# MIMIC 数据库智能挖掘研究概述

张家艳[1],郑建立[1],郑西川[2],夏 涛[1]
(1. 上海理工大学,上海 200093;
2. 上海交通大学,上海 200233)

**摘 要**:开源数据库-重症特别护理信息集 MIMIC 数据库包含了大量的医学数据,自它发布之日起,便得到了众多研究人员的青睐。但低效的挖掘方法很难发现内部的隐含信息,这使得 MIMIC 数据库得不到很好的利用,造成了资源的浪费。探索新兴的挖掘方法进行知识发现便显得异常重要。文中对围绕 MIMIC 数据库的各种挖掘方法进行综述,重点阐述了新出现的机器学习和深度学习方法。同时将传统统计学模型与新出现的人工智能技术包括机器学习和深度学习技术进行比较分析。结果发现相比传统的统计学模型,机器学习和深度学习技术在预测病人的早期死亡率、发现疾病影响因素等方面普遍效果更好,这有助于改善医疗质量、帮助医生进行辅助诊断,在一定程度上也减少了病人的医疗费用。

**关键词**:MIMIC 数据库;人工智能;机器学习;深度学习;数据挖掘;医疗质量
**中图分类号**:TP392 **文献标识码**:A **文章编号**:1673-629X(2020)01-0144-05
doi:10.3969/j. issn. 1673-629X. 2020. 01. 026

# Application of Artificial Intelligence Technology in MIMIC Database Mining

ZHANG Jia-yan[1], ZHENG Jian-li[1], ZHENG Xi-chuan[2], XIA Tao[1]
(1. University of Shanghai for Science and Technology, Shanghai 200093, China;
2. Shanghai Jiaotong University, Shanghai 200233, China)

**Abstract**: The open source database, the medical information mart for intensive care (MIMIC), which contains a large amount of medical data, has been favored by many researchers since its release. However, inefficient mining methods are difficult to find internal hidden information, which makes the MIMIC database not well utilized and causes resource waste. It is extremely important to explore emerging mining methods for knowledge discovery. We summarize the various mining methods around the MIMIC database, focusing on emerging machine learning and deep learning methods. At the same time, the traditional statistical model is compared with the emerging artificial intelligence technologies including machine learning and deep learning. It was found that machine learning and deep learning generally perform better in predicting early mortality and finding factors affecting the disease than traditional statistical models, which helps improve the quality of medical care, assist doctors in diagnosis, and reduce the cost of medical care for patients to some extent.

**Key words**: MIMIC database; artificial intelligence; machine learning; deep learning; data mining; medical quality

## 0 引 言

数据挖掘也称作数据库的知识发现(knowledge discovery in databases, KDD)[1],目的是从大量的数据中抽取出有价值的知识。医院数字系统普及产生大量医疗数据,挖掘分析这些医疗数据能够发现相关规律。Ghassemi[2]等使用数据挖掘发现在入院前服用血清素摄取抑制剂或血清去甲状腺素摄取抑制剂的 ICU 住院病人比一般病人有更高的住院死亡率。

近年来,随着机器学习、深度学习的兴起,将这些算法用到医学领域,能改善挖掘结果。Wu C 运用决策树可视化方法发现了老年焦虑病人的影响因素[3]。但数据集过少时,用人工智能技术挖掘结果有时并不理想。深度学习适合数据量和数据维度比较大的情况[4],以至于业界流传一句话为得数据者得天下。而医疗领域,由于医学数据的私密性,研究人员更难获取大量的医学数据。为解决数据量少的问题,文中研究

**收稿日期**:2019-03-05 **修回日期**:2019-07-08 **网络出版时间**:2019-09-25
**基金项目**:上海市科委科技支撑项目(15441900604)
**作者简介**:张家艳(1995-),女(苗族),硕士研究生,研究方向为医学信息处理;郑建立,博士,副教授,通讯作者,研究方向为医学信息集成;郑西川,高级工程师,研究方向为影像电子病历、区域信息集成。
**网络出版地址**:http://kns. cnki. net/kcms/detail/61. 1450. TP. 20190925. 1520. 018. html

的数据集为由贝斯以色列女执事医疗中心和麻省理工大学计算生理实验室和飞利浦共同支持的重症监护医学信息集(MIMIC-III)。MIMIC 包含了在 2001 年到 2012 年间 53 423 个进入重症监护病房的成年病人(年龄在 16 岁以上),以及在 2001 年到 2008 年间的 7 870 名新生儿的数据[5]。

对拥有庞大数据集的 MIMIC 数据库进行挖掘,人工智能技术便能发挥巨大的优势。文中旨在介绍围绕 MIMIC 数据库的内容和研究、深度学习及机器学习在 MIMIC 数据库挖掘研究的应用领域和不足。

## 1　MIMIC 数据库简介

最近发布的 MIMIC 版本是 MIMIC-III(medical information mart for intensive care),version1.4,它是在 MIMIC-II 基础上的扩展。MIMIC-II 包括在 2001 到 2008 年之前几乎所有进入贝斯以色列女执事医疗中心重症监护病房的成年患者[6]。在数据库数据整合进 MIMIC 数据库之前,需依据美国的 HIPAA 标准进行去身份化处理[7],进行结构化数据清洗和数据转换。每个病人的住院日期随机转换成了 2100 年到 2200 年期间。在 HIPAA 规则下,这些病人出现在数据库中的年龄都超过了 300 年。

MIMIC-III 是一个由 26 张表组成的关系数据库。表通过标识符连接,通常会有 ID 后缀。例如:SUBJECT_ID 是指一个单独的病人。像备注、实验室测试和液平衡等事件信息都存储在事件表中,例如 OUTPUTEVENTS 表包含了与患者输出相关的所有测量值,而 LABEVENTS 表中包含了一个患者实验室测量结果。前缀有'D_'的表是字典表,包含标识符的定义。具体可查看 http://mimic.physionet.org/mimictables。

MIMIC 数据库免费开放给大众,但在获取数据库之前需签署数据使用协议,完成相应题目。在 2012 年末,已经超过 500 个用户得到批准使用。获取 MIMIC 关系数据库的两个工具为:基于网上的 QueryBuilder 和可下载的虚拟机(VM)镜像[8]。QueryBuilder 可以让使用者使用结构化查询语句(sql)在电脑或者移动端的 web 浏览器查询自己想要的数据,查询后的结果数据集以 CSV 的形式输出。但为了防止用户过度消耗 QueryBuilder 上的共享资源,MIMIC-III,v1.4 数据库系统设置每次查询仅返回前 5 000 行数据,查询中运行时间不得超过 15 分钟,超过了将显示超时,且不返回结果。具体可查看官方文档[9]。由于 MIMIC 数据库使用者的增多和 QueryBuilder 的一些限制,官网提供了可供下载的虚拟机(VM),让用户在自己的计算机上运行关系数据库副本。

## 2　围绕 MIMIC 数据库的数据挖掘

自 MIMIC 开始发布至今,人们围绕数据库做了不同主题的挖掘研究,也采用了各种挖掘方法对 MIMIC 数据库进行研究,下面分别对这些方法进行介绍。

### 2.1　传统的数据库挖掘研究方法

开始人们采用统计分析的方法对 MIMIC 数据库的数据进行挖掘研究。采用像 Simplified Acute Physiology Score (SAPS)[10]、Acute Physiology and Chronic Health Evaluation (APACHE)[11]、Sequential Organ Failure Assessment (SOFA)[12]等重大疾病计分系统和它们的改进版本来预测结果。SAPS 和 SOFA 的 AUROCs 能达到 0.658(±0.1)和 0.633(±0.09)[13]。相对于未加处理的 ICU 数据,SAPS 和 SOFA 达到的效果还是比较可取的。

### 2.2　采用机器学习方法对 MIMIC 数据库挖掘的研究

随着机器学习的出现,机器学习被用于挖掘研究。机器学习是计算机科学的人工智能领域,该方法能够让计算机自己学习相关特征[14]。在机器学习模型中,每个模型都有其适合的场合。支持向量机最原始的目的就是用于二分类,在二分类问题中,K. M. D. M. Karunarathna[15]比较了几种机器学习模型的优劣,结果支持向量机比其他模型有更高的精度。G. Khalili-Zadeh-Mahani 等[16]对五种分类技术进行比较,发现在下消化道出血患者中,支持向量机方法有较好的灵敏度和类别加权精确度。Aya Awad 等[17]引入集成学习方法,使用了集成学习随机森林、预测决策树、概率贝叶斯和基于规则的射影自适应共振理论模型,发现随机森林具有更高的精确率。这些机器学习模型的表现都要优于传统方法。Joshua Parreco 等[18]将梯度提升决策树与传统方法进行比较,发现机器学习方法的 AUCs 最大。Aya Awad 等[17]提出方法的结果优于如 SOFA 等标准计分系统。表 1 对上述研究人员所推崇的模型的挖掘结果进行了详细的展示。

### 2.3　采用深度学习方法对 MIMIC 数据库挖掘的研究

随着信息时代来临,数据量变得越来越大,传统的浅层机器学习方法已无法更好地处理大数据,深度学习就此产生。深度学习模仿了生物神经系统间的信息交流,利用人工神经网络来抽取简单的特征。

与现有的机器学习模型相比,大多数深度学习得到的结果都比较好。文献[4]将自归一化神经网络(SNN)、SAPS、SOFA、LR 计分、随机森林、广义加性模型、贝叶斯自适应回归树、超学习方法的预测结果进行比较,最后发现 SNN 的 AUROC 是所有模型中最高的。文献[19]引进一个新的深度学习模型叫做 GRU-D,

最后得到的 AUC 分数是所有模型中最高的。Gehrmann 等[20]研究人员比较了卷积神经网络(CNNs)和其他常用模型的概念抽取方法。在大多数任务中,CNN 表现都优于概念抽取方法,在 F1-score 中上升了 26,在 ROC 曲线中上升了 7%。S. Nemati 等[21]采用了深度强化学习的方法,从回顾性数据学习到的序列模型算法的结果比临床指南期望的结果更好。表 2 对每个模型的预测任务和结果进行了展示。

表 1 机器学习模型应用评估

| Model | Prediction task | Evaluation result(Accurac y,AUROC,CWA,AUCS) |
|---|---|---|
| SVM+Clustering data | mortality | 79.8%(平均 Accuracy) |
| ensemble learning random forest | Early mortality | 0.82(AUROC) |
| SVM | Unnecessary lab tests | 0.879(CWAw=0.9,upper gastrointestinal) |
| FuzzyTS | Unnecessary lab tests | 0.905(CWAw=0.9,lower gastrointestinal) |
| gradient-boosted decision trees | Prolonged mechanical ventilation and tracheostomy placement | 0.82(AUCs,prolonged mechanical ventila tion)<br>0.83(AUCs,tracheostomy placement) |

表 2 深度学习模型评估

| Model | Prediction task | Evaluation result(AUROC、AUC、F1) |
|---|---|---|
| Deep Rule-Based Fuzzy System (DRBFS) | Mortality | 73.9%(AUROC) |
| Self-Normalizing Neural Networks (SNN) | Mortality | 0.86(AUROC) |
| GRU-D(Gated Recurrent Unit) | Mortality | 0.852 7±0.003(AUC) |
| Convolutional Neural Networks (CNNs) | Multiple predictive tasks such as heart disease | 0.91(AUC)<br>0.74(F1) |

### 2.4 采用结合模型的方法对 MIMIC 数据库挖掘的研究

单个模型都有各自的缺点,结合模型综合了这些模型的优点来避免模型的缺点。Sanjay Purushotham 等[22]将 multilayer feedforward network (FNN)和 recurrent neural networks (RNN)两种深度模型相结合,该方法比其他方法的预测结果要好。J. Venugopalan 等[23]结合了逻辑回归和前馈神经网络模型的静态模型和条件随机域的暂态模型,组合模型的结果比单个模型的表现要好。表 3 展示了这些组合模型的评估结果和任务。

表 3 组合模型应用评估

| Model | Prediction task | Evaluation result(AUROC、Accuracy) |
|---|---|---|
| combination of FFN and RNN | in-hospital mortality | 0.873 0±0.006 5(AUROC) |
| Logistic regression andfeedforward neural networks+ conditional random fields | Multiple predictive tasks such as mortality | 0.95±0.001(Accuracy) |

### 2.5 其 他

目前,除了采用上述方法对数据库数据进行挖掘分析之外,还有一些其他的方法。Alharbi 等[24]通过过程挖掘模型得到比较好的结果。文献[25]引进存活主题模型更好地显示了病人状况。文献[26]提出了一种暂态数据挖掘方法,运用 SW-MATFD 挖掘者挖掘重症监护领域的临床数据。Z. He 等[27]采用 ICD-9-CM 编码算法,对老年人口进行分类。关联规则能够在大量的数据中发现有趣的关联关系,转化成供人决策的知识。C. Cheng 等[28]首次在 ICU 中将关联规则运用到 CDSS(clinical decision support system)中。

## 3 围绕 MIMIC 数据库的挖掘应用

### 3.1 死亡率预测

现存文献中,对 MIMIC 进行数据挖掘的一个常见应用领域就是预测死亡率,包括住院死亡率、入院初期死亡率等。

预测 ICU 病人死亡率能够改善医生治疗效果。文献[15]中通过识别病人死亡的独立因子来预测 ICU 病人的死亡率。文献[17]预测了入院初期的 24 小时内的死亡率。J. Venugopalan[23]也通过处理混合的暂态数据和静态数据来预测 ICU 病人死亡率。

### 3.2 优化药物用量

在临床中,有些药物的用量有着严格的要求,一旦取量不精确,将会导致无法预计的后果。一些研究人员挖掘研究 MIMIC 数据库数据得到优化的推荐用量。S. Nemati 等[21]通过对大量电子病历数据中样品剂量试验和相关结果进行学习,得到一个优化的肝素剂量策略。该推荐肝素用量的结果比临床指南期望的结果更好。

### 3.3 电子病历提取语义分析

将 MIMIC 出院小结里的语义信息提取出来,有利

于下一步的临床决策。Gehrmann 等[20]对和医疗状况相关的各种短语进行识别和突出。Sanjay Purushotham[22]也采用了其他方法进行 ICD-9code 分类预测。文献[29]对病例信息进行分析,发现病人积极情感,从而监控病人心理健康状况。Alharbi 等[24]对病例信息进行处理,发现一些不易发现的隐藏过程。

### 3.4 其　他

除了上述应用方面,还有一些方面会围绕 MIMIC 挖掘研究。文献[23]对 ICU 病人进行了再入院预测。文献[19]引入了一个新的学习模型来处理多元时间序列缺失值的问题。医生关注的不只是患者的死亡率,还有出院率,文献[25]采用了一种模型来预测病人的出院率。M. Dunitz 等[30]开发一种实时的算法将感染性病人分成不同的风险类别来进行感染性休克研究。Z. He[27]研究发现老年人口患的并发症和现在临床研究相对较少的矛盾,从而指导人们花更多的精力开展这方面的研究。

## 4 工作进展

由于对 MIMIC 数据库的挖掘研究改善了医疗服务,但这些数据毕竟是国外的,有些并不一定适合国内人群体质,在对 MIMIC 数据库进行充分的学习研究及参考相关论文之后,采用某三甲医院数据中心的数据参考 MIMIC 数据库建库的技术手段建立数据仓库。

在建立数据仓库之前,首先需要分析数据仓库的主要用途,确定相应的表结构。目前已经确定了大致的表结构。具体会进行进一步的分析完全确定。确定结构之后,就会对医院的数据进行抽取、清洗、转换,进入数据仓库。

数据抽取的工作难点主要在于医院数据中心数据库比较多,数据库下面的表也比较多,而且有些数据库没有相应的数据字典,对于有些字段的含义就只能靠猜测加验证,从如此庞杂的表中找到所需要的数据是一个费时的过程,还需要将得到的数据抽取转换出来。目前确定的数据抽取工具是 kettle,该工具是一款国外开源的 etl 工具,使用比较方便。

在建好数据仓库之后,会对数据库进行相应的挖掘研究,以期发现一些隐藏的医学信息。

## 5 结束语

MIMIC 数据库包含着丰富的临床信息,对其进行挖掘研究,发现其中隐含的疾病关系,能够改善医疗质量。文中简要介绍了 MIMIC 数据库,描述了现今对 MIMIC 数据库进行挖掘研究的方法以及在医学各个领域的应用,其中着重描述了基于人工智能技术机器学习及深度学习对 MIMIC 数据库进行挖掘研究。

目前机器学习、深度学习对 MIMIC 数据库信息的挖掘分析研究的领域比较广泛,比如各种疾病的预测、对缺失数据的处理、提取电子病历的语义信息等等。尤其是近年来的论文中,已经很少有研究人员采用传统的计分系统去发现数据库中的医学数据规律。一大批的研究人员都采用人工智能的方法进行挖掘研究,也取得了相对可观的结果,技术手段也相对越来越成熟。

虽然将人工智能技术(机器学习、深度学习等)用于 MIMIC 数据库挖掘分析已经硕果累累,但是从技术上看,也都存在各自的缺陷。首先机器学习对于小数据集会比较好,对于大规模的数据集,最好使用深度学习。其次由于深度学习对于深层网络的不可解释性,很难调整深层网络来得到一个较好的结果。在文献[26]中,在一些测试数据集中得到的结果反而不如统计机器学习得到的结果好。而且从应用上看,挖掘分析主要集中于死亡率预测和电子病历提取语义分析相关的方面,集中领域比较单一,挖掘应用的广度和深度不够,没有充分应用 MIMIC 数据库的丰富资源。

然而机器学习和深度学习方法的结合模型能够结合各个模型的优点,得到更好的结果,具有较大的发展潜力。但是现今结合模型在 MIMIC 数据库挖掘研究应用还较少,研究的领域还比较窄。在将来的工作中,首先可以在 MIMIC 挖掘研究中更多地使用结合模型。其次应该扩大应用领域,而不仅仅关注死亡率预测那几个方向,大胆应用到医疗的其他领域。最后,应该注重挖掘研究的深度,发现更多的隐含信息。

**参考文献:**


[1] SRIKANT R, AGRAWAL R. Mining sequential patterns: generalizations and performance improvements[C]//Proceedings of 5th conference on extended database technology (EDBT'96). [s. l.]: Springer-Verlag, 1996: 3-17.

[2] GHASSEMI M, MARSHALL J, SINGH N, et al. Leveraging a critical care database: selective serotonin reuptake inhibitor use prior to ICU admission is associated with increased hospital mortality[J]. CHEST, 2014, 145(4): 745-752.

[3] WU Chunshan, WENG Yongqiang, JIANG Qiawei, et al. Applied research on visual mining technology in medical data [C]//4th international conference on cloud computing and intelligence systems. Beijing: IEEE, 2016.

[4] ZAHID M A H, JOON L. Mortality prediction with self normalizing neural networks in intensive care unit patients[C]// 2018 IEEE EMBS international conference on biomedical & health informatics. Las Vegas, NV, USA: IEEE, 2018.

[5] JOHNSON A E W, POLLARD T J, SHEN L, et al. MIMIC-III, a freely accessible critical care database[J]. Scientific Data, 2016, 3: 160035.

[6] LEE J, SCOTT D J, VILLARROEL M, et al. Open-access MIMIC-II database for intensive care research[C]//2011 annual international conference of the IEEE engineering in medicine and biology society. Boston, MA: IEEE, 2011: 8315-8318.

[7] SAEED M, VILLARROEL M, REISNER A T, et al. Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database[J]. Critical Care Medicine, 2011, 39(5): 952-960.

[8] SCOTT D J, LEE J, SILVA I, et al. Accessing the public mimic-ii intensive care relational database for clinical research[J]. BMC Medical Informatics and Decision Making, 2013, 13(1): 9.

[9] MIMIC-III critical care database: doucumentation and website [EB/OL]. 2018. http://mimic. physionet. org.

[10] LE GALL J R, LOIRAT P, ALPEROVITCH A, et al. A simplified acute physiology score for ICU patients[J]. Critical Care Medicine, 1984, 12(11): 975-977.

[11] KNAUS W A, ZIMMERMAN J E, WAGNER D P, et al. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system[J]. Critical Care Medicine, 1981, 9(8): 591-597.

[12] VINCENT J L, MORENO R, TAKALA J, et al. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure[J]. Intensive Care Medicine, 1996, 22(7): 707-710.

[13] LEE J, MASLOVE D M, DUBIN J A. Personalized mortality prediction driven by electronic medical data and a patient similarity metric[J]. PLoS One, 2015, 10(5): e0127428.

[14] PARRECO J P, HIDALGO A E, BADILLA A D, et al. Predicting central line-associated bloodstream infections and mortality using supervised machine learning[J]. Journal of Critical Care, 2018, 45: 156-162.

[15] KARUNARATHNA K M D M. Predicting ICU death with summarized patient data[C]//8th annual computing and communication workshop and conference (CCWC). [s. l.]: IEEE, 2018.

[16] KHALILI-ZADEH-MAHANI G, ZARE-MIRAKABAD M R, DERHAMI V. Necessity of laboratory blood tests in intensive care unit using data mining[C]//2015 5th international conference on computer and knowledge engineering (ICCKE). Mashhad: IEEE, 2015: 176-180.

[17] AWAD A, BADER-EL-DEN M, MCNICHOLAS J, et al. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach[J]. International Journal of Medical Informatics, 2017, 108: 185-195.

[18] PARRECO J, HIDALGO A, PARKS J J, et al. Using artificial intelligence to predict prolonged mechanical ventilation and tracheostomy placement[J]. Journal of Surgical Research, 2018, 228: 179-187.

[19] CHE Z, PURUSHOTHAM S, CHO K, et al. Recurrent neural networks for multivariate time series with missing values[J]. Scientific Reports, 2018, 8: 6085-6097.

[20] GEHRMANN S, DERNONCOURT F, LI Y, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives[J]. PLOS One, 2018, 13(2): e0192360.

[21] NEMATI S, GHASSEMI M M, CLIFFORD G D. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach[C]//2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC). Orlando, FL: IEEE, 2016: 2978-2981.

[22] PURUSHOTHAM S, MENG Chuizheng, CHE Zhengping, et al. Benchmarking deep learning models on large healthcare datasets[J]. Journal of Biomedical Informatics, 2018, 83: 112-134.

[23] VENUGOPALAN J, CHANANI N, MAHER K, et al. Combination of static and temporal data analysis to predict mortality and readmission in the intensive care[C]//39th annual international conference of the IEEE engineering in medicine and biology society (EMBC). South Korea: IEEE, 2017.

[24] ALHARBI A, BULPITT A, JOHNSON O A. Towards unsupervised detection of process models in healthcare[J]. Studies in Health Technology and Informatics, 2018, 247: 381-385.

[25] ZHANG Y, JIANG R, PETZOLD L. Suevival topic models for predicting outcomes for trauma patients[C]//2017 IEEE 33rd international conference on data engineering (ICDE). San Diego, CAL: IEEE, 2017: 1497-1504.

[26] COMBI C, MANTOVANI M, SALA P. Discovering quantitative temporal functional dependencies on clinical data[C]//2017 IEEE international conference on healthcare informatics (ICHI). Park City, UT: IEEE, 2017: 248-257.

[27] HE Z, CHARNESS N, BIAN J, et al. Assessing the comorbidity gap between clinical studies and prevalence in elderly patient populations[C]//2016 IEEE-EMBS international conference on biomedical and health informatics (BHI). Las Vegas, NV: IEEE, 2016: 136-139.

[28] CHENG C W, CHANANI N, VENUGOPALAN J, et al. icuARM-an ICU clinical decision support system using association rule mining[J]. IEEE Journal of Translational Engineering in Health and Medicine, 2013, 1: 4400110.

[29] GHASSEMI M M, MARK R G, NEMATI S. A visualizat-ion of evolving clinical sentiment using vector representations of clinical notes[C]//2015 computing in cardiology conference (CinC). Nice: IEEE, 2015: 629-632.

[30] DUNITZ M, VERGHESE G, HELDT T. Predicting hyperlactatemia in the MIMIC II database[C]//2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). Milan: IEEE, 2015: 985-988.