(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110051324 A
(43)申请公布日 2019.07.26

(21)申请号 201910194628.2

(22)申请日 2019.03.14

(71)申请人 深圳大学
地址 518060 广东省深圳市南山区南海大道3688号
申请人 南方医科大学深圳医院

(72)发明人 黄炳升 梁栋 刘勇 邹儒诗 黄树华 余夏夏

(74)专利代理机构 广州嘉权专利商标事务所有限公司 44205
代理人 胡辉 黎扬鹏

(51)Int.Cl.
A61B 5/00(2006.01)
A61B 5/08(2006.01)

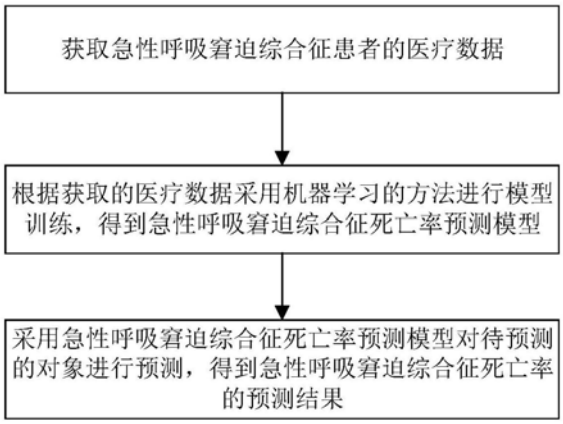
权利要求书2页 说明书11页 附图3页

(54)发明名称

一种急性呼吸窘迫综合征死亡率预测方法及系统

(57)摘要

本发明公开了一种急性呼吸窘迫综合征死亡率预测方法及系统，方法包括：获取急性呼吸窘迫综合征患者的医疗数据；根据获取的医疗数据采用机器学习的方法进行模型训练，得到急性呼吸窘迫综合征死亡率预测模型；采用急性呼吸窘迫综合征死亡率预测模型对待预测的对象进行预测，得到急性呼吸窘迫综合征死亡率的预测结果。本发明通过机器学习的方法训练出急性呼吸窘迫综合征死亡率预测模型，再采用急性呼吸窘迫综合征死亡率预测模型来预测ARDS患者的死亡率，将机器学习应用于ARDS患者死亡率预测上，能通过机器学习训练的模型准确和客观地预测出ARDS患者的死亡率，为临床医师提供了更有效和可行的预测信息，可广泛应用于医学数据挖掘领域。



CN 110051324 A

1.一种急性呼吸窘迫综合征死亡率预测方法,其特征在于:包括以下步骤:

获取急性呼吸窘迫综合征患者的医疗数据;

根据获取的医疗数据采用机器学习的方法进行模型训练,得到急性呼吸窘迫综合征死亡率预测模型;

采用急性呼吸窘迫综合征死亡率预测模型对待预测的对象进行预测,得到急性呼吸窘迫综合征死亡率的预测结果。

2.根据权利要求1所述的一种急性呼吸窘迫综合征死亡率预测方法,其特征在于:所述获取急性呼吸窘迫综合征患者的医疗数据这一步骤,具体为:

从MIMIC-III数据库中下载急性呼吸窘迫综合征患者的医疗数据。

3.根据权利要求1所述的一种急性呼吸窘迫综合征死亡率预测方法,其特征在于:所述根据获取的医疗数据采用机器学习的方法进行模型训练,得到急性呼吸窘迫综合征死亡率预测模型这一步骤,具体包括:

对获取的医疗数据进行预处理,所述预处理包括样本筛选和特征提取;

根据预处理后的数据采用机器学习的方法进行模型训练,得到急性呼吸窘迫综合征死亡率预测模型,所述急性呼吸窘迫综合征死亡率预测模型包括住院死亡率预测模型、30天死亡率预测模型和一年死亡率预测模型。

4.根据权利要求3所述的一种急性呼吸窘迫综合征死亡率预测方法,其特征在于:所述对获取的医疗数据进行预处理这一步骤,具体包括:

按照纳入标准和排除标准对获取的医疗数据进行样本筛选,得到筛选好的样本,所述纳入标准包括年龄大于等于18周岁的入住重症监护室且经柏林标准诊断为急性呼吸窘迫综合征的患者,所述排除标准包括MIMIC-III数据库中数据记录不完整的数据、年龄小于18周岁的患者、采用姑息疗法的患者和ICU记录时间小于48小时的患者中的任意一个;

对筛选好的样本提取每个样本用于建模的变量特征。

5.根据权利要求4所述的一种急性呼吸窘迫综合征死亡率预测方法,其特征在于:所述对获取的医疗数据进行预处理这一步骤,还具体包括以下步骤:

对获取的医疗数据中的缺失数据进行多重插补。

6.根据权利要求3所述的一种急性呼吸窘迫综合征死亡率预测方法,其特征在于:所述根据预处理后的数据采用机器学习的方法进行模型训练,得到急性呼吸窘迫综合征死亡率预测模型这一步骤,具体包括:

根据患者生存天数对预处理后的数据进行分类,分别得到3个死亡率预测模型的阳性组和阴性组,所述3个死亡率预测模型包括住院死亡率预测模型、30天死亡率预测模型和一年死亡率预测模型;

分别对3个死亡率预测模型的阳性组和阴性组进行组间分析,筛选出组间差异显著的特征;

根据组间差异显著的特征采用随机森林算法建立急性呼吸窘迫综合征死亡率预测模型。

7.根据权利要求6所述的一种急性呼吸窘迫综合征死亡率预测方法,其特征在于:所述3个死亡率预测模型的阳性组和阴性组具体为:住院死亡率预测模型的阳性组是住院内死亡的患者数据,住院死亡率预测模型的阴性组是住院期间存活的患者数据;30天死亡率预

测模型的阳性组是住院后30天内死亡的患者数据,30天死亡率预测模型的阴性组是住院后30天内存活的患者数据;一年死亡率预测模型的阳性组是住院后一年内死亡的患者数据,一年死亡率预测模型的阴性组是住院后一年内存活的患者数据。

8.根据权利要求6所述的一种急性呼吸窘迫综合征死亡率预测方法,其特征在于:所述根据组间差异显著的特征采用随机森林算法建立急性呼吸窘迫综合征死亡率预测模型这一步骤,具体包括:

采用K折交叉验证法将组间差异显著的特征划分为第一训练集和测试集;

采用K折交叉验证法将第一训练集划分为第二训练集和验证集;

根据第二训练集和验证集采用网格寻优的方法寻找出最佳的模型参数,进而根据最佳的模型参数采用随机森林算法构建若干个急性呼吸窘迫综合征死亡率预测模型;

分别采用各个急性呼吸窘迫综合征死亡率预测模型对测试集进行测试,得到各折的预测结果;

对各折的预测结果求平均值,得到各个急性呼吸窘迫综合征死亡率预测模型对应的预测性能结果;

根据得到的预测性能结果,从各个急性呼吸窘迫综合征死亡率预测模型中选择预测性能结果最好的模型作为最终的急性呼吸窘迫综合征死亡率预测模型。

9.一种急性呼吸窘迫综合征死亡率预测系统,其特征在于:包括以下模块:

获取模块,用于获取急性呼吸窘迫综合征患者的医疗数据;

训练模块,用于根据获取的医疗数据采用机器学习的方法进行模型训练,得到急性呼吸窘迫综合征死亡率预测模型;

预测模块,用于采用急性呼吸窘迫综合征死亡率预测模型对待预测的对象进行预测,得到急性呼吸窘迫综合征死亡率的预测结果。

10.一种急性呼吸窘迫综合征死亡率预测系统,其特征在于:包括:

至少一个处理器;

至少一个存储器,用于存储至少一个程序;

当所述至少一个程序被所述至少一个处理器执行,使得所述至少一个处理器实现如权利要求1-8任一项所述的急性呼吸窘迫综合征死亡率预测方法。

一种急性呼吸窘迫综合征死亡率预测方法及系统

技术领域

[0001] 本发明涉及医学数据挖掘领域，尤其是一种急性呼吸窘迫综合征死亡率预测方法及系统。

背景技术

[0002] 急性呼吸窘迫综合征(Acute Respiratory Distress Syndrome，ARDS)是一种常见危重症，是指与暴露于危险因素有关的急性弥漫性肺损伤，常常伴随着肺部炎症导致的肺血管通透性增加以及含气肺组织减少。在临床上各种危重症的患者均存在发生ARDS的潜在风险，且发生ARDS后的住院死亡率在34.9%-46.1%之间，严重威胁重症患者的生命并影响其生存质量。因此建立ARDS患者死亡率预测模型，可以区分病情严重程度，从而决定不同的救治策略；同时利用预测模型可以探究各种变量对患者死亡率的影响因素，对提高患者生存率有着积极的意义。

[0003] 当前存在的预测ICU患者的生存率(probability of survival)的方法，都基于传统分析方法，包括APACHE II、OSI、OI以及LIS分析等。这些传统分析方法通常在一个或多个医学中心收集数据，再基于疾病专家的经验和统计方法(最常用的是逻辑回归)得到相关的变量，最后通过所得变量去构建并验证预测模型。然而这类方法存在如下问题：(1)由专家经验或统计分析得到的变量，会存在主观性与数据偏差；(2)影响ARDS发生与发展的因素极为复杂，很难结合多维变量做统计分析；(3)这些方法并非专为ARDS设计，目前尚未存在有效的评分模型适用于预测ARDS患者的死亡率。

发明内容

[0004] 为解决上述技术问题，本发明的目的在于：提供一种客观和准确的急性呼吸窘迫综合征死亡率预测方法及系统。

[0005] 本发明一方面所采取的技术方案是：

[0006] 一种急性呼吸窘迫综合征死亡率预测方法，包括以下步骤：

[0007] 获取急性呼吸窘迫综合征患者的医疗数据；

[0008] 根据获取的医疗数据采用机器学习的方法进行模型训练，得到急性呼吸窘迫综合征死亡率预测模型；

[0009] 采用急性呼吸窘迫综合征死亡率预测模型对待预测的对象进行预测，得到急性呼吸窘迫综合征死亡率的预测结果。

[0010] 进一步，所述获取急性呼吸窘迫综合征患者的医疗数据这一步骤，具体为：

[0011] 从MIMIC-III数据库中下载急性呼吸窘迫综合征患者的医疗数据。

[0012] 进一步，所述根据获取的医疗数据采用机器学习的方法进行模型训练，得到急性呼吸窘迫综合征死亡率预测模型这一步骤，具体包括：

[0013] 对获取的医疗数据进行预处理，所述预处理包括样本筛选和特征提取；

[0014] 根据预处理后的数据采用机器学习的方法进行模型训练，得到急性呼吸窘迫综合

征死亡率预测模型，所述急性呼吸窘迫综合征死亡率预测模型包括住院死亡率预测模型、30天死亡率预测模型和一年死亡率预测模型。

[0015] 进一步，所述对获取的医疗数据进行预处理这一步骤，具体包括：

[0016] 按照纳入标准和排除标准对获取的医疗数据进行样本筛选，得到筛选好的样本，所述纳入标准包括年龄大于等于18周岁的入住重症监护室且经柏林标准诊断为急性呼吸窘迫综合征的患者，所述排除标准包括MIMIC-III数据库中数据记录不完整的数据、年龄小于18周岁的患者、采用姑息疗法的患者和ICU记录时间小于48小时的患者中的任意一个；

[0017] 对筛选好的样本提取每个样本用于建模的变量特征。

[0018] 进一步，所述对获取的医疗数据进行预处理这一步骤，还具体包括以下步骤：

[0019] 对获取的医疗数据中的缺失数据进行多重插补。

[0020] 进一步，所述根据预处理后的数据采用机器学习的方法进行模型训练，得到急性呼吸窘迫综合征死亡率预测模型这一步骤，具体包括：

[0021] 根据患者生存天数对预处理后的数据进行分类，分别得到3个死亡率预测模型的阳性组和阴性组，所述3个死亡率预测模型包括住院死亡率预测模型、30天死亡率预测模型和一年死亡率预测模型；

[0022] 分别对3个死亡率预测模型的阳性组和阴性组进行组间分析，筛选出组间差异显著的特征；

[0023] 根据组间差异显著的特征采用随机森林算法建立急性呼吸窘迫综合征死亡率预测模型。

[0024] 进一步，所述3个死亡率预测模型的阳性组和阴性组具体为：住院死亡率预测模型的阳性组是住院内死亡的患者数据，住院死亡率预测模型的阴性组是住院期间存活的患者数据；30天死亡率预测模型的阳性组是住院后30天内死亡的患者数据，30天死亡率预测模型的阴性组是住院后30天内存活的患者数据；一年死亡率预测模型的阳性组是住院后一年内死亡的患者数据，一年死亡率预测模型的阴性组是住院后一年内存活的患者数据。

[0025] 进一步，所述根据组间差异显著的特征采用随机森林算法建立急性呼吸窘迫综合征死亡率预测模型这一步骤，具体包括：

[0026] 采用K折交叉验证法将组间差异显著的特征划分为第一训练集和测试集；

[0027] 采用K折交叉验证法将第一训练集划分为第二训练集和验证集；

[0028] 根据第二训练集和验证集采用网格寻优的方法寻找出最佳的模型参数，进而根据最佳的模型参数采用随机森林算法构建若干个急性呼吸窘迫综合征死亡率预测模型；

[0029] 分别采用各个急性呼吸窘迫综合征死亡率预测模型对测试集进行测试，得到各折的预测结果；

[0030] 对各折的预测结果求平均值，得到各个急性呼吸窘迫综合征死亡率预测模型对应的预测性能结果；

[0031] 根据得到的预测性能结果，从各个急性呼吸窘迫综合征死亡率预测模型中选择预测性能结果最好的模型作为最终的急性呼吸窘迫综合征死亡率预测模型。

[0032] 本发明另一方面所采取的技术方案是：

[0033] 一种急性呼吸窘迫综合征死亡率预测系统，包括以下模块：

[0034] 获取模块，用于获取急性呼吸窘迫综合征患者的医疗数据；

[0035] 训练模块,用于根据获取的医疗数据采用机器学习的方法进行模型训练,得到急性呼吸窘迫综合征死亡率预测模型;

[0036] 预测模块,用于采用急性呼吸窘迫综合征死亡率预测模型对待预测的对象进行预测,得到急性呼吸窘迫综合征死亡率的预测结果。

[0037] 本发明另一方面所采取的技术方案是:

[0038] 一种急性呼吸窘迫综合征死亡率预测系统,包括:

[0039] 至少一个处理器;

[0040] 至少一个存储器,用于存储至少一个程序;

[0041] 当所述至少一个程序被所述至少一个处理器执行,使得所述至少一个处理器实现如本发明所述的急性呼吸窘迫综合征死亡率预测方法。

[0042] 本发明的有益效果是:本发明急性呼吸窘迫综合征死亡率预测方法及系统,获取急性呼吸窘迫综合征患者的医疗数据后,通过机器学习的方法训练出急性呼吸窘迫综合征死亡率预测模型,最后采用急性呼吸窘迫综合征死亡率预测模型来预测ARDS患者的死亡率,将机器学习应用于ARDS患者死亡率预测上,能通过机器学习训练的模型准确和客观地预测出ARDS患者的死亡率,为临床医师提供了更有效和可行的预测信息作为参考。

附图说明

[0043] 图1为本发明实施例提供的急性呼吸窘迫综合征死亡率预测方法流程图;

[0044] 图2为本发明实施例的数据筛选流程图;

[0045] 图3为随机森林算法示意图;

[0046] 图4为本发明实施例采用机器学习的方法进行死亡率预测模型训练的具体流程图;

[0047] 图5为本发明实施例最终三个死亡率预测模型的训练以及应用于临床上新数据预测的流程图。

具体实施方式

[0048] 首先对本发明所涉及到的名词及术语进行说明:

[0049] EHR:electronic health record,个人电子健康记录。

[0050] 柏林标准:Berlin definition,2011年ARDS定义工作组提出的诊断ARDS的通用标准。

[0051] PEEP:positive end expiratory pressure,呼气末正压。

[0052] APPS:plateau pressure,气道平台压。

[0053] CPAP:Continuous Positive Airway Pressure,持续正压通气。

[0054] SAPS:Simplified Acute Physiology Score,简化急性生理评分。

[0055] LIS:the Lung Injury Score,肺损伤分数。

[0056] APACHE:Acute Physiology and Chronic Health Evaluation,急性生理及慢性健康状况评分。

[0057] OI:Oxygenation index,氧合指数,计算方法:[PaO2/FiO2]。

[0058] OSI:Oxygenation Saturation Index,氧饱和度指数,计算方法:[FiO2/SpO2]。

[0059] ROC:Receiver Operating Characteristic Curve,接受者操作特征曲线。

[0060] AUROC:Area Under the Receiver Operating Characteristic Curve,接受者操作特征曲线下面积。

[0061] RF:random forest,随机森林,是机器学习(ML)的一种分类算法。

[0062] 住院死亡:in-hospital mortality,指ARDS患者在住院期间死亡,以入院信息记录时间为准。

[0063] 30天死亡:30-day mortality,指ARDS患者入院后30天内死亡。

[0064] 一年死亡:1-year mortality,指ARDS患者入院后一年内死亡。

[0065] 下面结合说明书附图和具体实施例对本发明作进一步解释和说明。

[0066] 参照图1,本发明实施例提供了一种急性呼吸窘迫综合征死亡率预测方法,包括以下步骤:

[0067] 获取急性呼吸窘迫综合征患者的医疗数据;

[0068] 根据获取的医疗数据采用机器学习的方法进行模型训练,得到急性呼吸窘迫综合征死亡率预测模型;

[0069] 采用急性呼吸窘迫综合征死亡率预测模型对待预测的对象进行预测,得到急性呼吸窘迫综合征死亡率的预测结果。

[0070] 具体地,急性呼吸窘迫综合征患者的医疗数据可以是从公共数据库(如MIMIC-III数据库等)下载的人口统计学资料、生命体征监测等医疗数据。待预测的对象是新的ARDS患者。

[0071] 机器学习是人工智能的一个分支,通过设计专门的算法令计算机自动完成数据分析以掌握规律(即“学习”),并利用规律对未知数据做出判断或预测。机器学习方法可通过不断“学习”来分析、掌握规律,也可轻而易举地完成信息的处理,相比于统计分析方法,机器学习在分析大数据量与高变量维度方面有着不可比拟的优势。机器学习的方法包括随机森林算法、支持向量机算法、深度学习算法等。

[0072] 本实施例应用了机器学习的方法,根据已有的ARDS患者医疗数据寻找ARDS患者死亡率的规律,得到ARDS患者死亡率预测模型,下次再有一些新的数据(即待预测的对象)就可以按照先前学习到的规律,让该预测模型自动预测ARDS患者的死亡率。

[0073] 本实施例使用机器学习训练模型,把更多更全的临床数据作为输入,受专家经验或统计分析的影响小,可以避免数据采集偏差,得到更多隐藏信息,从而提高预测的准确率;此外,机器学习可以在不同规格或尺度的数据中进行训练,可以发现一些新的具有预测价值或不具备价值的变量,为临床治疗提供额外的灵感。

[0074] 进一步作为优选的实施方式,所述获取急性呼吸窘迫综合征患者的医疗数据这一步骤,具体为:

[0075] 从MIMIC-III数据库中下载急性呼吸窘迫综合征患者的医疗数据。

[0076] 具体地,MIMIC-III是由MIT的physionet实验室、BIDMC(Beth Israel Deaconess MedicalCenter)和飞利浦公司共同建设的针对重症监护患者的数据库,目前数据库包含了53423名成年患者的医疗数据(其中ARDS患者数据有3186例),包括基本的人口统计学资料、生命体征监测等数据。

[0077] 本实施例利用了MIMIC-III中结构化的临床数据进行训练,有效避免了模糊不清

的临床定义和数据采集的偏差，提高了ARDS死亡率预测的准确率。

[0078] 进一步作为优选的实施方式，所述根据获取的医疗数据采用机器学习的方法进行模型训练，得到急性呼吸窘迫综合征死亡率预测模型这一步骤，具体包括：

[0079] 对获取的医疗数据进行预处理，所述预处理包括样本筛选和特征提取；

[0080] 根据预处理后的数据采用机器学习的方法进行模型训练，得到急性呼吸窘迫综合征死亡率预测模型，所述急性呼吸窘迫综合征死亡率预测模型包括住院死亡率预测模型、30天死亡率预测模型和一年死亡率预测模型。

[0081] 具体地，样本筛选，是为了筛选出符合当前ARDS诊断标准，且满足后续构建模型的要求，以及记录内容可以在医院ICU实现(确保模型可以实际应用)的样本。特征提取，是为了从样本中提取出用于进行模型训练的特征。

[0082] 进一步作为优选的实施方式，所述对获取的医疗数据进行预处理这一步骤，具体包括：

[0083] 按照纳入标准和排除标准对获取的医疗数据进行样本筛选，得到筛选好的样本，所述纳入标准包括年龄大于等于18周岁的入住重症监护室且经柏林标准诊断为急性呼吸窘迫综合征的患者，所述排除标准包括MIMIC-III数据库中数据记录不完整的数据、年龄小于18周岁的患者、采用姑息疗法的患者和ICU记录时间小于48小时的患者中的任意一个；

[0084] 对筛选好的样本提取每个样本用于建模的变量特征。

[0085] 具体地，柏林标准诊断为急性呼吸窘迫综合征的依据为：1)急性发作；2)PEEP(或CPAP)≥5cm H2O时，OI(PaO2/FiO2)<300mmHg；3)胸部影像双侧浸润影；4)呼吸衰竭无法用心力衰竭来完全解释。

[0086] 采用姑息疗法的患者，是指没有接受积极治疗的患者。

[0087] 提取的变量特征包括患者的年龄、性别、APACHE II分数等可以直接读取的数据，也包括机械通气时间、患者生理信息等变量特征。

[0088] 进一步作为优选的实施方式，所述对获取的医疗数据进行预处理这一步骤，还具体包括以下步骤：

[0089] 对获取的医疗数据中的缺失数据进行多重插补。

[0090] 具体地，多重插补是一种对缺失数据进行调整的方法，采用一系列可能的数据集来填充每一个缺失数据值，然后使用完全数据的标准去分析多重插补数据集，最后对这些分析结果归纳综合。本实施例可使用SPSS软件完成多重插补操作，对个别数据缺失的ARDS数据进行补充。

[0091] 进一步作为优选的实施方式，所述根据预处理后的数据采用机器学习的方法进行模型训练，得到急性呼吸窘迫综合征死亡率预测模型这一步骤，具体包括：

[0092] 根据患者生存天数对预处理后的数据进行分类，分别得到3个死亡率预测模型的阳性组和阴性组，所述3个死亡率预测模型包括住院死亡率预测模型、30天死亡率预测模型和一年死亡率预测模型；

[0093] 分别对3个死亡率预测模型的阳性组和阴性组进行组间分析，筛选出组间差异显著的特征；

[0094] 根据组间差异显著的特征采用随机森林算法建立急性呼吸窘迫综合征死亡率预测模型。

[0095] 具体地，组间分析可使用SPSS软件来进行。本实施例通过对样本特征的组间分析找出不同类别样本间差异显著的特征作为输入，可以降低数据冗余，减少模型计算量，找出更有意义的特征。本实施例分别对三个预测模型中的阳性组与阴性组进行组间分析，具体操作是：对于符合正态分布的连续变量采用t检验(Student t-test)分析，而非正态分布的连续变量采用非参数检验(Mann-Whitney U test)；对于离散变量则通过卡方检验(Chi2test)或Fisher's exact test进行分析。经过上述检验的P值小于0.05的变量特征可认为组间差异显著，则保留该特征；对于P值大于0.05的认为该特征在阳性与阴性的组间差异不明显，对预测模型的分类影响较小，可以删去。

[0096] 随机森林是利用样本对多棵决策树进行训练并预测样本结果的一种分类器，决策树的训练过程采用的是自上而下的递归方法，其基本思想是以信息熵为度量构建一颗熵值下降最快的树，直到叶子节点的熵值为零，此时每个叶子节点的样本都属于同一类别。当输入新样本时，随机森林中各决策树分别判断投票，得票数最多的便作为最终的分类结果。随机森林通过决策树的集成学习与多数投票机制，拥有较好的抗噪声能力和不易过拟合，能较好地对ARDS患者的死亡率进行预测。

[0097] 进一步作为优选的实施方式，所述3个死亡率预测模型的阳性组和阴性组具体为：住院死亡率预测模型的阳性组是住院内死亡的患者数据，住院死亡率预测模型的阴性组是住院期间存活的患者数据；30天死亡率预测模型的阳性组是住院后30天内死亡的患者数据，30天死亡率预测模型的阴性组是住院后30天内存活的患者数据；一年死亡率预测模型的阳性组是住院后一年内死亡的患者数据，一年死亡率预测模型的阴性组是住院后一年内存活的患者数据。

[0098] 进一步作为优选的实施方式，所述根据组间差异显著的特征采用随机森林算法建立急性呼吸窘迫综合征死亡率预测模型这一步骤，具体包括：

[0099] 采用K折交叉验证法将组间差异显著的特征划分为第一训练集和测试集；

[0100] 采用K折交叉验证法将第一训练集划分为第二训练集和验证集；

[0101] 根据第二训练集和验证集采用网格寻优的方法寻找出最佳的模型参数，进而根据最佳的模型参数采用随机森林算法构建若干个急性呼吸窘迫综合征死亡率预测模型；

[0102] 分别采用各个急性呼吸窘迫综合征死亡率预测模型对测试集进行测试，得到各折的预测结果；

[0103] 对各折的预测结果求平均值，得到各个急性呼吸窘迫综合征死亡率预测模型对应的预测性能结果；

[0104] 根据得到的预测性能结果，从各个急性呼吸窘迫综合征死亡率预测模型中选择预测性能结果最好的模型作为最终的急性呼吸窘迫综合征死亡率预测模型。

[0105] 具体地，采用随机森林算法构建的急性呼吸窘迫综合征死亡率预测模型可不止一个，故可根据各个预测模型的预测性能结果进行模型筛选。

[0106] 在K折交叉验证法的每一折使用训练集构建模型时，会使用网格寻优的方法来寻找最佳的模型参数(RF的参数包括决策树数目n_estimators、分枝标准criterion、最小叶子样本数min_sample_leaf等)。所以本实施例会把第一训练集再次划分为第二训练集与验证集两部分，循环测试每组参数的效果，选择AUROC效果最佳的一组参数构建随机森林分类器模型(即急性呼吸窘迫综合征死亡率预测模型)，对测试集进行测试，得到本折预测结果；

然后对K折的结果求平均值,得到各个模型的预测性能结果;最后选择预测性能结果最好的模型作为最终的死亡率预测模型。

[0107] 为了提高ARDS患者死亡率预测的准确率,为临床医师提供实时和可行的预测信息,本具体实施例提出了急性呼吸窘迫综合征死亡率预测方案。下面将对该方案的具体实现流程以及使用流程进行描述。

[0108] (一)具体实现流程

[0109] 本具体实施例的方案具体实现流程可分为三步:(1)数据收集及数据预处理;(2)预测模型建立与测试;(3)结果评估与比较。

[0110] 1.数据收集及数据预处理

[0111] 1.1数据来源

[0112] 本具体实施例从公共数据库MIMIC-III(Medical Information Mart for Intensive Care)中下载3186例ARDS患者数据。MIMIC-III是由MIT的physionet实验室、BIDMC(Beth IsraelDeaconess Medical Center)和飞利浦公司共同建设的针对重症监护患者的数据库,目前数据库包含了53423名成年患者的医疗数据,包括基本的人口统计学资料、生命体征监测等数据。

[0113] 1.2数据预处理

[0114] 本具体实施例按照图2所示的纳入标准和排除标准对ARDS患者数据进行筛选,以选出符合当前ARDS诊断标准,且满足后续构建模型的要求,以及记录内容可以在合作医院ICU实现(确保模型可以实际应用)的样本。

[0115] 纳入标准具体包括以下两个要求(需同时满足两个要求才将该数据纳入):

[0116] 1)年龄大于18周岁(含18周岁)的入住重症监护室患者;

[0117] 2)经柏林标准诊断为ARDS患者,诊断标准:a)急性发作;b)PEEP(或CPAP)≥5cmH2O时,OI(PaO2/FiO2)<300mmHg;c)胸部影像双侧浸润影;d)呼吸衰竭无法用心力衰竭来完全解释。

[0118] 排除标准具体为以下四个要求中的任一个(只要有1个要求不满足即将该数据排除):

[0119] 1)数据记录不完整,如接受无创通气(不含机械通气数据)的患者等;

[0120] 2)年龄小于18周岁;

[0121] 3)姑息疗法,没有接受积极治疗;

[0122] 4)ICU记录时间小于48小时。

[0123] 筛除了不符合要求的患者数据后,还剩下475位患者用于建立预测模型。为了保证用于建模的变量具有意义,本具体实施例在临床医生的确认下提取临床上与ARDS有关的临床数据,共101个变量信息,其中包括患者的年龄、性别、APACHE II分数等可以直接读取的数据,也包括机械通气时间、患者生理信息等变量作为建模的输入变量。

[0124] 这101个变量中的机械通气时间(length of mechanical ventilation)为患者确诊ARDS后的首次机械通气持续时间。无机械天数(days free of mechanical ventilation)为ICU记录中患者没有接受(任何形式的)机械通气的天数,如果患者在拔管后24小时内去世,无机械天数认定为0。生理信息(physiologic information)是ARDS发作之前记录的生理数据。通气环境(Ventilator settings)取决于患者接受机械通气最初24

小时的仪器设置。ARDS发生24小时后，在标准通气环境(FiO2≥0.5;PEEP≥5cm H2O)下计算PaO2/FiO2值，作为氧合指数OI。

[0125] 由于仪器和数据管理的原因，存在个别数据缺失的情况，本具体实施例采用多重插补(Multiple imputation)的方法补充缺失的数据。多重插补是一种对缺失数据进行调整的方法，采用一系列可能的数据集来填充每一个缺失数据值，然后使用完全数据的标准去分析多重插补数据集，最后对这些分析结果归纳综合。本具体实施例可使用SPSS软件完成多重插补操作。

[0126] 以上步骤即为本具体实施例的数据预处理流程，最终可得到每位患者的101种变量特征用于建立预测模型。

[0127] 2.预测模型建立与测试

[0128] 2.1样本分类

[0129] 根据患者生存天数，可以把数据分为三类，并使用三个预测模型(住院死亡率预测模型、30天死亡率预测模型以及一年死亡率预测模型)分别进行预测：

[0130] 模型1(住院死亡率预测模型)：住院内死亡vs.住院期间存活，其中前者为阳性，后者为阴性；

[0131] 模型2(30天死亡率预测模型)：住院后30天内死亡vs.30天后存活，前者为阳性，后者为阴性；

[0132] 模型3(一年死亡率预测模型)：住院后一年内死亡vs.一年后存活，前者为阳性，后者为阴性。

[0133] 这三个预测模型的特征筛选、分类器训练(即模型训练)与测试、结果评估的流程基本一致。

[0134] 2.2特征筛选

[0135] 本具体实施例使用SPSS软件对特征进行组间分析，找出不同类别样本间差异显著的特征作为输入，可以降低数据冗余，减少模型计算，找出更有意义的特征。本具体实施例分别对分类得到的三个预测模型中的阳性组与阴性组进行组间分析，具体操作是：对于符合正态分布的连续变量采用t检验(Student t-test)分析，而非正态分布的连续变量采用非参数检验(Mann-Whitney U test)；对于离散变量则通过卡方检验(Chi2test)或Fisher's exact test进行分析。经过上述检验的P值小于0.05的变量特征可认为组间差异显著，则保留该特征；对于P值大于0.05的认为该特征在阳性组与阴性组的组间差异不明显，对预测模型的分类影响较小，可以删去。

[0136] 2.3模型训练与测试

[0137] 本具体实施例使用随机森林(Random Forest,RF)分别对ARDS患者的住院死亡率、30天死亡率以及一年死亡率建立预测模型。RF是一种机器学习算法，能随机生成多棵决策树(Decision tree)，每棵决策树都是一个分类器，会通过一系列决策对输入的数据进行预测，分配标签，最后RF的输出结果则通过决策树“投票”产生。本具体实施例采用了scikit-learn Python library工具包实现RF算法，如图3所示。图3中包含若干棵决策树，对每个样本x，每棵树都会给出自己的预测结果，每棵树“投票”决定最终结果y。

[0138] 建立预测模型主要分为两步：训练(training)与测试(testing)。

[0139] 为了保证预测模型的可靠与稳定，本具体实施例使用了八折交叉验证法(8-folds

cross-validation)对模型的预测效果进行评价,总体流程如图4所示,具体包括:

[0140] 首先把数据分为类别比例接近的8份,每折使用其中7份数据作为训练集训练模型,剩下1份数据用于测试模型效果(即测试集)。每一折训练的测试集和训练集都不同,一共循环8次(如图4中的Loop 2),模型的最终结果会在八折交叉验证的基础上求出。该流程的特征筛选在机器学习交叉验证过程中的每一折训练集中是分别进行的,故可能会出现每一折所使用的特征不同的情况。

[0141] 在每一折使用训练集构预测建模型时,可使用网格寻优的方法来寻找最佳的模型参数(RF的参数包括决策树数目n_estimators、分枝标准criterion、最小叶子样本数min_sample_leaf等)。所以本具体实施例会把训练集再次划分为训练集与验证集两部分,循环测试每组参数的效果(如图4中的Loop 1),选择AUROC效果最佳的一组参数来构建各个分类器模型(即预测模型),然后对测试集进行测试,得到本折预测结果。

[0142] 再对8折的结果求平均值,得到各个预测模型的预测性能结果。

[0143] 最后从各个预测模型中选择预测性能结果最好的模型作为最终的预测模型。由于本具体实施例将数据分为三类且分别训练了3个不同的模型,所以每一类都会得到一个最终的预测结果。

[0144] 3.结果评估与比较

[0145] 3.1分类结果(即模型预测结果)评估

[0146] 本具体实施例分类结果(即模型预测结果)的评估标准是AUROC。ROC曲线的横轴是假阳率(False Positive Rate,FPR),纵轴是真阳率(True Positive Rate,TPR),曲线上的点是根据样本的概率输出在不同的分类阈值下所表现出来的TPR与FPR所决定(当输出概率大于等于设定阈值时,该样本预测为阳性,否则为阴性)。随着分类阈值逐渐减小,越来越多的样本被预测为阳性,但是这些阳性中同样掺杂着真正的阴性样本,即TPR与FPR会同时增大。当阈值最大时,对应ROC曲线坐标点为(0,0),阈值最小时对应坐标点(1,1),而理想目标是TPR＝1,FPR＝0,对应的ROC曲线坐标点是(0,1),所以本具体实施例在模型训练中选择最接近坐标点(0,1)的ROC曲线上的点所代表的概率值作为分类阈值。

[0147] AUROC是ROC曲线下面积,可用于评价分类器性能。随机挑选一个阳性样本与阴性样本,输入到预测模型之中,输出两个样本的预测概率,由大到小排列将阳性样本排在阴性样本前面的概率就是AUC值(也就是阳性样本输出概率大于阴性样本输出概率的可能性)。

[0148] 同时,根据最佳分类阈值,可以求得分类的准确率、敏感度、特异度,计算公式如下:

[0149] 准确率:Accuracy＝(TP+TN)/(TP+TN+FP+FN)

[0150] 敏感度:Sensitivity＝TP/(TP+FN)

[0151] 特异度:Specificity＝TN/(TN+FP)

[0152] 其中,TP:True Positive,真阳性,即实际为阳性,预测为阳性的样本。

[0153] FP:False Positive,假阳性,即实际为阴性,预测为阳性的样本。

[0154] TN:True Negative,真阴性,即实际为阴性,预测为阴性的样本。

[0155] FN:False Negative,假阴性,即实际为阳性,预测为阴性的样本。

[0156] 使用基于RF的机器学习方法训练死亡率预测模型,经测试可得结果如下:

[0157] 1)住院死亡率预测模型的AUROC值为0.854(95％置信区间为0.835-0.874,p＜

0.001)；

[0158] 2)30天死亡率预测模型的AUROC值为0.817(95％置信区间为0.796-0.839,p＜0.001)；

[0159] 3)一年死亡率预测模型的AUROC值为0.817(95％置信区间为0.800-0.834,p＜0.001)。

[0160] 3.2方法比较

[0161] 为了比较本发明与现有方法的预测效果,本具体实施例还分别复现了已有研究中提到的关于ARDS死亡率预测的模型,对同一批的数据进行预测。已有的预测模型包括SAPS II、OI、OSI以及APPS,结果表现为AUROC,并分别与RF的预测结果比较。

[0162] 预测时使用SPSS软件进行统计处理。符合正态分布的计量数据以均值±标准差(mean±std)表示;不符合正态分布的计量数据则以中位数(四分位数)表示;计数数据以分数或百分比形式表现。以年龄,性别与APACHE II分数作为控制变量,采用多变量逻辑回归方法分析各因素与死亡率的关系,并绘制ROC曲线。

[0163] 上述各种预测方法的结果如下表1所示：

[0164] 表1

[0165]

预测方法（模型/评分）	预测结果（AUROC）		
	住院死亡	30 天死亡	一年死亡
SAPS II	0.686	0.670	0.696
APPS	0.657	0.649	0.679
OSI	0.647	0.705	0.600
OI	0.610	0.661	0.550
RF	0.843	0.861	0.747

[0166] 显然,从表1可以看出,本发明使用基于RF的机器学习方法预测ARDS患者死亡率,效果普遍比现有方法好。

[0167] 4.预测模型的应用

[0168] 由上述结果可知,使用机器学习方法训练的预测模型,能有效地对ARDS在住院死亡、30天死亡以及一年死亡进行预测,且相对于现有的其他预测方法,本发明的方法的预测准确率有显著的提高。该结果说明了本具体实施例的模型设计方案与流程是可行的,所以本具体实施例可利用现有的全部ARDS数据,按照上述方案与流程,训练一个最终模型,对新来的数据进行预测,以实现在临床上的应用,具体流程如图5所示,通过对现有数据的分组、特征筛选、参数寻优,可以得到三个预测模型,分别是模型1(用于预测ARDS患者住院死亡的概率)、模型2(预测30天内死亡的概率)和模型3(预测一年内死亡的概率)。

[0169] 当需要对新数据进行预测时,首先对新数据进行特征筛选(根据训练时的三种筛选策略),选出一样的特征,再分别输入到对应的模型中,进而实现对新数据的预测。

[0170] 与图1的方法相对应,本发明实施例还提供了一种急性呼吸窘迫综合征死亡率预测系统,包括以下模块：

[0171] 获取模块,用于获取急性呼吸窘迫综合征患者的医疗数据；

[0172] 训练模块,用于根据获取的医疗数据采用机器学习的方法进行模型训练,得到急性呼吸窘迫综合征死亡率预测模型；

[0173] 预测模块，用于采用急性呼吸窘迫综合征死亡率预测模型对待预测的对象进行预测，得到急性呼吸窘迫综合征死亡率的预测结果。

[0174] 上述方法实施例中的内容均适用于本系统实施例中，本系统实施例所具体实现的功能与上述方法实施例相同，并且达到的有益效果与上述方法实施例所达到的有益效果也相同。

[0175] 与图1的方法相对应，本发明实施例还提供了一种急性呼吸窘迫综合征死亡率预测系统，包括：

[0176] 至少一个处理器；

[0177] 至少一个存储器，用于存储至少一个程序；

[0178] 当所述至少一个程序被所述至少一个处理器执行，使得所述至少一个处理器实现如本发明所述的急性呼吸窘迫综合征死亡率预测方法。

[0179] 上述方法实施例中的内容均适用于本系统实施例中，本系统实施例所具体实现的功能与上述方法实施例相同，并且达到的有益效果与上述方法实施例所达到的有益效果也相同。

[0180] 以上是对本发明的较佳实施进行了具体说明，但本发明并不限于所述实施例，熟悉本领域的技术人员在不违背本发明精神的前提下还可做作出种种的等同变形或替换，这些等同的变形或替换均包含在本申请权利要求所限定的范围内。

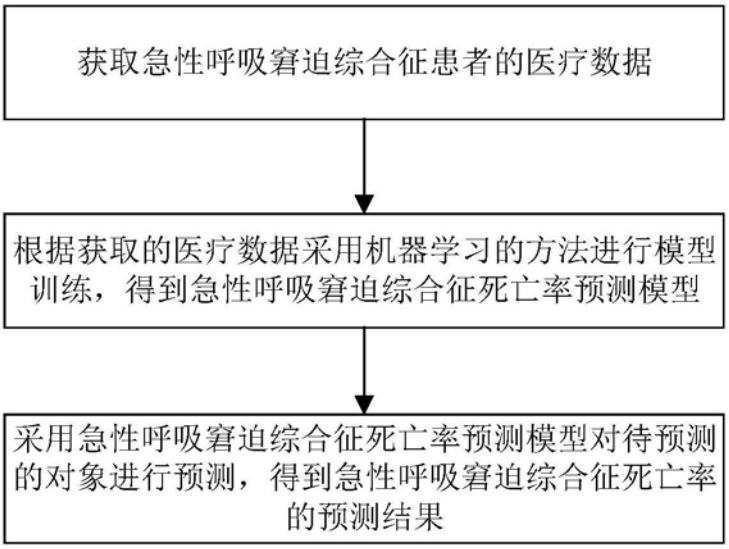

获取急性呼吸窘迫综合征患者的医疗数据
根据获取的医疗数据采用机器学习的方法进行模型训练，得到急性呼吸窘迫综合征死亡率预测模型
采用急性呼吸窘迫综合征死亡率预测模型对待预测的对象进行预测，得到急性呼吸窘迫综合征死亡率的预测结果

图1

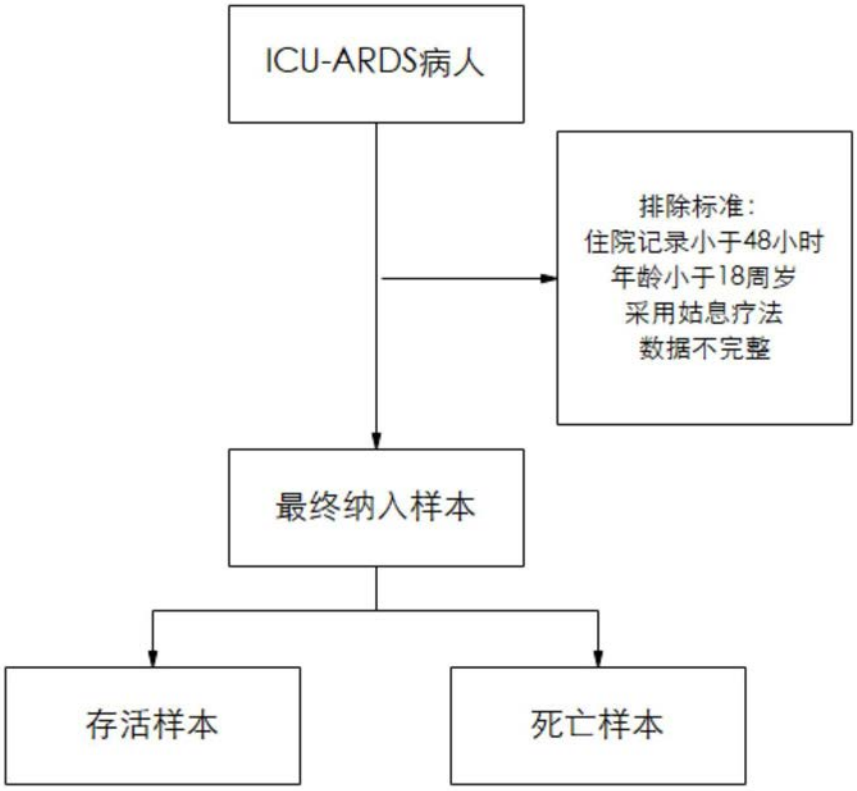

ICU-ARDS病人
排除标准：
住院记录小于48小时
年龄小于18周岁
采用姑息疗法
数据不完整
最终纳入样本
存活样本
死亡样本

图2

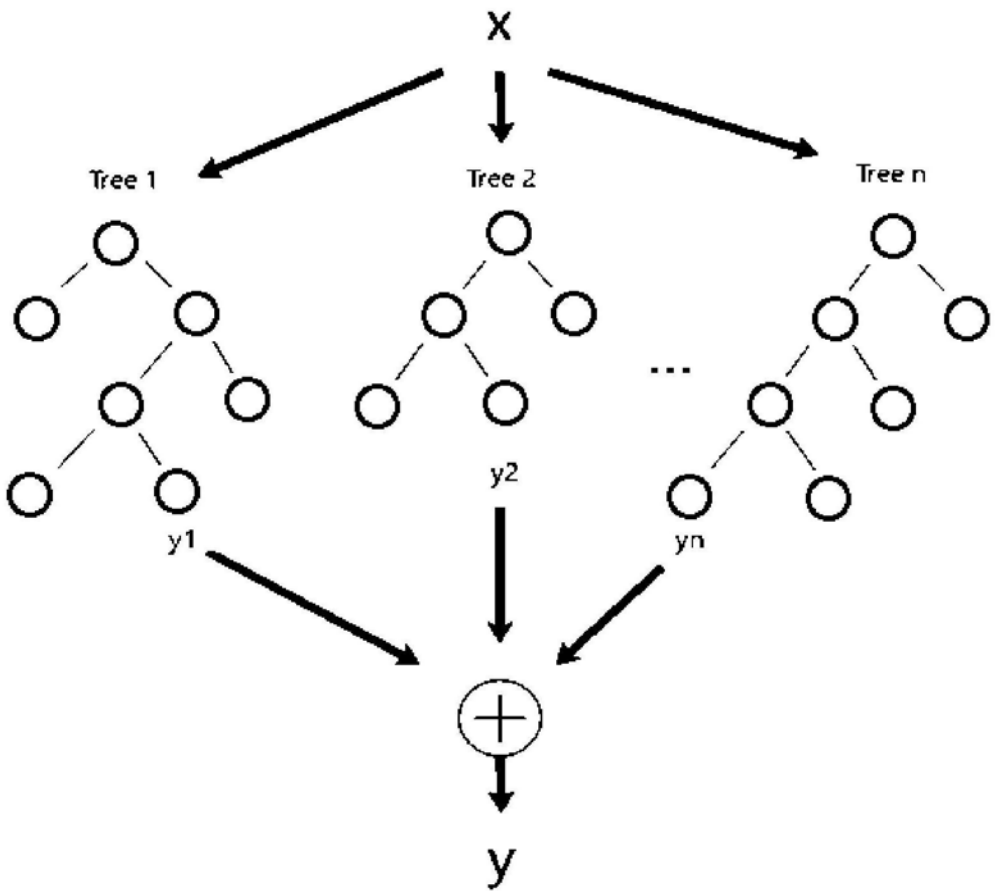

X
Tree 1
Tree 2
Tree n
y1
y2
yn
y

图3

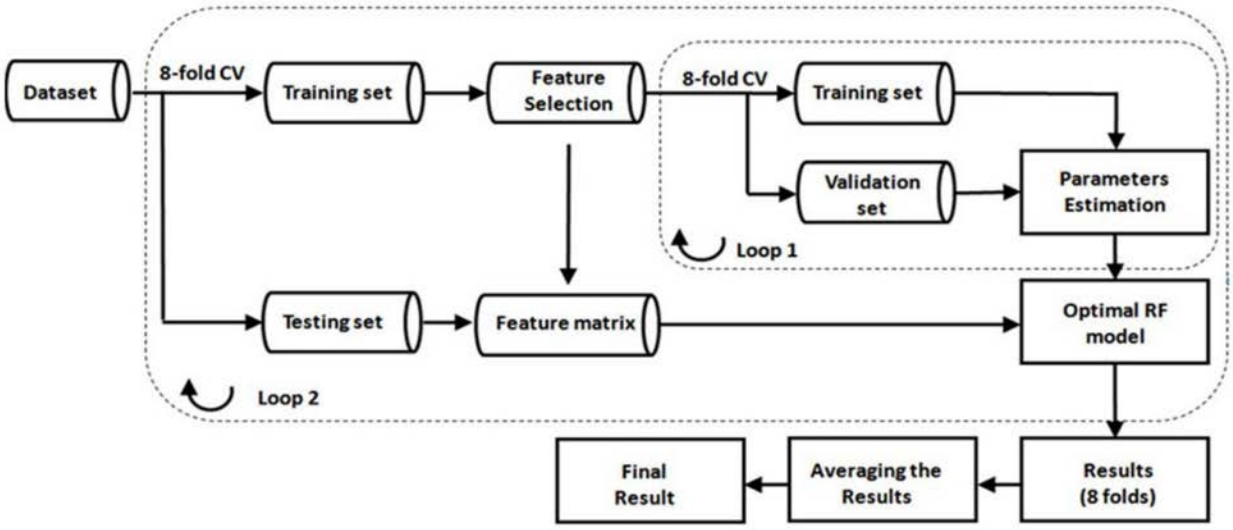

Dataset
8-fold CV
Training set
Feature Selection
8-fold CV
Training set
Validation set
Parameters Estimation
Loop 1
Testing set
Feature matrix
Optimal RF model
Loop 2
Final Result
Averaging the Results
Results (8 folds)

图4

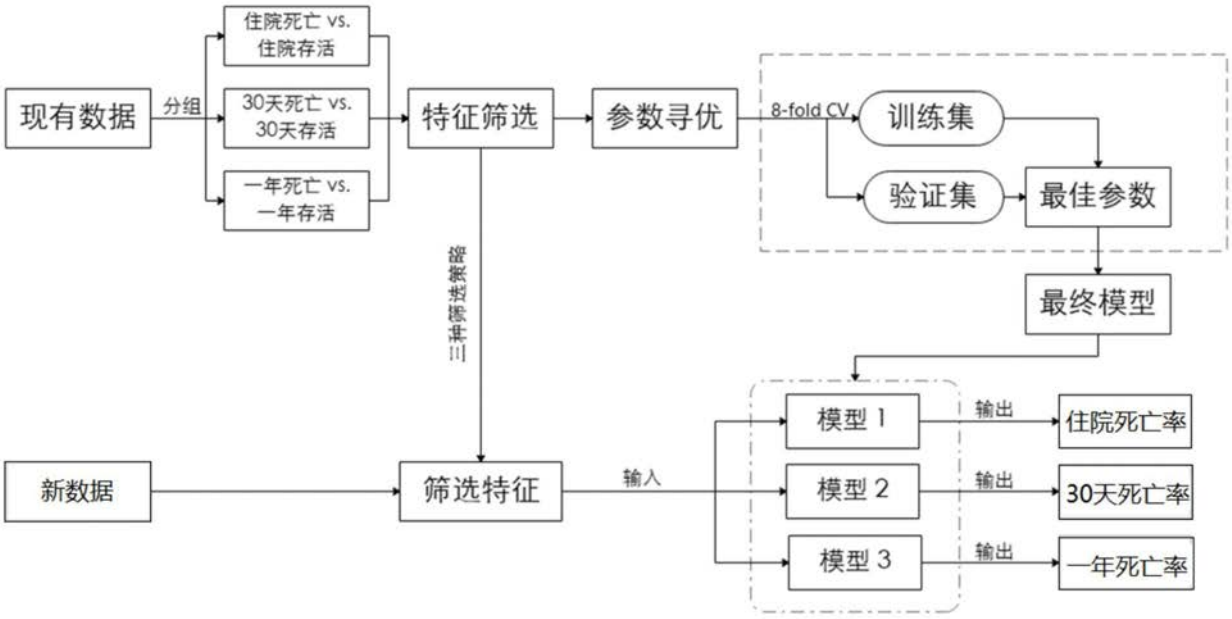

现有数据
分组
住院死亡 vs.
住院存活
30天死亡 vs.
30天存活
一年死亡 vs.
一年存活
特征筛选
三种筛选策略
参数寻优
8-fold CV
训练集
验证集
最佳参数
最终模型
新数据
筛选特征
输入
模型 1
模型 2
模型 3
输出
输出
输出
住院死亡率
30天死亡率
一年死亡率

图5

专利名称(译)	一种急性呼吸窘迫综合征死亡率预测方法及系统		
公开(公告)号	CN110051324A	公开(公告)日	2019-07-26
申请号	CN201910194628.2	申请日	2019-03-14
[标]申请(专利权)人(译)	深圳大学		
申请(专利权)人(译)	深圳大学		
当前申请(专利权)人(译)	深圳大学		
[标]发明人	黄炳升 梁栋 刘勇 黄树华 余夏夏		
发明人	黄炳升 梁栋 刘勇 邹儒诗 黄树华 余夏夏		
IPC分类号	A61B5/00 A61B5/08		
CPC分类号	A61B5/0826 A61B5/7267 A61B5/7275		
代理人(译)	胡辉		
外部链接	Espacenet SIPO		

摘要(译)

本发明公开了一种急性呼吸窘迫综合征死亡率预测方法及系统，方法包括：获取急性呼吸窘迫综合征患者的医疗数据；根据获取的医疗数据采用机器学习的方法进行模型训练，得到急性呼吸窘迫综合征死亡率预测模型；采用急性呼吸窘迫综合征死亡率预测模型对待预测的对象进行预测，得到急性呼吸窘迫综合征死亡率的预测结果。本发明通过机器学习的方法训练出急性呼吸窘迫综合征死亡率预测模型，再采用急性呼吸窘迫综合征死亡率预测模型来预测ARDS患者的死亡率，将机器学习应用于ARDS患者死亡率预测上，能通过机器学习训练的模型准确和客观地预测出ARDS患者的死亡率，为临床医师提供了更有效和可行的预测信息，可广泛应用于医学数据挖掘领域。