的具体体现。职业病信息管理系统初步建立后，可迅速调取出所需信息，并对信息进行维护及分类，进而对职业危害进行动态分析评价，为职业病危害治理提供可靠的科学依据；对建立健全职业卫生管理体制，贯彻《中华人民共和国职业病防治法》具有积极而深远的影响。

目前，用人单位普遍存在职业卫生档案建档率低、内容不全和缺乏有效管理等问题，直接影响了职业卫生信息的完整性、准确性和时效性，束缚了我国职业危害监管网络化的建设和应用。因此，应当加强职业卫生管理人员的业务培训，不断提高他们的能力及技术水平，加强职业卫生档案的建立、管理和利用，促进职业卫生管理规范化，为职业卫生的信息化管理和应用奠定良好的基础，从而更加有力地防控职业危害、保护劳动者健康。

参考文献：

[1] 刘玉琴. 职业卫生信息管理系统的研制开发 [J]. 职业与健康，2004，20 (6)：106-107.
[2] 徐益珊，冯勇，李兆明，等. 劳动卫生职业病预防信息管理集成系统的研究 [J]. 环境与职业医学，2003，20 (1)：15-16.
[3] 何家禧，黄先青. 深圳市职业卫生管理系统软件的开发应用 [J]. 中国职业医学，2000，27 (1)：43.
[4] 张荣军，王跃平. 铝行业职业安全卫生管理信息系统的研究 [J]. 工业安全与环保，2003，29 (5)：45-46.
[5] 张金龙，姚健，焦建栋，等. 职业卫生管理信息服务平台的开发与应用 [J]. 中华劳动卫生职业病杂志，2010，28 (2)：128-130.
[6] 焦建栋，姚健，张恒东，等. 地市级职业病防治信息网络管理平台的研发与应用 [J]. 南京医科大学学报. 2009，35 (2)：125-130.
[7] 常德强，柳静献，陈宝智. 生产企业职业卫生管理信息系统开发 [J]. 工业安全与环保，2011，37 (9)：9-11.
[8] 黄缪，倪淑萍，曾德才，等. 职业卫生基础档案信息平台开发 [J]. 中国职业医学，2010，37 (1)：53-54.
[9] 于永中，高星，雷卫星，等. 北京市劳动卫生与职业病信息计算机管理系统的研究 [J]. 中华劳动卫生职业病杂志，2000，18 (4)：255-256.

数据挖掘技术在职业健康监护信息管理中的应用

周浩，潘明伟，张远辉，肖吕武，吴琳，刘移民

（广州市职业病防治院/广东省“十二五”医学重点专科职业健康监护科，广东 广州 510620）

关键词：健康监护；信息管理；数据挖掘

中图分类号：R197.32 文献标识码：C

文章编号：1002－221X(2014)03－0229－03

DOI:10.13631/j.cnki.zggyyx.2014.03.033

如何有效利用职业健康监护业已建立的庞大数据信息，从逐步积累的人群健康信息中去探寻相关职业病危害因素健康效应新的知识与潜在关系，通过内在规律对相关职业损害的发生进行趋势预测分析有着重要的实用价值和现实意义。数据仓库能提供面向主题的、集成的、稳定的和随时间变化的数据集合；数据挖掘又称知识发现（knowledge discovery in database，KDD），能从大量数据中提取或“挖掘”知识[1]；两项技术当前都被金融、电信等诸多领域的信息管理所广泛采用。本研究将以我院职业健康监护信息数据库为基础，对其进行数据挖掘的实践方法进行深入探讨。

1 材料与方法

1.1 研究对象

广州市职业病防治院职业健康监护中心信息数据库系统中近 2 年的部分职业健康检查结果及其相关个人信息，共包含 16 万多名受检者、1000多万条检查项目及相关信息项目。

1.2 方法

1.2.1 数据抽取与预处理 分析数据库结构确立研究所需数据清单，包括个人健康检查结果、职业因素接触及其他非职业相关数据等，应用数据抽取、集合、数据导入创建职业健康分析临时数据库，并对入库数据进行数据清洗，采用专家经验、预测等手段对缺失及异常数据进行补充，处理噪声数据及无法补充数据，完成上述处理后创建职业健康分析数据仓库。

1.2.2 健康检查数据的特征分析 使用 Microsoft 公司 SQL Server 2008 中 Analysis Services 多维数据集工具，通过对职业健康数据仓库中各类健康相关数据进行分层、聚集等分析，并对数据进行上卷、下钻、切片、切块及转轴等联机分析处理（OLAP），了解职业健康损害发生的特征。

1.2.3 数据挖掘 使用 Microsoft 公司 SQL Server 2008 的 Analysis Services 服务数据挖掘程序，对集中的数据进行影响因素、关联、预测等挖掘分析。

2 结果

2.1 数据挖掘概念模型创建

根据职业健康检查数据挖掘的目标问题，分析数据挖掘的结构特征。再从现有的职业健康监护信息关系数据库中通过抽取、整合、派生等数据处理方法，创建职业健康分析数据仓库，分析数据特征，并以此作为数据基础创建数据挖掘模型，调整模型参数进行信息挖掘，并对模型中发现的关联规则进行分析解释及使用验证数据对模型效能进行检测，概念模型见图 1。

收稿日期：2014－03－26

基金项目：广州市医药卫生科技项目（20121A011100）；广州市医药卫生科技重点项目（2012A021017）

作者简介：周浩（1978—），男，主治医师，硕士，主要从事职业健康监护工作。

通讯作者：刘移民，E-mail：ymliu61@163.com。

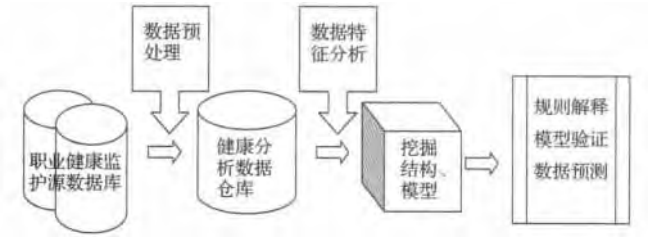

图 1 职业健康检查数据挖掘概念模型

2.1.1 数据预处理 由源数据库导出职业人群健康源数据表缓存至临时存储数据库中，主要项目涉及性别、年龄、职业病危害因素接触情况、吸烟情况、饮酒情况、饮食习惯、体育锻炼情况、既往疾病史、家族史、血压、耳鼻喉检查、心电图、EB 病毒检测结果、纯音听阈测试结果等，对数据进行清洗，包括对缺失值无法赋值者均以“Unknown”或“0”填充、异常值的处理如年龄“20”而工龄“25 年”等类似情况以该工龄段中平均年龄予以填充，如无法采用上述方式处理者也以“Unknown”或“0”填充，根据本次的数据挖掘主题对需要转换的属性值进行离散数据与连续值数据之间的相互转换、数据值的分组分类、数据项之间的计算组合，使用聚合对数据进行缩减，部分主要分析项目转换对照见表 1。

表 1 主要分析项目数值转换对照

项目名称	转换结果
既往疾病史	1-既往体健，2-耳鼻喉疾病，3-心血管疾病，4-营养代谢类疾病，5-呼吸系统疾病，6-消化系统疾病，7-神经系统疾病，8-肿瘤疾病，9-其他疾病
家族疾病史	1-家族无疾病史，2-耳鼻喉疾病，3-心血管疾病，4-营养代谢类疾病，5-呼吸系统疾病，6-消化系统疾病，7-神经系统疾病，8-肿瘤疾病，9-其他疾病
血压	1-正常，2-偏低，3-偏高（舒张压≥90 mm Hg，收缩压≥140 mm Hg）
耳鼻喉检查	1-正常，2-炎症，3-器质性病变，4-其他
心电图	1-正常，2-心房、心室肥大，3-心肌缺血，4-心律失常，5-其他异常
纯音听阈测试	1-听力正常，2-高频听阈提高平均＜40 dBHL、语频听力正常，3-高频听阈提高平均≥40 dBHL、语频听力正常，4-高频、语频听力下降，5-单侧听力损失

2.1.2 创建职业健康分析数据仓库 根据挖掘主题建立纯音听阈测试分析的多维数据集，采用星形模型结构，包括 1 个职业健康分析事实表与时间维、人员基本信息维、职业史维、职业病危害维、疾病史维、检查项目维、健康评价维 7 个维表，体检人员基本信息维表存放体检者的性别、年龄、吸烟、饮酒、运动习惯等信息，职业情况维表中记录了工作所在企业的经济类型、所属行业等信息，职业病危害维表中存放所接触的职业病危害分类、名称、接触工龄等信息，事实表中 FK 为各维表外键，检查值计数列为度量值。详见图 2。

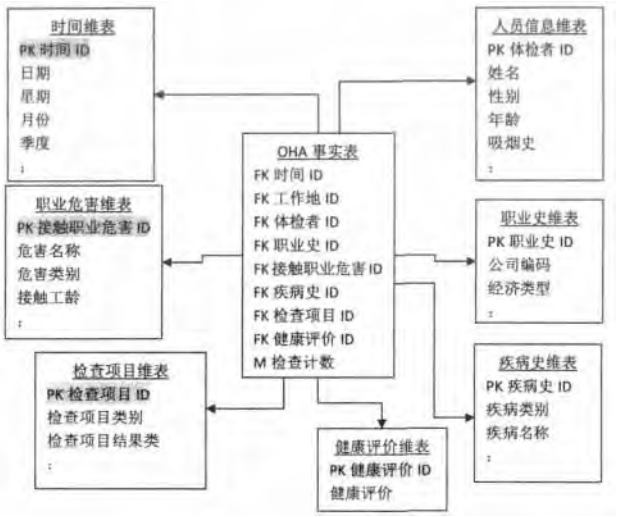

图 2 职业健康分析数据星型架构逻辑模型

2.2 职业健康多维数据集的 OLAP 分析

建立好健康分析数据集市，并对数据集各维度与事实表进行处理后，可快速对数据方中数据进行多维度分析，便于了解数据特征。如对某一类职业病危害人检查结果异常项目在年龄、接害工龄、企业经济类型等方面的分析比较，根据分析结果可进一步下钻或是上卷至不同数据层逐步做出分析，分析中可在听力异常、企业类型等不同纬度的不同层次间进行上卷与下钻，如在以听力正常、异常分析后，可进一步向下一层次不同分类听力异常进行下钻。表 2 显示了 2013 年度机构进行噪声作业人群各类听力异常分布特征情况的数据分析，可观察到工龄、年龄有随发生听力损失程度提高而增高的趋势，私营经济类型企业的听力异常检出高于其他两种类型，而异常检出率性别构成中男性显著高于女性。

表 2 噪声作业人群听力异常特征分析

分析维度	听力正常	高频听阈提高水平＜40 dBHL	高频听阈提高平均≥40 dBHL	高频及语频均提高	单侧耳听阈提高
平均工龄(年)					
私营	5.57	7.99	9.39	10.37	8.35
国有	8.20	10.96	10.83	13.85	10.84
外资	4.67	5.92	7.16	8.37	5.58
合计	5.55	7.25	8.25	10.68	7.47
平均年龄(岁)					
私营	31.92	34.71	38.70	40.69	34.53
国有	32.18	36.89	35.85	41.70	37.73
外资	28.17	30.46	32.30	35.52	31.30
合计	29.34	32.30	33.65	38.42	33.59
检出率(%)					
私营	73.21	19.01	5.68	1.08	1.01
国有	77.87	13.92	5.59	1.18	1.44
外资	79.33	14.14	5.07	0.53	0.92
男	76.32	15.86	6.09	0.74	0.99
女	90.28	6.87	0.79	0.66	1.40
合计	78.58	14.40	5.24	0.72	1.05

2.3 职业健康检查健康损害的影响因素分析

在数据集上利用决策树挖掘模型，可对不同职业健康损害

的影响因素进行分析，以噪声作业人员电测听检查不同异常结果为例进行影响因素分析发现，发生高频听力下降语频听力正常的关键影响因素包括 30～45 岁、46～59 岁、男性、血压偏高、家族高血压病史及有吸烟习惯者，其影响度由高至低依次为年龄、性别、血压。高频听力损失伴语频听力损失的关键影响因素包括 46～59 岁、伴有耳部疾患、血压偏高，其影响度由高至低依次为年龄、耳部疾病、血压。

2.4 职业健康检查异常结果关联规则分析

在数据集上以检查项目异常作为预测属性创建关联模型，可查找不同检查项目异常结果中的关联。以噪声作业人员异常项目为例进行关联规则分析，结果提示检查项目异常关联的预测规则有 16 条，与纯音听阈测试结果异常有关的置信度与支持度均较高的关联规则有：血压异常 + 心率异常⇒电测听异常、心电图异常⇒电测听异常、耳鼻喉异常⇒电测听异常等，分析中也存在提示肝胆脾 B 超异常⇒听力异常这样的置信度与支持度均较低的关联规则。

2.5 职业健康损害预测分析

使用数据集的时间维进行相应的聚合，可得到需分析的任意一项异常检出的时间序列数据，再使用时序预测模型加以分析可得到该类异常未来发生趋势的预测。图 3 是纯音听阈测试两类检查异常率（‰）未来 5 个月的趋势预测，提示听力异常检出率将在未来数月中趋于下降。

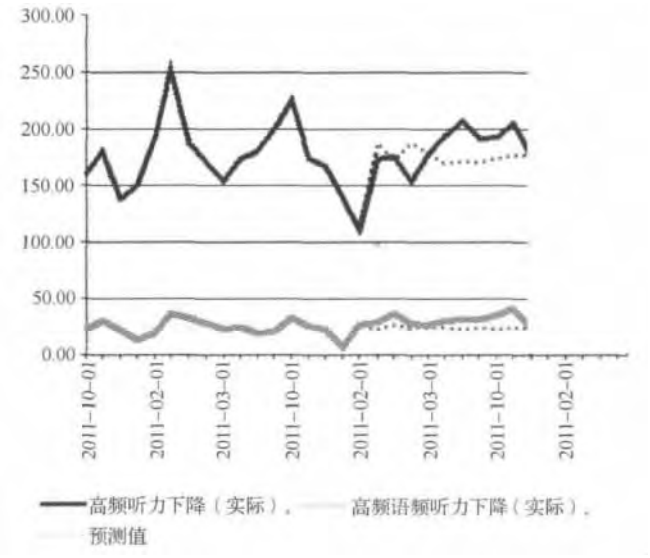

图 3 噪声作业听力异常检出率趋势预测

3 讨论

伴随着职业健康监护信息化进程的持续推进，越来越多的医疗机构将工作转移到了集成的信息系统服务平台上，这为开展信息分析提供了良好的数据基础，借助数据挖掘的相关技术手段分析相关因素，在更多区域、更大层面上对资料进行整合与利用，将是我国职业卫生信息化发展的趋势之一[2]。近年来，数据挖掘在医疗领域中的运用正逐步增多，对大量的临床数据进行计算机辅助挖掘就是当前医学信息利用与知识发现的一项研究热点[3]。 此外，利用数据挖掘技术进行医院管理信息的再处理及医疗质量的管理也是当前比较集中的应用方向[4,5]，在职业卫生领域中近年来也有在工程分析等专题上的应用[6]。而目前针对职业健康监护信息挖掘的工作尚少有探讨，职业健康监护数据不仅包含个人的健康检查结果，也有职业特征信息、个人相关信息数据，且这类信息多有良好层次性与延续性，易于数据挖掘利用。本研究详细分析了职业健康检查各类数据的结构层次与特征，以此为基础建立职业健康监护信息数据的星形架构数据仓库模型，并利用 OLAP 及数据挖掘的分析技术尝试了对职业健康监护信息进行分析，重构的多维数据集能高效地执行联机分析处理，快速响应各类复杂的分析要求并为数据挖掘提供数据基础。尽管健康监护数据库拥有较好的数据优势，但因传统关系型数据库的设计架构及原数据库系统开发时前瞻性的不足，要在其基础上进行数据分析挖掘并适合，仍有数据清洗、转换及模型重构等大量工作要做，而合理的模型结构及准确的数据，对数据挖掘的效率与发现信息的质量起着至关重要的作用。

在新的模型结构下应用数据挖掘模型进行知识探索，进一步提高了分析问题的方式与深度，在影响度分析及关联规则分析中提示噪声作业人群听力损害发生的相关知识，与传统流行病学分析的结论颇为一致[7]。此外时序模型进行异常检出率的预测分析中，虽预测值较实际值偏低并有所偏倚，但总体趋势较符合，其准确度与历史训练数据的数量有关。

本次研究的结果提示数据挖掘技术能较好地利用职业健康监护数据，在职业健康监护信息利用方面有着极强的实用价值与广阔的应用前景，其应用需解决好数据质量及模型结构等问题，在大数据上部署分析与挖掘技术将促进职业健康监护数据管理向着更高效率、更全面的决策支持与知识发现系统发展。

参考文献：

[1] 韩家炜. 数据挖掘概念与技术 [M]. 加拿大：机械工业出版社，2007：10.
[2] 杜燮祎，张敏. 我国职业卫生信息化发展趋势探讨 [J]. 中国安全生产科学技术，2010，6（4）：128-134.
[3] 余晖，张力新，刘文耀. 计算机辅助医学知识发现系统研究——糖尿病并发症流行病学数据挖掘 [J]. 生物医学工程学杂志，2008，(2)：295-299.
[4] 欧崇阳，曹宏伟，黄小琴，等. 基于 HIS 数据挖掘的医疗风险预警建模研究 [J]. 解放军医院管理杂志，2011，18（7）：626-627.
[5] 杨海青. 数据挖掘技术在医院管理中的应用 [J]. 中华医院管理杂志，2005，21（7）：497-499.
[6] 麦海明，罗海铭，常会友，等. 数据挖掘技术在职业卫生工程分析中的应用研究 [J]. 中国卫生工程学，2007，6（2）：70-72.
[7] 张维森，周浩，肖吕武，等. 噪声作业工人听力损伤与血压和高血压的相关性研究 [J]. 中华劳动卫生职业病杂志，2012，30（7）：517-520.