病，年龄均值为 72.52 岁，女性患者占 63.6%。第 5 类特征——均无肺炎、无脑梗塞、无糖尿病、无冠状动脉粥样硬化性心脏病，年龄均值为 63.28 岁，女性患者占 60.4%。

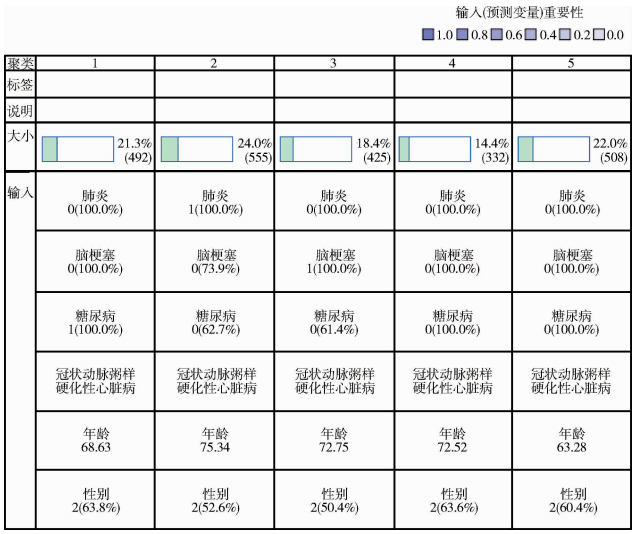

图 4 聚类总体分布

4 讨论

4.1 高血压的预测变量重要性分析

通过两步聚类分析，挖掘出肺炎和脑梗塞是高血压最重要的预测变量。目前，多项临床和流行病学研究显示，高血压患者肺部感染率要显著高于非高血压患者，肺炎支原体感染和高血压之间具有密切的联系，并且超重、性别和血脂异常与肺炎支原体感染存在的交互作用对高血压的影响十分显著[8]，并且发现高血压与脑梗塞和脑出血密切相关，脑出血组发病前有高血压病史的患者占 66.7%，脑梗塞组为 57.4%，脑出血组加梗塞组为 58.3%[15]。高血压还有可能增加患糖尿病的风险[16]，我国糖尿病人群高血压发病率高于普通人群并随年龄增高[17]。

4.2 聚类特征分析

通过 5 类可以发现，45% 左右的高血压患者同时伴有冠状动脉粥样硬化性心脏病。研究表明，高血压与冠状动脉粥样硬化性心脏病（冠心病）密切相关，是冠心病的独立危险因素，同时也是冠心病最常见的合并症之一[18]。而且发现难治性高血压与性别无关，与年龄密切相关，高血压患者的年龄主要分布在 60 ~ 80 岁之间，并且女性稍多于男性[19]。

5 结语

本文以高血压患者电子病历为研究对象，采用 SQL 数据库技术对电子病历中的基本信息、病程记录进行预处理，利用 SPSS 19.0 软件中的两步聚类算法进行分析，得到了高血压较为重要的预测变量，得到了每一类的共同特征。可为高血压的诊断和治疗提供参考依据。

参考文献

1 刘力生. 中国高血压防治指南 2010 [J]. 中华高血压杂志，2011，19（8）：701 - 743.
2 周亚东，刘晓红，张永强，等. 陕西省农村老年人高血压患者知晓率治疗率和控制率的现况调查研究 [J]. 中国预防医学杂志，2016，17（3）：170 - 172.
3 中华预防医学会慢性病预防与控制分会. 慢性病的流行形势和防治对策 [J]. 中国慢性病预防与控制杂志，2005，13（1）：2 - 3.
4 李伟明. 电子病历档案应用现状及前景的探讨 [J]. 广州档案，2010，(3)：38 - 39.
5 张岩，霍勇. 伴同型半胱氨酸升高的高血压——“H 型”高血压 [J]. 心血管病学进展，2011，32（1）：3 - 6.
6 苗保霞. 探讨高血压及合并冠心病或脑梗塞患者的动态血压特点 [D]. 济南：山东大学，2014.
7 陈永刚，李云，安利杰，等. 高血压对糖尿病人群心脑血管事件的影响 [J]. 中华高血压杂志，2013，21（4）：346 - 351.
8 项燕凌，兰青，陈嘉利，等. 肺炎支原体感染与高血压患者的血清流行病学关系研究 [J]. 中华医院感染学杂志，2015，25（20）：4636 - 4638.
9 Chen Y K, Mani S, Xu H. Applying Active Learning to Assertion Classification of Concepts in Clinical Text [J]. Journal of Biomedical Informatics, 2012, 45 (2): 265 - 272.
10 Jonnalagadda S, Cohen T, et al. Enhancing Clinical Concept Extraction with Distributional Semantics [J]. Journal of Biomedical Informatics, 2012, 45 (1): 129 - 140.

（下转第 41 页）

则上可以按介于可控网络与不可控网络的中间区域对待，策略配置参照 DMZ 区。从应用系统角度，也应建议相关软件厂商进行加密处理。

5 结语

医疗行业将不可避免地走向“互联网 +”的时代，从简单的收费系统到临床大数据，越来越多的医院业务依托于信息技术运行。而医院信息系统显然还没有为此做好充足的准备。正如有关信息安全专家所说：“当前医疗行业采取的安全防护手段，与其所掌握的数据资源不匹配。”有专家认为当前国内医院的信息化安全投入与国际通常的比例相差 1 ~ 2 个数量级。本文所关注的安全边界控制问题只是安全防护的第 1 道关口，一个安全的信息系统还应建立更加完善的纵深防御体系，包括主机配置的安全基线、按应用系统划分的网络域隔离、数据加密传输存储、用户行为监控与审计等，最终达到进不来、拿不走、看不懂、改不了、走不脱、有记录的网络信息安全建设目标。

参考文献

1 何玲，刘曰波，魏津瑜. 信息安全四十年三大步［J］. 信息系统工程，2007，(12)：64 - 65.

2 李亚子，尤斌，王晖，等. 医疗保险信息泄露案例分析及对我国安全隐私保护的借鉴［J］. 医学信息学杂志，2014，35 (2)：6 - 12

3 孟晓阳，郭杰峰. 使用 IT 运行监控系统保障医院信息系统的高可用性［J］. 医学信息学杂志，2015，36 (2)：23 - 26.

4 IT 治理俱乐部. 对程稚瀚案的分析［EB/OL］.［2016 - 01 - 10］. http：//blog. vsharing. com/itgov/A884579. html.

5 赵韡. 赵韡：医院信息安全威胁案例及应对方法［EB/OL］.［2016 - 01 - 10］. http：//news. hc3i. cn/art/201411/31752. htm.

6 McCann E. Slideshow：10 biggest HIPAA data breaches in the U. S［EB/OL］.［2016 - 01 - 10］. http：//www. he-althcareitnews. com/slideshow/slideshow - 10 - biggest - hipaa - data - breaches - us.

7 陈卫平. DMZ 区安全建设模型初探［J］. 现代电视技术，2013，(2)：125 - 128.

8 百度百科. VPN 虚拟专网［EB/OL］.［2016 - 01 - 10］. http：//baike. baidu. com/link？ url = LUtKs9S9gNTbYvYu2hvR9dNLrs4NHeEcn8bpS2 - JJh3mm4vs4Dusnluzoe4pxXny1wyQ7tyVWxo3OOKlWmKxLa.

9 张益钊，朱卫国，孟晓阳，等. 医院信息系统等级保护测评实践［J］. 医学信息学杂志，2015，36 (10)：14 - 18.

(上接第 17 页)

11 李准，冯思佳，杨美洁，等. 关联规则技术在冠心病电子病历中的应用［J］. 医学信息学杂志，2015，36 (1)：58 - 62.

12 Wu X，Zhan F B，Zhang K，et al. Application of a Two - step Cluster Analysis and the Apriori Algorithm to Classify the Deformation States of Two Typical Colluvial Landslides in the Three Gorges，China［J］. Environmental Earth Sciences，2016，75 (2)：1 - 16.

13 施卓敏，孙健英，何晓涛. 基于两步聚类分析方法的 ARP 系统用户分析［J］. 计算机与现代化，2014，(3)：73 - 76.

14 Zhang T，Ramakrishnan R，Livny M. BIRCH：an efficient data clustering method for very large databases［J］. ACM Sigmod Record，1996，25 (2)：103 - 114.

15 杨遇春，林志国，戴钦舜，等. 高血压脑出血及脑梗塞危险因素 Logistic 回归分析［J］. 中国慢性病预防与控制，1995，3 (1)：3 - 6，46.

16 刘湘琳，吕淑荣，张凤云，等. 高血压合并糖尿病的相关危险因素分析［J］. 南京医科大学学报：自然科学版，2013，33 (1)：68 - 72.

17 孙静，黄玉艳，吴雷，等. 糖尿病人群高血压的发病率及影响因素［J］. 中华高血压杂志，2013，21 (7)：654 - 658.

18 霍勇，付洪喜，金振刚，等. 高血压伴冠状动脉粥样硬化性心脏病患者降压治疗的选择［J］. 中华高血压杂志，2011，19 (4)：305 - 311.

19 刘爱兵，李俊锋，卢艳芬. 对难治性高血压患者在不同性别及年龄中病因构成的探究［J］. 当代医学，2013，19 (19)：47 - 48.