(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 108364106 A
(43)申请公布日 2018.08.03

(21)申请号 201810161565.6

(22)申请日 2018.02.27

(71)申请人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区八卦岭工业区平安大厦六楼

(72)发明人 袁军 陆源 魏尧东

(74)专利代理机构 深圳众鼎专利商标代理事务所(普通合伙) 44325
代理人 阳开亮

(51)Int.Cl.
G06Q 10/04(2012.01)
G06Q 10/06(2012.01)
G06Q 40/00(2012.01)

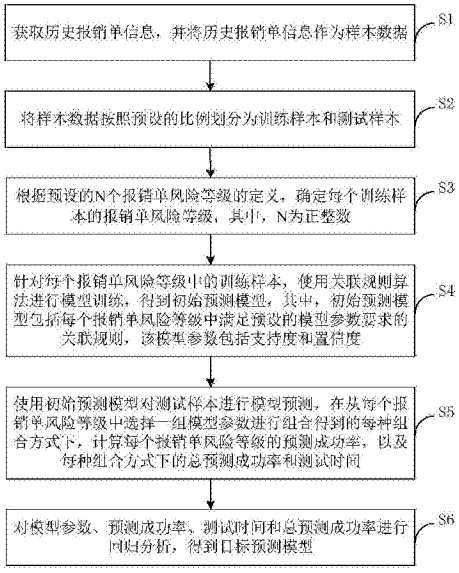
权利要求书3页 说明书16页 附图6页

(54)发明名称

一种报销单风险预测方法、装置、终端设备及存储介质

(57)摘要

本发明公开了一种报销单风险预测方法、装置、设备及介质。该方法包括:获取历史报销单信息作为样本数据,并按照预设的比例划分为训练样本和测试样本;根据预设的报销单风险等级,确定每个训练样本的风险等级;针对每个风险等级中的训练样本,使用关联规则算法进行模型训练,得到初始预测模型;使用初始预测模型对测试样本进行预测,在从每个风险等级中选择一组模型参数进行组合得到的每种组合方式下,计算每个风险等级的预测成功率,以及每种组合方式下的总预测成功率和测试时间;对模型参数、预测成功率、测试时间和总预测成功率作回归分析,得到目标预测模型,从而辅助工作人员高效地识别报销单的风险级别,提高预测报销单风险等级的准确率。



CN 108364106 A

1.一种报销单风险预测方法，其特征在于，所述报销单风险预测方法包括：

获取历史报销单信息，并将所述历史报销单信息作为样本数据；

将所述样本数据按照预设的比例划分为训练样本和测试样本；

根据预设的N个报销单风险等级的定义，确定每个所述训练样本的报销单风险等级，其中，N为正整数；

针对每个所述报销单风险等级中的所述训练样本，使用关联规则算法进行模型训练，得到初始预测模型，其中，所述初始预测模型包括每个所述报销单风险等级中满足预设的模型参数要求的关联规则，所述模型参数包括支持度和置信度；

使用所述初始预测模型对所述测试样本进行模型预测，在从每个所述报销单风险等级中选择一组所述模型参数进行组合得到的每种组合方式下，计算每个所述报销单风险等级的预测成功率，以及每种所述组合方式下的总预测成功率和测试时间；

对所述模型参数、所述预测成功率、所述测试时间和所述总预测成功率进行回归分析，得到目标预测模型。

2.如权利要求1所述的报销单风险预测方法，其特征在于，所述针对每个所述报销单风险等级中的所述训练样本，使用关联规则算法进行模型训练，得到初始预测模型包括：

对每个所述报销单风险等级中的所述训练样本进行数据预处理，得到每个所述报销单风险等级中的待处理数据集；

对所述待处理数据集使用关联规则算法进行数据挖掘，得到每个所述报销单风险等级中的多个项集；

针对每个所述报销单风险等级，从该报销单风险等级中的所述项集中筛选出满足所述模型参数要求的目标项集，并根据该目标项集建立关联规则；

根据所述关联规则和所述关联规则对应的所述模型参数要求，构建所述初始预测模型。

3.如权利要求1或2所述的报销单风险预测方法，其特征在于，所述使用所述初始预测模型对所述测试样本进行模型预测，在从每个所述报销单风险等级中选择一组所述模型参数进行组合得到的每种组合方式下，计算每个所述报销单风险等级的预测成功率，以及该组合方式下的总预测成功率和测试时间包括：

根据所述预设的N个报销单风险等级的定义，确定每个所述测试样本的报销单风险等级，以及每个所述报销单风险等级的测试样本数；

按照如下公式计算所述测试样本中每个报销单风险等级的概率：

$$P_i=\frac{R_i}{S}$$

其中，$i\in[1,N]$，P_i为所述测试样本中第i个报销单风险等级的概率，R_i为第i个所述报销单风险等级的测试样本数，S为所述测试样本的总数；

从每个所述报销单风险等级中选择一组所述模型参数进行组合，得到L种组合方式，其中，L为正整数；

针对每种所述组合方式，按照所述概率由高到低的顺序，使用所述初始预测模型对所述测试样本进行报销单风险等级预测，得到每个所述测试样本的预测结果，并获取在该组合方式下的进行报销单风险等级预测的测试时间；

将每个所述测试样本的所述预测结果与该测试样本的报销单风险等级进行对比，若两者相同则确认该测试样本预测成功，并统计在每种所述组合方式下每个所述报销单风险等级下的测试样本预测成功的个数；

按照如下公式计算每种所述组合方式下每个所述报销单风险等级的预测成功率：

$$hitrate_i = \frac{M_i}{R_i}$$

其中，$hitrate_i$为第i个所述报销单风险等级的预测成功率，M_i为第i个所述报销单风险等级下的测试样本预测成功的个数；

按照如下公式计算每种所述组合方式下的总预测成功率：

$$hitRate = \frac{\sum_{i=1}^{N} M_i}{S}$$

其中，hitRate为所述总预测成功率。

4.如权利要求3所述的报销单风险预测方法，其特征在于，所述对所述模型参数、所述预测成功率、所述测试时间和所述总预测成功率进行回归分析，得到目标预测模型包括：

将每个所述报销单风险等级中的所述模型参数，以及所述预测成功率和所述测试时间作为设计变量，将所述总预测成功率作为目标变量，使用所述设计变量和所述目标变量进行函数拟合，得到拟合函数；

对所述拟合函数进行求解，根据求解结果将所述总预测成功率最高并且所述模型参数的值最高的一组设计变量作为模型配置参数，并根据所述模型配置参数构建目标预测模型，其中，所述目标预测模型的模型精确度为最高的所述总预测成功率。

5.如权利要求4所述的报销单风险预测方法，其特征在于，所述对所述模型参数、所述预测成功率、所述测试时间和所述总预测成功率进行回归分析，得到目标预测模型之后，所述报销单风险预测方法还包括：

将所述样本数据分割成K个子样本数据；

从所述K个子样本数据中，选择一个所述子样本数据作为所述测试样本，剩余K-1个所述子样本数据作为所述训练样本，进行所述模型训练、所述模型预测和所述回归分析，得到K个所述目标预测模型和每个所述目标预测模型的所述模型精确度，其中，K为正整数；

将所述模型精确度最高的目标预测模型作为合理模型。

6.一种报销单风险预测装置，其特征在于，所述报销单风险预测装置包括：

样本数据采集模块，用于获取历史报销单信息，并将所述历史报销单信息作为样本数据；

第一划分模块，用于将所述样本数据按照预设的比例划分为训练样本和测试样本；

风险等级预设模块，用于根据预设的N个报销单风险等级的定义，确定每个所述训练样本的报销单风险等级，其中，N为正整数；

初始预测模型获取模块，用于针对每个所述报销单风险等级中的所述训练样本，使用关联规则算法进行模型训练，得到初始预测模型，其中，所述初始预测模型包括每个所述报销单风险等级中满足预设的模型参数要求的关联规则，所述模型参数包括支持度和置信度；

初始预测模型测试模块，用于使用所述初始预测模型对所述测试样本进行模型预测，在从每个所述报销单风险等级中选择一组所述模型参数进行组合得到的每种组合方式下，计算每个所述报销单风险等级的预测成功率，以及每种所述组合方式下的总预测成功率和测试时间；

目标预测模型获取模块，用于对所述模型参数、所述预测成功率、所述测试时间和所述总预测成功率进行回归分析，得到目标预测模型。

7.如权利要求6所述的报销单风险预测装置装置，其特征在于，所述初始预测模型获取模块包括：

数据预处理单元，用于对每个所述报销单风险等级中的所述训练样本进行数据预处理，得到每个所述报销单风险等级中的待处理数据集；

训练样本挖掘单元，用于对所述待处理数据集使用关联规则算法进行数据挖掘，得到每个所述报销单风险等级中的多个项集；

关联规则获取单元，用于针对每个所述报销单风险等级，从该报销单风险等级中的所述项集中筛选出满足所述模型参数要求的目标项集，并根据该目标项集建立关联规则；

初始预测模型构建单元，用于根据所述关联规则和所述关联规则对应的所述模型参数要求，构建所述初始预测模型。

8.如权利要求6所述的报销单风险预测装置装置，其特征在于，所述报销单风险预测装置装置还包括：

第二划分模块，用于将所述样本数据分割成K个子样本数据；

交叉验证模块，用于从所述K个子样本数据中，选择一个所述子样本数据作为所述测试样本，剩余K-1个所述子样本数据作为所述训练样本，进行所述模型训练、所述模型预测和所述回归分析，得到K个所述目标预测模型和每个所述目标预测模型的所述模型精确度，其中，K为正整数；

合理模型获取模块，用于将所述模型精确度最高的目标预测模型作为合理模型。

9.一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1至5任一项所述报销单风险预测方法的步骤。

10.一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至5任一项所述报销单风险预测方法的步骤。

一种报销单风险预测方法、装置、终端设备及存储介质

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种报销单风险预测方法、装置、终端设备及存储介质。

背景技术

[0002] 在日常的费用报销中会存在着一些恶意报销，虚假报销的情况，为了加强风险管理，目前大多使用基于关联规则的挖掘算法建立报销单风险等级预测模型，来进行预测报销单的风险等级。但是当报销单风险等级数据分布不均时，低概率风险等级的报销单在训练数据中所占比例很小，传统的基于关联规则的挖掘算法会把低概率风险等级的报销单数据当做噪声处理而丢弃，导致所建模型无法训练学习得到低概率风险等级报销单数据的特征，使得所建模型用于预测新的报销单的风险等级时，其预测准确率较低。

发明内容

[0003] 本发明实施例提供一种报销单风险预测方法，以解决目前报销单风险等级预测模型对报销单的风险等级预测准确率低的问题。

[0004] 第一方面，本发明实施例提供一种报销单风险预测方法，包括：获取历史报销单信息，并将所述历史报销单信息作为样本数据；

[0005] 将所述样本数据按照预设的比例划分为训练样本和测试样本；

[0006] 根据预设的N个报销单风险等级的定义，确定每个所述训练样本的报销单风险等级，其中，N为正整数；

[0007] 针对每个所述报销单风险等级中的所述训练样本，使用关联规则算法进行模型训练，得到初始预测模型，其中，所述初始预测模型包括每个所述报销单风险等级中满足预设的模型参数要求的关联规则，所述模型参数包括支持度和置信度；

[0008] 使用所述初始预测模型对所述测试样本进行模型预测，在从每个所述报销单风险等级中选择一组所述模型参数进行组合得到的每种组合方式下，计算每个所述报销单风险等级的预测成功率，以及每种所述组合方式下的总预测成功率和测试时间；

[0009] 对所述模型参数、所述预测成功率、所述测试时间和所述总预测成功率进行回归分析，得到目标预测模型。

[0010] 第二方面，本发明实施例提供一种报销单风险预测装置，包括：

[0011] 样本数据采集模块，用于获取历史报销单信息，并将所述历史报销单信息作为样本数据；

[0012] 第一划分模块，用于将所述样本数据按照预设的比例划分为训练样本和测试样本；

[0013] 风险等级预设模块，用于根据预设的N个报销单风险等级的定义，确定每个所述训练样本的报销单风险等级，其中，N为正整数；

[0014] 初始预测模型获取模块，用于针对每个所述报销单风险等级中的所述训练样本，

使用关联规则算法进行模型训练，得到初始预测模型，其中，所述初始预测模型包括每个所述报销单风险等级中满足预设的模型参数要求的关联规则，所述模型参数包括支持度和置信度；

[0015] 初始预测模型测试模块，用于使用所述初始预测模型对所述测试样本进行模型预测，在从每个所述报销单风险等级中选择一组所述模型参数进行组合得到的每种组合方式下，计算每个所述报销单风险等级的预测成功率，以及每种所述组合方式下的总预测成功率和测试时间；

[0016] 目标预测模型获取模块，用于对所述模型参数、所述预测成功率、所述测试时间和所述总预测成功率进行回归分析，得到目标预测模型。

[0017] 第三方面，本发明实施例提供一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现所述报销单风险预测方法的步骤。

[0018] 第四方面，本发明实施例提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时实现所述报销单风险预测方法的步骤。

[0019] 本发明实施例提供的一种报销单风险预测方法、装置、终端设备及存储介质中，通过获取历史报销单信息作为样本数据，并将样本数据按照预设的比例划分为训练样本和测试样本，能够通过测试样本来评价训练样本训练得到的模型的质量；在对报销单风险等级进行定义，确定每个训练样本的报销单风险等级后，针对每个报销单风险等级中的训练样本，使用关联规则算法进行模型训练，获取各报销单风险等级中满足预设的模型参数要求的目标关联规则，构建初始预测模型，这种按照不同报销单风险等级进行模型训练的方式能够学习到样本数据中所占比例较小的报销单数据的特征，避免这部分报销单数据被当做噪声处理而丢弃的情况，从而提高模型的精确度；最后再使用初始预测模型对测试样本进行模型预测，在从每个报销单风险等级中选择一组模型参数进行组合得到的每种组合方式下，计算每种组合方式下的每个报销单风险等级的预测成功率、总预测成功率和测试时间，并对这些离散型数据作回归分析，得到目标预测模型，通过模型预测和回归分析得到精准的模型配置参数，使得目标预测模型能够辅助工作人员精准高效地识别报销单的风险级别，有效提高预测报销单风险等级的准确率。

附图说明

[0020] 为了更清楚地说明本发明实施例的技术方案，下面将对本发明实施例的描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0021] 图1是本发明实施例1中提供的报销单风险预测方法的流程图；

[0022] 图2是本发明实施例1中提供的报销单风险预测方法中步骤S4的实现流程图；

[0023] 图3是本发明实施例1中提供的报销单风险预测方法中步骤S5的实现流程图；

[0024] 图4是本发明实施例1中提供的报销单风险预测方法中步骤S6的实现流程图；

[0025] 图5是本发明实施例1中提供的报销单风险预测方法中使用交叉验证方法测试目

标预测模型精确度的实现流程图；

[0026] 图6是本发明实施例2提供的报销单风险预测装置的示意图；

[0027] 图7是本发明实施例4提供的终端设备的示意图。

具体实施方式

[0028] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0029] 实施例1

[0030] 请参阅图1，图1示出了本发明实施例提供的报销单风险预测方法的实现流程。该报销单风险预测方法应用在各个企事业单位的报销单审核系统中，用于识别报销单的风险级别，提高预测报销单风险等级的准确率。如图1所示，该报销单风险预测方法包括步骤S1至步骤S6，详述如下：

[0031] S1：获取历史报销单信息，并将历史报销单信息作为样本数据。

[0032] 在本发明实施例中，样本数据是从报销单数据库的历史报销单中采集，获取历史报销单信息。

[0033] 历史报销单是企事业单位在生产经营过程中存储在报销单数据库中的数据。每个历史报销单信息包括从报销单上获得的信息和在处理报销单过程中所产生的信息，具体地，历史报销单信息包括但不限于报销单编号、报销单名称、经办人中文姓名、报销人中文姓名、部门名称、报销金额、合计金额、附单据张数等多种属性信息，以历史报销单信息作为样本数据进行挖掘学习。

[0034] 具体地，在对报销单的样本数据进行采集、存储和处理加工时，使用Hadoop大数据平台实现从报销单数据库中存储的历史报销单中采集样本数据。

[0035] Hadoop是一种分布式系统基础架构，实现了一个分布式文件系统(Hadoop Distributed File System,HDFS)，HDFS能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。在对样本数据的采集过程中，通过采用分布式文件系统HDFS和数据仓库工具hive进行数据处理，其中，hive是基于Hadoop的一个数据仓库工具，用于存储、查询和分析存储在Hadoop中的大规模数据，使得采用Hadoop大数据平台进行样本数据的采集具有采集效率高的优点。

[0036] S2：将样本数据按照预设的比例划分为训练样本和测试样本。

[0037] 在本发明实施例中，预先设置用于对样本数据进行划分的比例。

[0038] 需要说明的是，该预设的比例可以是根据历史经验获取的比例，也可以是根据对样本数据进行分析得到的比例，其具体可以根据实际应用的需要进行设置，此处不做限制。

[0039] 训练样本是用于机器学习的样本数据集，进行数据特征学习，即采用训练样本中的数据信息进行训练机器学习模型，以确定机器学习模型的参数，测试样本是用于测试完成训练的机器学习模型的分辨能力，如报销单风险等级的预测成功率。

[0040] 具体地，按照预设的比例将样本数据划分为训练样本和测试样本。例如，按照9:1的比例对样本数据进行划分，即将90％的样本数据作为训练样本，剩余10％的数据作为测

试样本。若采集到的总样本数据为605万份，则按照9:1的比例，将其中544.5万份样本数据作为训练样本进行特征学习，剩余60.5万份样本数据作为测试样本进行预测其报销单风险等级，验证模型的预测成功率。

[0041] S3：根据预设的N个报销单风险等级的定义，确定每个训练样本的报销单风险等级，其中，N为正整数。

[0042] 在本发明实施例中，预先设置N个报销单风险等级的定义，用于区分报销单的风险，其中，N为正整数，报销单风险等级的定义可以根据实际应用的需要进行设置，此处不做限制。报销单的风险等级越大，报销单存在的风险越高。

[0043] 具体地，根据预设的报销单风险等级的定义确定每个训练样本的报销单风险等级，并为每个训练样本标识对应的报销单风险等级的标识信息。

[0044] 为了更好的理解本步骤，下面以一个具体的报销单风险等级分类为例加以说明。如表一所示，表一示出了报销单的风险等级分为0、1、2、3四个风险等级的分类标准。

[0045] 表一

[0046]

等级	标准
3	未通过报销，即退回后删除的报销单
2	退回报销单，退回环节为部门预算员或中心会计，退回意见中含“假发票”、“连号”、“重号”等严重违反制度的报销单
1	退回报销单，但不属于退回环节为部门预算员或中心会计，退回意见中不含“假发票”、“连号”、“重号”等严重违反制度的报销单
0	一次通过无退回的报销单

[0047] S4：针对每个报销单风险等级中的训练样本，使用关联规则算法进行模型训练，得到初始预测模型，其中，初始预测模型包括每个报销单风险等级中满足预设的模型参数要求的关联规则，该模型参数包括支持度和置信度。

[0048] 具体地，根据步骤S3中标识的每个训练样本的报销单风险等级标识信息，将采集整理得到的训练样本按照预设的报销单风险等级分类的标准进行分组，分别使用关联规则算法进行机器学习。对每一组训练样本预设模型参数要求，该模型参数要求包括但不限于预设的支持度阈值和置信度阈值，根据该模型参数要求，筛选出满足该支持度阈值和该置信度阈值的模型参数及其对应的关联规则，并根据该模型参数和该模型参数对应的关联规则，构建得到初始预测模型。

[0049] 需要说明的是，在预设的模型参数要求中可以预设一组支持度阈值和置信度阈值，也可以预设多组支持度阈值和置信度阈值，预设的支持度阈值和置信度阈值可以根据历史经验取值，也可以根据数据的分布情况进行取值，此处不做限制。

[0050] 例如，当报销单风险等级预设为0、1、2、3四个等级时，具体分组如下：

[0051] P_0:$sup_0=x_0$,$confid_0=y_0$

[0052] P_1:$sup_1=x_1$,$confid_1=y_1$

[0053] P_2:$sup_2=x_2$,$confdi_2=y_2$

[0054] P_3:$sup_3=x_3$,$confid_3=y_3$

[0055] 其中，P_0、P_1、P_2、P_3分别为训练样本按0、1、2、3四个报销单风险等级分类的分组，sup_i为支持度阈值，$confid_i$为置信度阈值，$x_i\in[0,1]$，$y_i\in[0,1]$，且$y_i\geq x_i$，i=0,1,2,3。例如，x_i和yi的具体取值可以是$x_0=0.6$，$y_0=0.8$；$x_1=0.1$，$y_1=0.7$；$x_2=0.6$，$y_2=0.95$；$x_3=0.1$，$y_3=0.7$或者$x_0=0.8$，$y_0=0.95$；$x_1=0.2$，$y_1=0.7$；$x_2=0.8$，$y_2=0.9$；$x_3=0.4$，$y_3=0.7$等。

[0056] S5：使用初始预测模型对测试样本进行模型预测，在从每个报销单风险等级中选择一组模型参数进行组合得到的每种组合方式下，计算每个报销单风险等级的预测成功率，以及每种组合方式下的总预测成功率和测试时间。

[0057] 在本发明实施例中，对每个报销单风险等级下的训练样本进行数据挖掘，在每个报销单风险等级的中预设了一组或多组模型参数要求进行筛选满足预设的模型参数要求的关联规则，在从每个报销单风险等级中选择一组模型参数进行组合得到的每种组合方式下，使用初始预测模型对测试样本进行模型预测，计算每种组合方式下的每个报销单风险等级的预测成功率和总预测成功率，并获取在该组合方式下完成全部测试样本的报销单风险等级预测的测试时间t。

[0058] S6：对模型参数、预测成功率、测试时间和总预测成功率进行回归分析，得到目标预测模型。

[0059] 在本发明实施例中，对步骤S5得到每种组合方式下的预测成功率、测试时间和总预测成功率等离散型数据，进行回归分析，确定变量间相互依赖的定量关系，得到一个连续的函数或者更加密集的离散方程，使该函数或该离散方程与离散型数据相吻合，并对该函数或该离散方程进行求解和分析，以总预测成功率最高并且模型参数的值最高的一组离散型数据作为模型最优配置参数，其中，支持度阈值和置信度阈值越大，得到的关联规则越准确，并根据模型最优配置参数以及对应满足该模型最优配置参数要求的关联规则构建目标预测模型，得到目标预测模型，用于预测报销单风险等级，提高报销单风险预测模型的准确率。

[0060] 在图1对应的实施例中，通过获取历史报销单信息作为样本数据，并将样本数据按照预设的比例划分为训练样本和测试样本，能够通过测试样本来评价训练样本训练得到的模型的质量；在对报销单风险等级进行定义，确定每个训练样本的报销单风险等级后，针对每个报销单风险等级中的训练样本，使用关联规则算法进行模型训练，获取各报销单风险等级中满足预设的模型参数要求的目标关联规则，构建初始预测模型，这种按照不同报销单风险等级进行模型训练的方式能够学习到样本数据中所占比例较小的报销单数据的特征，避免这部分报销单数据被当做噪声处理而丢弃的情况，从而提高模型的精确度；最后再使用初始预测模型对测试样本进行模型预测，在从每个报销单风险等级中选择一组模型参数进行组合得到的每种组合方式下，计算每种组合方式下的每个报销单风险等级的预测成功率、总预测成功率和测试时间，并对这些离散型数据作回归分析，得到目标预测模型，通

过模型预测和回归分析得到精准的模型配置参数，使得目标预测模型能够辅助工作人员精准高效地识别报销单的风险级别，有效提高预测报销单风险等级的准确率。

[0061] 接下来，在图1对应的实施例的基础之上，下面通过一个具体的实施例对步骤S4中提及的针对每个报销单风险等级中的训练样本，使用关联规则算法进行模型训练，得到初始预测模型的具体实现方法进行详细说明。

[0062] 请参阅图2，图2示出了本发明实施例提供的步骤S4的具体实现流程，详述如下：

[0063] S41：对每个报销单风险等级中的训练样本进行数据预处理，得到每个报销单风险等级中的待处理数据集。

[0064] 在本发明实施例中，数据预处理的过程包括对训练样本进行数据清理、数据集成和数据转换。

[0065] 数据清理是选取训练样本中需要的属性信息作为特征值进行训练学习。数据集成是将每个报销单风险等级的训练样本的数据集成到一个数据文件中作为数据集。数据转换是将数据集中训练样本的数据类型转换为统一的格式，例如，关联规则算法一般适用于对布尔型数据进行挖掘，则将数据类型全部转换为布尔型数据。

[0066] 对每个报销单风险等级中的训练样本进行数据预处理后，得到每个报销单风险等级中的待处理数据集，提高，训练样本的数据质量。

[0067] S42：对待处理数据集使用关联规则算法进行数据挖掘，得到每个报销单风险等级中的多个项集。

[0068] 在本发明实施例中，使用关联规则算法对每个待处理数据集进行数据挖掘，每个报销单训练样本为一个事务，记为T，并为每个训练样本标识对应的事务标识信息，事务的集为事务集合，记为D，报销单中每个属性为一个项，记为W，每个事务包括多个属性，项的集合为项集，项集$W=\{w_1,w_2,\ldots,w_j\}$，j为项集中项的个数。在对待处理数据集中的每个训练样本进行标识后，每个事务的标识信息对应一个项集，得到每个报销单风险等级中的多个项集。

[0069] S43：针对每个报销单风险等级，从该报销单风险等级中的项集中筛选出满足模型参数要求的目标项集，并根据该目标项集建立关联规则。

[0070] 在本发明实施例中，针对每个报销单风险等级训练样本的训练学习，预设对应的一组或多组支持度阈值和置信度阈值，从每个数据集中筛选出支持度大于等于支持度阈值的目标项集作为频繁项集，再由频繁项集产生初步规则，计算初步规则的置信度，获取置信度大于等于置信度阈值的规则，作为关联规则。

[0071] 需要说明的是，支持度是事务集合D中同时包含事务A和事务B的事务数所占总事务数的百分比，置信度是事务集合D中同时包含事务A和事务B的事务数与包含事务A事务数的百分比，规则可以用式子$A \Rightarrow B$表示，反映事务A与事务B之间的关联性。

[0072] 具体地，支持度可以按照公式(1)进行计算：

[0073] $$sup=\frac{\|\{T\in D\mid A\cup B\subseteq T\}\|}{\|D\|} \quad \text{公式（1）}$$

[0074] 其中，sup为支持度，$\|\{T\in D\mid A\cup B\subseteq T\}\|$为事务集合D中同时包含事务A和事务B的事务数，||D||为事务集合D中的事务数。

[0075] 具体地，置信度可以按照公式(2)进行计算：

[0076] $$confid(A \Rightarrow B) = \frac{\|\{T \in D \mid A \cup B \subseteq T\}\|}{\|\{T \in D \mid A \subseteq T\}\|}$$ 公式（2）

[0077] 其中，$confid(A \Rightarrow B)$为规则$A \Rightarrow B$的置信度，$\|\{T \in D \mid A \subseteq T\}\|$为事务集合D中包含事务A的事务数。

[0078] S44：根据关联规则和关联规则对应的模型参数要求，构建初始预测模型。

[0079] 具体地，在根据预设的支持度阈值和置信度阈值，使用关联规则算法对训练样本进行数据挖掘得到关联规则的基础上，以预设的支持度阈值和置信度阈值作为模型参数，对得到的关联规则进行汇总，生成初始预测模型，该初始预测模型用于预测测试样本中报销单的风险等级。

[0080] 在图2对应的实施例中，通过对每个报销单风险等级中的训练样本进行数据预处理，提高用于训练机器学习模型的数据的质量，再对每个报销单风险等级预设支持度阈值和置信度阈值作为模型参数要求，使用关联规则算法对每个报销单风险等级的训练样本进行数据挖掘，挖掘数据之间的关联性，得到关联规则，结合所预设的模型参数，生成初始预测模型，用于预测报销单的风险等级。采用按照不同报销单风险等级进行模型训练的方式能够学习到样本数据中所占比例较小的报销单数据的特征，避免这部分报销单数据被当做噪声处理而丢弃的情况，从而提高模型的精确度。

[0081] 在图1或图2对应的实施例的基础之上，下面通过一个具体的实施例对步骤S5中提及使用初始预测模型对测试样本进行模型预测，在从每个报销单风险等级中选择一组所述模型参数进行组合得到的每种组合方式下，计算每个报销单风险等级的预测成功率，以及该组合方式下的总预测成功率和测试时间的具体实现方法进行详细说明。

[0082] 请参阅图3，图3示出了本发明实施例提供的步骤S5的具体实现流程，详述如下：

[0083] S51：根据预设的N个报销单风险等级的定义，确定每个测试样本的报销单风险等级，以及每个报销单风险等级的测试样本数。

[0084] 在本发明实施例中，根据步骤S3预设的N个报销单风险等级的定义，确定测试样本中每个报销单的风险等级，并为每个测试样本标识对应的报销单风险等级的标识信息，根据该标识信息统计得到每个报销单风险等级的测试样本数。

[0085] 通过使用训练学习得到的初始预测模型对测试样本预测其报销单风险等级，校验和修正模型生成过程中产生的规则。

[0086] S52：按照公式(3)计算测试样本中每个报销单风险等级的概率：

[0087] $$P_i = \frac{R_i}{S}$$ 公式（3）

[0088] 其中，$i \in [1, N]$，P_i为测试样本中第i个报销单风险等级的概率，R_i为第i个报销单风险等级的测试样本数，S为测试样本的总数。

[0089] S53：从每个报销单风险等级中选择一组模型参数进行组合，得到L种组合方式，其中，L为正整数。

[0090] 在本发明实施例中，根据每个报销单风险等级中预设的一组或者多组模型参数要求，对每个报销单风险等级的训练样本进行关联规则挖掘，模型参数包括支持度阈值和置信度阈值。在每个报销单风险等级的训练样本中，根据每组模型参数均能够筛选得到对应

的满足该模型参数要求的关联规则。

[0091] 具体地，在N个报销单风险等级的多组模型参数中，从每个报销单风险等级中选择一组模型参数进行组合，得到L种不同的组合方式，其中，L为正整数。

[0092] 例如，当报销单风险等级预设为0、1、2、3四个等级，每个报销单风险等级的模型参数分别预设为：

[0093] $P_0:(sup_0,confid_0)=\{(x_{01},y_{01}),(x_{02},y_{02}),(x_{03},y_{03})\}$

[0094] $P_1:(sup_1,confid_1)=\{(x_{11},y_{11}),(x_{12},y_{12})\}$

[0095] $P_2:(sup_2,confid_2)=\{(x_{21},y_{21})\}$

[0096] $P_3:(sup_3,confid_3)=\{(x_{31},y_{31}),(x_{32},y_{32})\}$

[0097] 则组合方式为：

[0098] $L_1:\{(x_{01},y_{01}),(x_{11},y_{11}),(x_{21},y_{21}),(x_{31},y_{31})\}$

[0099] $L_2:\{(x_{01},y_{01}),(x_{12},y_{12}),(x_{21},y_{21}),(x_{31},y_{31})\}$

[0100] $L_3:\{(x_{01},y_{01}),(x_{11},y_{11}),(x_{21},y_{21}),(x_{32},y_{32})\}$

[0101] ⋮

[0102] 一共有3×2×1×2=12种组合方式。

[0103] S54：针对每种组合方式，按照概率由高到低的顺序，使用初始预测模型对测试样本进行报销单风险等级预测，得到每个测试样本的预测结果，并获取在该组合方式下的进行报销单风险等级预测的测试时间。

[0104] 在本发明实施例中，按照公式(3)计算得到的测试样本中每个报销单风险等级的概率，针对每种组合方式，按照概率由高到低的顺序，使用训练得到的初始预测模型对测试样本进行报销单风险等级预测，得到每个测试样本的预测结果，并获取在该组合方式下完成对全部测试样本进行报销单风险等级预测的测试时间，一共得到L种组合方式下每个测试样本的预测结果，以及对应的测试时间，用于进一步分析初始预测模型的精确度。

[0105] S55：将每个测试样本的预测结果与该测试样本的报销单风险等级进行对比，若两者相同则确认该测试样本预测成功，并统计在每个报销单风险等级下的测试样本预测成功的个数。

[0106] 具体地，根据步骤S54预测得到的每个测试样本的报销单风险等级的预测结果，与该测试样本的报销单风险等级的标识信息进行对比分析，若两者报销单风险等级相同则确认该测试样本预测成功，若两者报销单风险等级不同则确认该测试样本预测失败。

[0107] 统计在每个报销单风险等级下的测试样本预测成功的个数，用于计算每种组合方式下的每个报销单风险等级的预测成功率。

[0108] S56：按照公式(4)计算每种组合方式下的每个报销单风险等级的预测成功率：

[0109] $$hitrate_i=\frac{M_i}{R_i}$$ 公式(4)

[0110] 其中，$hitrate_i$为第i个报销单风险等级的预测成功率，M_i为第i个报销单风险等级下的测试样本预测成功的个数，R_i为第i个报销单风险等级的测试样本数。

[0111] S57：按照公式(5)计算每种组合方式下的总预测成功率：

[0112] $$hitRate=\frac{\sum_{i=1}^{N}M_i}{S}$$ 公式（5）

[0113] 其中，hitRate为总预测成功率，M_i为第i个报销单风险等级下的测试样本预测成功的个数，S为测试样本的总数。

[0114] 例如，当报销单风险等级预设为0、1、2、3四个等级，对采集的605790个报销单测试样本，进行报销单风险等级预测，根据预设的报销单风险等级的定义，标识并统计每个报销单风险等级的测试样本数，其中，0风险等级的报销单样本数有561627个，1风险等级的报销单样本数有34818个，2风险等级的报销单样本数有13个，3风险等级的报销单样本数有9332个。

[0115] 当使用sup_0＝0.8，$confid_0$＝0.95，sup_1＝0.4，$confid_1$＝0.7，sup_2＝0.4，$confdi_2$＝0.95，sup_3＝0.4，$confid_3$＝0.7作为预设的模型参数要求，对测试样本进行报销单风险等级预测，并将每个测试样本的预测结果与该测试样本的标识信息标识的报销单风险等级进行对比后，得到各风险等级预测成功的结果为：0风险等级的报销单个数为561527个，1风险等级的报销单个数为30821个，2风险等级的报销单个数为1个，3风险等级的报销单个数为1532个，总共预测成功的报销单个数为593881个。

[0116] 按照公式(4)计算得到：0风险等级的报销单预测成功率$hitrate_0$为561527/561627＝99.98219％，1风险等级的报销单预测成功率$hitrate_1$为30821/34818＝88.52285％，2风险等级的报销单预测成功率$hitrate_2$为1/13＝7.69230％，3风险等级的报销单预测成功率$hitrate_3$为1532/9332＝16.41663％。按照公式(5)计算得到总预测成功率hitRate为593881/605790＝98.03413％。

[0117] 在图3对应的实施例中，通过计算测试样本中每个报销单风险等级的概率，从每个报销单风险等级中选择一组模型参数进行组合，按照概率由高到低的顺序，使用初始预测模型对测试样本进行报销单风险等级预测，检验初始预测模型的识别率，提高了模型测试的效率。将每个测试样本的预测结果与预先标识的报销单风险等级进行对比，得到在每个报销单风险等级下的测试样本预测成功的个数，并计算每种组合方式下的每个报销单风险等级的预测成功率和总预测成功率，以便根据预测成功率、测试时间和总预测成功率进一步分析初始预测模型的精确度，进行校验和修正模型生成过程中产生的规则，实现对初始预测模型的优化，得到精准的目标预测模型，使得目标预测模型能够辅助工作人员精准高效地识别报销单的风险级别，有效提高预测报销单风险等级的准确率。

[0118] 在图3对应的实施例的基础之上，下面通过一个具体的实施例对步骤S6中提及的对模型参数、预测成功率、测试时间和总预测成功率进行回归分析，得到目标预测模型的具体实现方法进行详细说明。

[0119] 请参阅图4，图4示出了本发明实施例提供的步骤S6的具体实现流程，详述如下：

[0120] S61：将每个报销单风险等级中的模型参数，以及预测成功率和测试时间作为设计变量，将总预测成功率作为目标变量，使用设计变量和目标变量进行函数拟合，得到拟合函数。

[0121] 在本发明实施例中，将每个报销单风险等级中的模型参数，以及预测成功率和测试时间作为设计变量，将总预测成功率作为目标变量，使用设计变量和目标变量进行函数

拟合，以每种组合方式下对测试样本预测的结果作为一组数据，对步骤S53中得到L组结果数据进行拟合，拟合的方式具体可以表示为：

[0122]

$$\langle sup_0, confid_0, hitrate_0, sup_1, confid_1, hitrate_1, ..., sup_{n-1}, confid_{n-1}, hitrate_{n-1}, t, \delta \rangle \Rightarrow hitRate$$

[0123] 其中，n表示报销单风险等级的个数，t为每种组合方式下完成全部测试样本的报销单风险等级预测的测试时间，δ为运行配置参数，δ是根据系统软硬件配置预设的一个常数，其具体可以根据实际应用的需要进行设置，此处不做限制。

[0124] 通过参数t和参数δ的组合可以对拟合过程的程序执行效率进行调节。

[0125] 具体地，函数拟合的方式可以使用办公软件(Microsoft Excel，excel)或者数学软件(Matrix Laboratory，matlab)等工具进行拟合，对包含支持度和置信度的模型参数、预测成功率和总预测成功率等离散型数据进行非线性回归分析，寻找设计变量与目标变量之间的关系，并根据该关系确定拟合函数的表达式f(x)，从而拟合出与离散型数据相吻合的离散方程。

[0126] S62：对拟合函数进行求解，根据求解结果将总预测成功率最高并且模型参数的值最高的一组设计变量作为模型配置参数，并根据模型配置参数构建目标预测模型，其中，目标预测模型的模型精确度为最高的总预测成功率。

[0127] 具体地，对拟合得到的拟合函数f(x)进行求解，根据求解结果将总预测成功率最高并且模型参数的值最高的一组设计变量作为模型配置参数，其中，支持度阈值和置信度阈值越大，得到的关联规则越准确，并根据模型配置参数以及满足该模型配置参数要求的关联规则构建目标预测模型。

[0128] 使用目标预测模型预测报销单风险等级时，以目标预测模型的模型精确度为最高的总预测成功率，作为评价模型质量的标准，模型的总预测成功率越高，模型精确度也越高。

[0129] 在图4对应的实施例中，通过将每个报销单风险等级中的模型参数，以及预测成功率和测试时间作为设计变量，将总预测成功率作为目标变量，作非线性回归分析进行函数拟合，以寻找设计变量与目标变量之间的关系，得到拟合函数的表达式，对拟合函数进行求解，根据求解结果将总预测成功率最高并且模型参数的值最高的一组设计变量作为模型配置参数，提高关联规则的准确性，并根据模型配置参数以及对应满足模型参数要求的关联规则进行构建目标预测模型，从而提高目标预测模型进行预测的准确率。

[0130] 在图4对应的实施例的基础之上，在步骤S6中提及的对模型参数、预测成功率、测试时间和总预测成功率进行回归分析，得到目标预测模型之后，还可以进一步的使用交叉验证的方法选择合理模型。

[0131] 如图5所示，该报销单风险预测方法还包括：

[0132] S71：将样本数据分割成K个子样本数据。

[0133] 在本发明实施例中，使用交叉验证的精度测试方法对拟合优化后的目标预测模型进行验证，将采集到的报销单样本数据采用随机分割的方式分割成K个子样本数据，通过机器学习的方式进行多次目标预测模型的构建，以及对构建得到的目标预测模型进行精确度评价，避免训练得到的模型出现过拟合的情况，其中，K为正整数。

[0134] 过拟合是指拟合函数与训练样本高度吻合，但是求解得到的模型配置参数用于预

测测试样本的报销单风险等级成功率却不高的情况。

[0135] 需要说明的是，交叉验证可以采用留出验证(holdout cross validation，holdout)，K折交叉验证(k-fold cross validatio)或者留一验证(leave-one-out cross validation，loocv)等方式，将样本数据切割成较小子样本之后，获取其中大部分样本进行模型构建，剩余的小部分样本用于对建立的模型进行测试。

[0136] S72：从K个子样本数据中，选择一个子样本数据作为测试样本，剩余K-1个子样本数据作为训练样本，进行模型训练、模型预测和回归分析，得到K个目标预测模型和每个目标预测模型的模型精确度，其中，K为正整数。

[0137] 在本发明实施例中，从K个子样本数据中，选择其中一个子样本数据作为验证模型的测试样本，其他K-1个子样本数据作为特征学习的训练样本，执行步骤S3至步骤S6的过程，进行模型训练、模型预测和回归分析，，完成一次目标预测模型的构建，得到目标预测模型及其模型精确度。按照该构建方式，将K个子样本数据中的每个子样本数据作为测试样本进行一次目标预测模型的构建，得到K个结果，包括K个目标预测模型和每个目标预测模型的模型精确度。

[0138] S73：将模型精确度最高的目标预测模型作为合理模型。

[0139] 具体地，对得到K个目标预测模型和每个目标预测模型的模型精确度进行对比分析，将模型精确度最高的目标预测模型作为合理模型，从而得到可靠稳定的合理模型。

[0140] 合理模型一方面能够拟合样本数据，另一方面能够以高准确率进行新的报销单数据的风险等级预测，该合理模型能够预测出准确的报销单风险等级，并将报销单数据和对应的报销单风险等级存储到报销单数据库中。

[0141] 进一步地，按照预设的时间间隔，该时间间隔可以是1个月、2个月或者其它时间范围，每隔预定的时间间隔从报销单数据库中随机获取历史报销单信息，重复执行步骤S1至步骤S6过程，完成自主机器学习，得到更新后的目标预测模型，从而进一步地优化模型的精确度，提高报销单风险等级预测成功率，实现报销单风险等级的精准预测。

[0142] 在图5对应的实施例中，通过交叉验证的方法对模型精度进行测试，运用随机分割的子样本数据进行多次训练和验证，避免训练得到的目标预测模型出现拟合不当的情况，并从验证结果中选择模型精确度最高的目标预测模型作为合理模型，即能够拟合样本数据，又能够以高准确率实现对新的报销单数据的预测，提高了报销单风险等级预测的准确率。

[0143] 应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本发明实施例的实施过程构成任何限定。

[0144] 实施例2

[0145] 对应于实施例1中的报销单风险预测方法，图6示出与实施例1所示的报销单风险预测方法一一对应的报销单风险预测装置，为了便于说明，仅示出了与本发明实施例相关的部分。

[0146] 如图6所示，该报销单风险预测装置包括样本数据采集模块61、样本数据划分模块62、风险等级预设模块63、初始预测模型获取模块64、初始预测模型测试模块65和目标预测模型获取模块66。各功能模块详细说明如下：

[0147] 样本数据采集模块61，用于获取历史报销单信息，并将历史报销单信息作为样本数据；

[0148] 第一划分模块62，用于将样本数据按照预设的比例划分为训练样本和测试样本；

[0149] 风险等级预设模块63，用于根据预设的N个报销单风险等级的定义，确定每个训练样本的报销单风险等级，其中，N为正整数；

[0150] 初始预测模型获取模块64，用于针对每个报销单风险等级中的训练样本，使用关联规则算法进行模型训练，得到初始预测模型，其中，初始预测模型包括每个报销单风险等级中满足预设的模型参数要求的关联规则，模型参数包括支持度和置信度；

[0151] 初始预测模型测试模块65，用于使用初始预测模型对测试样本进行模型预测，在从每个报销单风险等级中选择一组模型参数进行组合得到的每种组合方式下，计算每个报销单风险等级的预测成功率，以及每种组合方式下的总预测成功率和测试时间；

[0152] 目标预测模型获取模块66，用于对模型参数、预测成功率、测试时间和总预测成功率进行回归分析，得到目标预测模型。

[0153] 进一步地，初始预测模型获取模块64包括：

[0154] 数据预处理单元641，用于对每个报销单风险等级中的训练样本进行数据预处理，得到每个报销单风险等级中的待处理数据集；

[0155] 训练样本挖掘单元642，用于对待处理数据集使用关联规则算法进行数据挖掘，得到每个报销单风险等级中的多个项集；

[0156] 关联规则获取单元643，用于针对每个报销单风险等级，从该报销单风险等级中的项集中筛选出满足模型参数要求的目标项集，并根据该目标项集建立关联规则；

[0157] 初始预测模型构建单元644，用于根据关联规则和关联规则对应的模型参数要求，构建初始预测模型。

[0158] 进一步地，初始预测模型测试模块65包括：

[0159] 第一统计单元651，用于根据预设的N个报销单风险等级的定义，确定每个测试样本的报销单风险等级，以及每个报销单风险等级的测试样本数；

[0160] 第一计算单元652，用于按照如下公式计算测试样本中每个报销单风险等级的概率：

[0161] $$P_i=\frac{R_i}{S}$$

[0162] 其中，$i\in[1,N]$，P_i为测试样本中第i个报销单风险等级的概率，R_i为第i个报销单风险等级的测试样本数，S为测试样本的总数；

[0163] 预测方式组合单元653，用于从每个报销单风险等级中选择一组模型参数进行组合，得到L种组合方式，其中，L为正整数；

[0164] 测试样本预测单元654，用于针对每种组合方式，按照概率由高到低的顺序，使用初始预测模型对测试样本进行报销单风险等级预测，得到每个测试样本的预测结果，并获取在该组合方式下的进行报销单风险等级预测的测试时间；

[0165] 第二统计单元655，用于将每个测试样本的预测结果与该测试样本的报销单风险等级进行对比，若两者相同则确认该测试样本预测成功，并统计在每种组合方式下每个报销单风险等级下的测试样本预测成功的个数；

[0166] 第二计算单元656，用于按照如下公式计算每种组合方式下每个报销单风险等级的预测成功率：

[0167] $$hitrate_i = \frac{M_i}{R_i}$$

[0168] 其中，$hitrate_i$为第i个报销单风险等级的预测成功率，M_i为第i个报销单风险等级下的测试样本预测成功的个数；

[0169] 第三计算单元657，用于按照如下公式计算每种组合方式下的总预测成功率：

[0170] $$hitRate = \frac{\sum_{i=1}^{N} M_i}{S}$$

[0171] 其中，hitRate为总预测成功率。

[0172] 进一步地，目标预测模型获取模块66包括：

[0173] 数据拟合单元661，用于将每个报销单风险等级中的模型参数，以及预测成功率和测试时间作为设计变量，将总预测成功率作为目标变量，使用设计变量和目标变量进行函数拟合，得到拟合函数；

[0174] 目标预测模型构建单元662，用于对拟合函数进行求解，根据求解结果将总预测成功率最高并且模型参数的值最高的一组设计变量作为模型配置参数，并根据模型配置参数构建目标预测模型，其中，目标预测模型的模型精确度为最高的总预测成功率。

[0175] 进一步地，该报销单风险预测装置还包括：

[0176] 第二划分模块67，用于将样本数据分割成K个子样本数据；

[0177] 交叉验证模块68，用于从K个子样本数据中，选择一个子样本数据作为测试样本，剩余K-1个子样本数据作为训练样本，进行模型训练、模型预测和回归分析，得到K个目标预测模型和每个目标预测模型的模型精确度，其中，K为正整数；

[0178] 合理模型获取模块69，用于将模型精确度最高的目标预测模型作为合理模型。

[0179] 本实施例提供的一种报销单风险预测装置中各模块实现各自功能的过程，具体可参考前述方法实施例1的描述，此处不再赘述。

[0180] 实施例3

[0181] 本实施例提供一计算机可读存储介质，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器执行时实现实施例1中报销单风险预测方法，为避免重复，这里不再赘述。或者，该计算机程序被处理器执行时实现实施例2中报销单风险预测装置中各模块/单元的功能，为避免重复，这里不再赘述。

[0182] 可以理解地，所述计算机可读存储介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、电载波信号和电信信号等。

[0183] 实施例4

[0184] 图7是本发明一实施例提供的终端设备的示意图。如图7所示，该实施例的终端设备7包括：处理器70、存储器71以及存储在存储器71中并可在处理器70上运行的计算机程序72，例如报销单风险预测程序。处理器70执行计算机程序72时实现上述各个报销单风险预

测方法实施例中的步骤，例如图1所示的步骤S1至步骤S6。或者，处理器70执行计算机程序72时实现上述各装置实施例中各模块/单元的功能，例如图6所示模块61至步骤66的功能。

[0185] 示例性的，计算机程序72可以被分割成一个或多个模块/单元，一个或者多个模块/单元被存储在存储器71中，并由处理器70执行，以完成本发明。一个或多个模块/单元可以是能够完成特定功能的一系列计算机程序指令段，该指令段用于描述计算机程序72在终端设备7中的执行过程。计算机程序72可以被分割成样本数据采集模块、样本数据划分模块、风险等级预设模块、初始预测模型获取模块、初始预测模型测试模块和目标预测模型获取模块。各模块详细说明如下：

[0186] 样本数据采集模块，用于获取历史报销单信息，并将历史报销单信息作为样本数据；

[0187] 第一划分模块，用于将样本数据按照预设的比例划分为训练样本和测试样本；

[0188] 风险等级预设模块，用于根据预设的N个报销单风险等级的定义，确定每个训练样本的报销单风险等级，其中，N为正整数；

[0189] 初始预测模型获取模块，用于针对每个报销单风险等级中的训练样本，使用关联规则算法进行模型训练，得到初始预测模型，其中，初始预测模型包括每个报销单风险等级中满足预设的模型参数要求的关联规则，模型参数包括支持度和置信度；

[0190] 初始预测模型测试模块，用于使用初始预测模型对测试样本进行模型预测，在从每个报销单风险等级中选择一组模型参数进行组合得到的每种组合方式下，计算每个报销单风险等级的预测成功率，以及每种组合方式下的总预测成功率和测试时间；

[0191] 目标预测模型获取模块，用于对模型参数、预测成功率、测试时间和总预测成功率进行回归分析，得到目标预测模型。

[0192] 进一步地，初始预测模型获取模块包括：

[0193] 数据预处理单元，用于对每个报销单风险等级中的训练样本进行数据预处理，得到每个报销单风险等级中的待处理数据集；

[0194] 训练样本挖掘单元，用于对待处理数据集使用关联规则算法进行数据挖掘，得到每个报销单风险等级中的多个项集；

[0195] 关联规则获取单元，用于针对每个报销单风险等级，从该报销单风险等级中的项集中筛选出满足模型参数要求的目标项集，并根据该目标项集建立关联规则；

[0196] 初始预测模型构建单元，用于根据关联规则和关联规则对应的模型参数要求，构建初始预测模型。

[0197] 进一步地，初始预测模型测试模块包括：

[0198] 第一统计单元，用于根据预设的N个报销单风险等级的定义，确定每个测试样本的报销单风险等级，以及每个报销单风险等级的测试样本数；

[0199] 第一计算单元，用于按照如下公式计算测试样本中每个报销单风险等级的概率：

[0200] $$P_i=\frac{R_i}{S}$$

[0201] 其中，$i\in[1,N]$，P_i为测试样本中第i个报销单风险等级的概率，R_i为第i个报销单风险等级的测试样本数，S为测试样本的总数；

[0202] 预测方式组合单元，用于从每个报销单风险等级中选择一组模型参数进行组合，

得到L种组合方式，其中，L为正整数；

[0203] 测试样本预测单元，用于针对每种组合方式，按照概率由高到低的顺序，使用初始预测模型对测试样本进行报销单风险等级预测，得到每个测试样本的预测结果，并获取在该组合方式下的进行报销单风险等级预测的测试时间；

[0204] 第二统计单元，用于将每个测试样本的预测结果与该测试样本的报销单风险等级进行对比，若两者相同则确认该测试样本预测成功，并统计在每种组合方式下每个报销单风险等级下的测试样本预测成功的个数；

[0205] 第二计算单元，用于按照如下公式计算每种组合方式下每个报销单风险等级的预测成功率：

[0206] $$hitrate_i = \frac{M_i}{R_i}$$

[0207] 其中，$hitrate_i$为第i个报销单风险等级的预测成功率，M_i为第i个报销单风险等级下的测试样本预测成功的个数；

[0208] 第三计算单元，用于按照如下公式计算每种组合方式下的总预测成功率：

[0209] $$hitRate = \frac{\sum_{i=1}^{N} M_i}{S}$$

[0210] 其中，hitRate为总预测成功率。

[0211] 进一步地，目标预测模型获取模块包括：

[0212] 数据拟合单元，用于将每个报销单风险等级中的模型参数，以及预测成功率和测试时间作为设计变量，将总预测成功率作为目标变量，使用设计变量和目标变量进行函数拟合，得到拟合函数；

[0213] 目标预测模型构建单元，用于对拟合函数进行求解，根据求解结果将总预测成功率最高并且模型参数的值最高的一组设计变量作为模型配置参数，并根据模型配置参数构建目标预测模型，其中，目标预测模型的模型精确度为最高的总预测成功率。

[0214] 进一步地，计算机程序72还可以被分割成：

[0215] 第二划分模块，用于将样本数据分割成K个子样本数据；

[0216] 交叉验证模块，用于从K个子样本数据中，选择一个子样本数据作为测试样本，剩余K-1个子样本数据作为训练样本，进行模型训练、模型预测和回归分析，得到K个目标预测模型和每个目标预测模型的模型精确度，其中，K为正整数；

[0217] 合理模型获取模块，用于将模型精确度最高的目标预测模型作为合理模型。

[0218] 终端设备7可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。终端设备7可包括，但不仅限于，处理器70、存储器71。本领域技术人员可以理解，图7仅仅是终端设备7的示例，并不构成对终端设备7的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如终端设备7还可以包括输入输出设备、网络接入设备、总线等。

[0219] 所称处理器70可以是中央处理单元(Central Processing Unit,CPU)，还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-

Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0220] 存储器71可以是终端设备7的内部存储单元,例如终端设备7的硬盘或内存。存储器71也可以是终端设备7的外部存储设备,例如终端设备7上配备的插接式硬盘,智能存储卡(Smart Media Card,SMC),安全数字(Secure Digital,SD)卡,闪存卡(Flash Card)等。进一步地,存储器71还可以既包括终端设备7的内部存储单元也包括外部存储设备。存储器71用于存储计算机程序以及终端设备7所需的其他程序和数据。存储器71还可以用于暂时地存储已经输出或者将要输出的数据。

[0221] 所属领域的技术人员可以清楚地了解到,为了描述的方便和简洁,仅以上述各功能单元、模块的划分进行举例说明,实际应用中,可以根据需要而将上述功能分配由不同的功能单元、模块完成,即将所述装置的内部结构划分成不同的功能单元或模块,以完成以上描述的全部或者部分功能。

[0222] 以上所述实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围,均应包含在本发明的保护范围之内。

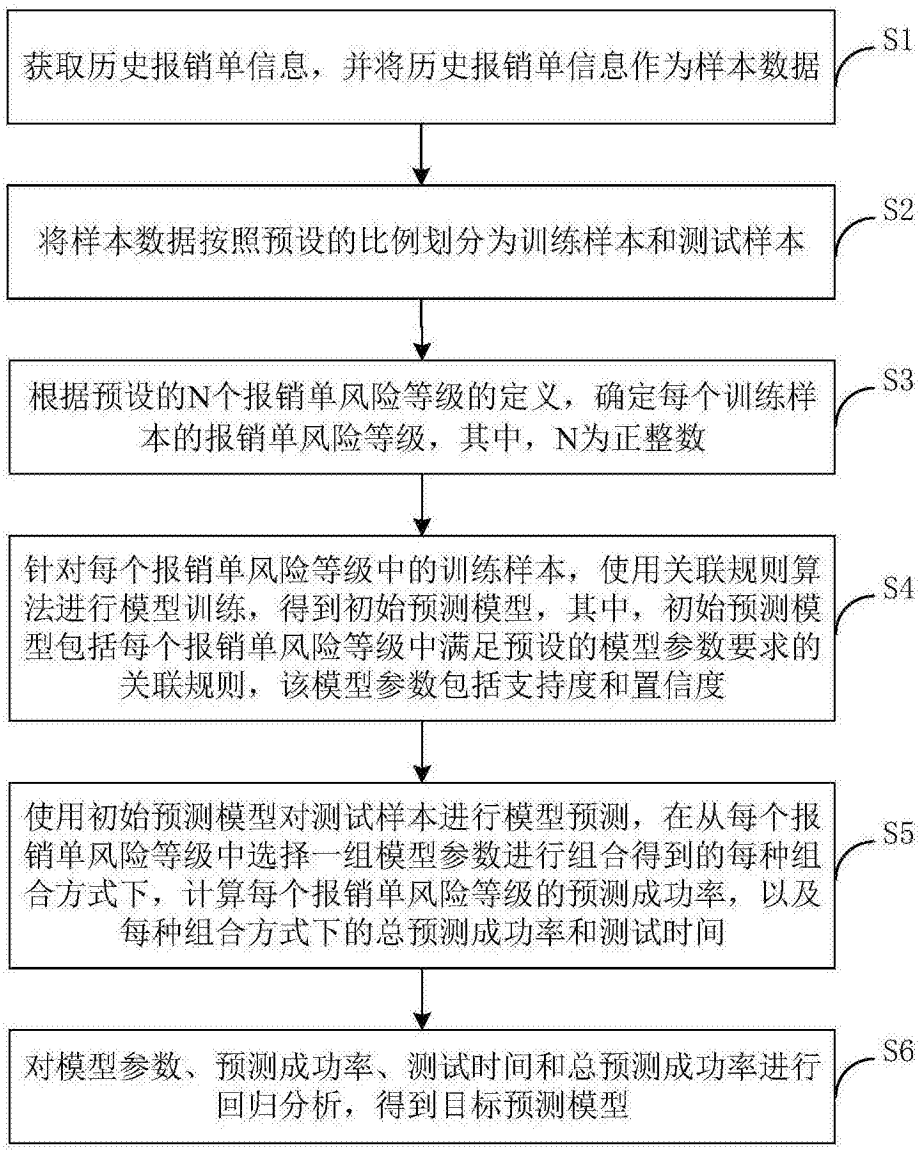

获取历史报销单信息，并将历史报销单信息作为样本数据
S1
将样本数据按照预设的比例划分为训练样本和测试样本
S2
根据预设的N个报销单风险等级的定义，确定每个训练样本的报销单风险等级，其中，N为正整数
S3
针对每个报销单风险等级中的训练样本，使用关联规则算法进行模型训练，得到初始预测模型，其中，初始预测模型包括每个报销单风险等级中满足预设的模型参数要求的关联规则，该模型参数包括支持度和置信度
S4
使用初始预测模型对测试样本进行模型预测，在从每个报销单风险等级中选择一组模型参数进行组合得到的每种组合方式下，计算每个报销单风险等级的预测成功率，以及每种组合方式下的总预测成功率和测试时间
S5
对模型参数、预测成功率、测试时间和总预测成功率进行回归分析，得到目标预测模型
S6

图1

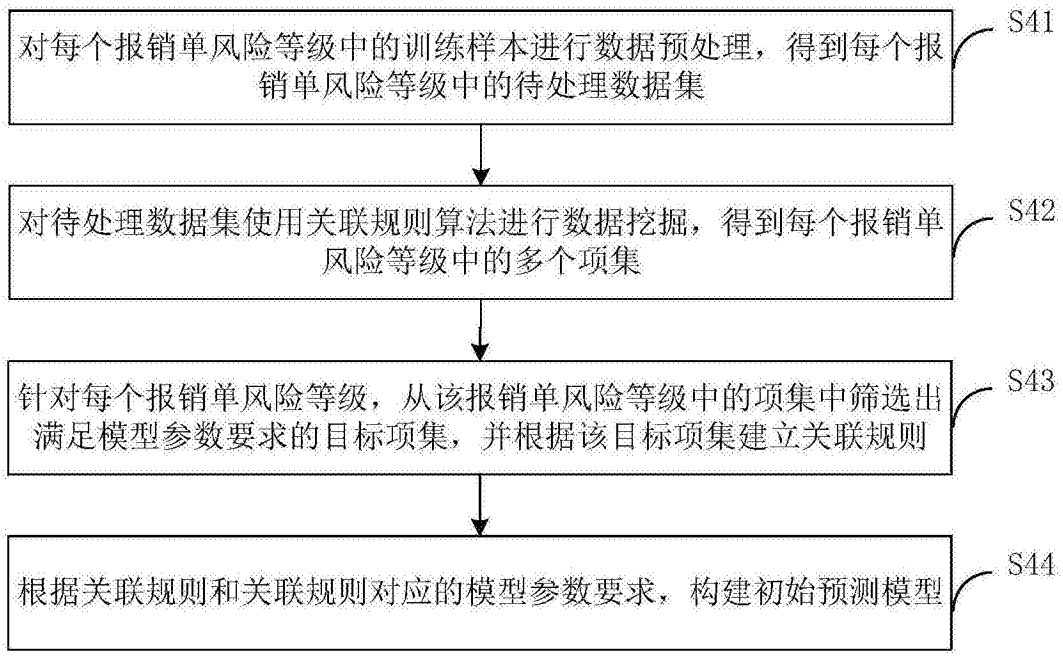

对每个报销单风险等级中的训练样本进行数据预处理，得到每个报销单风险等级中的待处理数据集
S41
对待处理数据集使用关联规则算法进行数据挖掘，得到每个报销单风险等级中的多个项集
S42
针对每个报销单风险等级，从该报销单风险等级中的项集中筛选出满足模型参数要求的目标项集，并根据该目标项集建立关联规则
S43
根据关联规则和关联规则对应的模型参数要求，构建初始预测模型
S44

图2

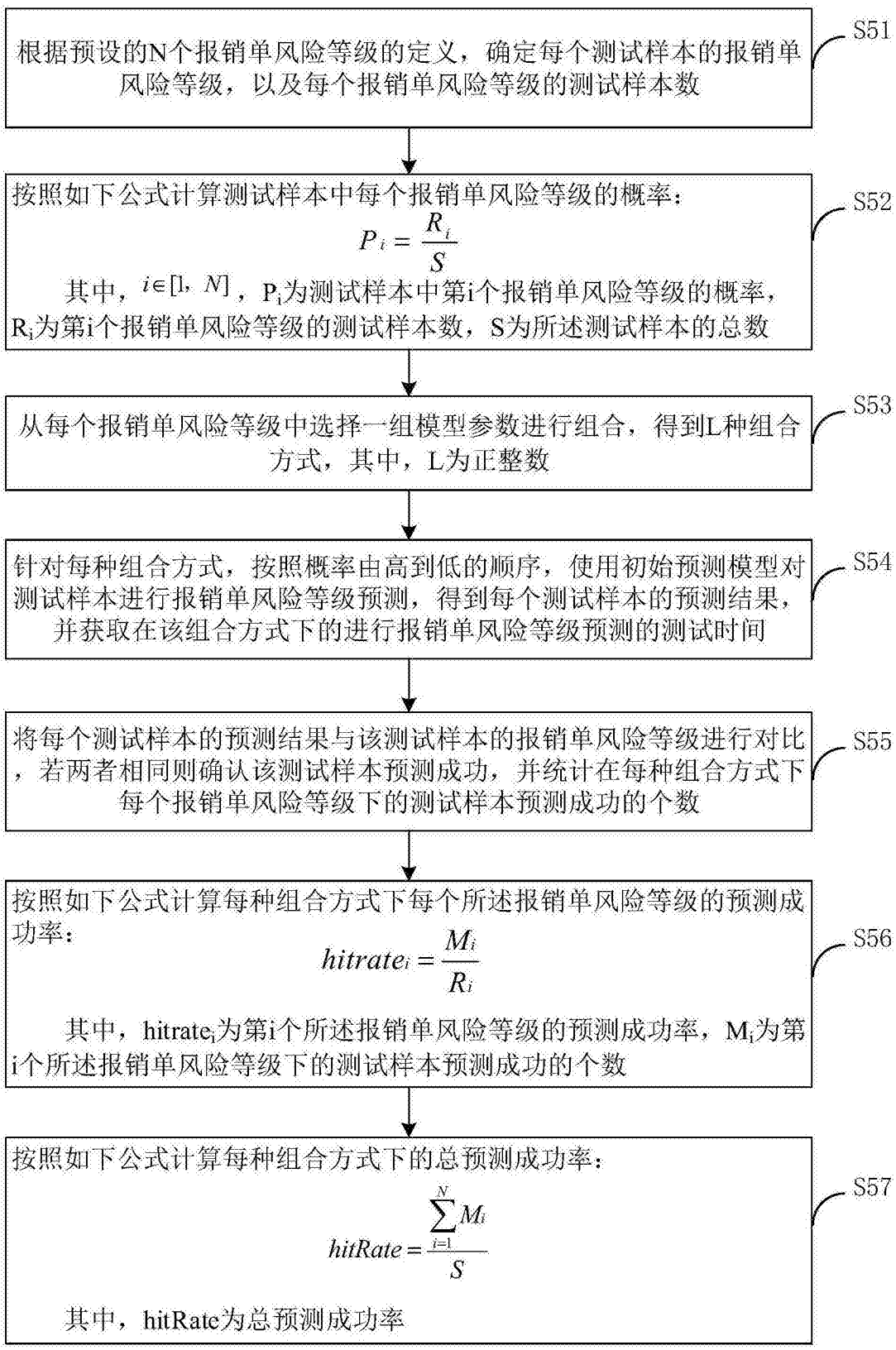

根据预设的N个报销单风险等级的定义，确定每个测试样本的报销单风险等级，以及每个报销单风险等级的测试样本数
S51
按照如下公式计算测试样本中每个报销单风险等级的概率：
$P_i = \frac{R_i}{S}$
其中，$i \in [1, N]$，P_i为测试样本中第i个报销单风险等级的概率，R_i为第i个报销单风险等级的测试样本数，S为所述测试样本的总数
S52
从每个报销单风险等级中选择一组模型参数进行组合，得到L种组合方式，其中，L为正整数
S53
针对每种组合方式，按照概率由高到低的顺序，使用初始预测模型对测试样本进行报销单风险等级预测，得到每个测试样本的预测结果，并获取在该组合方式下的进行报销单风险等级预测的测试时间
S54
将每个测试样本的预测结果与该测试样本的报销单风险等级进行对比，若两者相同则确认该测试样本预测成功，并统计在每种组合方式下每个报销单风险等级下的测试样本预测成功的个数
S55
按照如下公式计算每种组合方式下每个所述报销单风险等级的预测成功率：
$hitrate_i = \frac{M_i}{R_i}$
其中，$hitrate_i$为第i个所述报销单风险等级的预测成功率，M_i为第i个所述报销单风险等级下的测试样本预测成功的个数
S56
按照如下公式计算每种组合方式下的总预测成功率：
$hitRate = \frac{\sum_{i=1}^{N} M_i}{S}$
其中，hitRate为总预测成功率
S57

图3

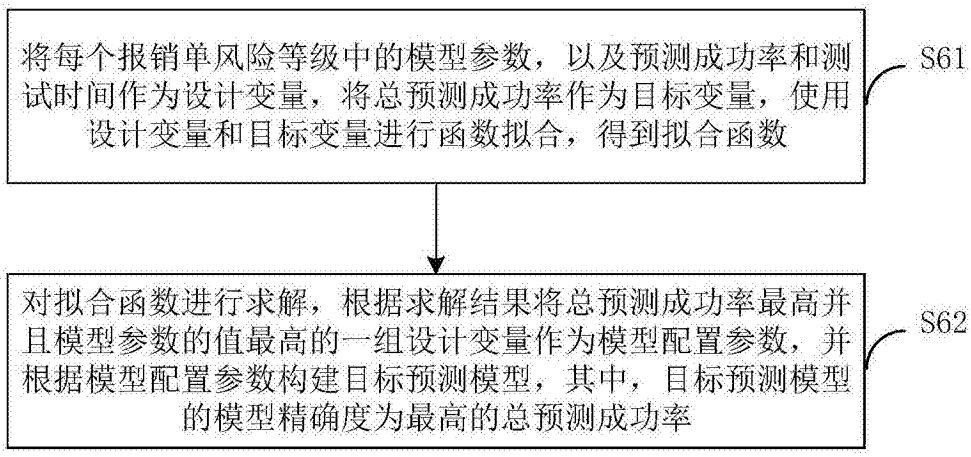

将每个报销单风险等级中的模型参数，以及预测成功率和测试时间作为设计变量，将总预测成功率作为目标变量，使用设计变量和目标变量进行函数拟合，得到拟合函数
S61
对拟合函数进行求解，根据求解结果将总预测成功率最高并且模型参数的值最高的一组设计变量作为模型配置参数，并根据模型配置参数构建目标预测模型，其中，目标预测模型的模型精确度为最高的总预测成功率
S62

图4

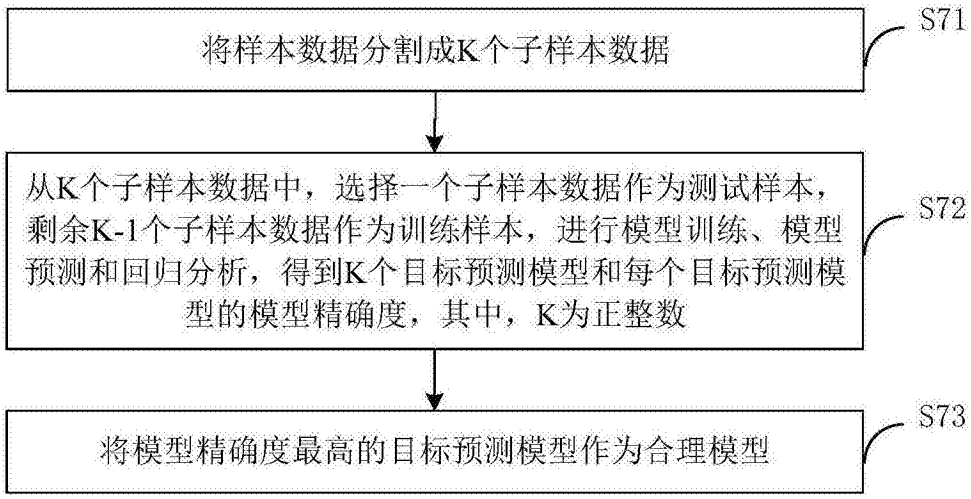

将样本数据分割成K个子样本数据
S71
从K个子样本数据中，选择一个子样本数据作为测试样本，剩余K-1个子样本数据作为训练样本，进行模型训练、模型预测和回归分析，得到K个目标预测模型和每个目标预测模型的模型精确度，其中，K为正整数
S72
将模型精确度最高的目标预测模型作为合理模型
S73

图5

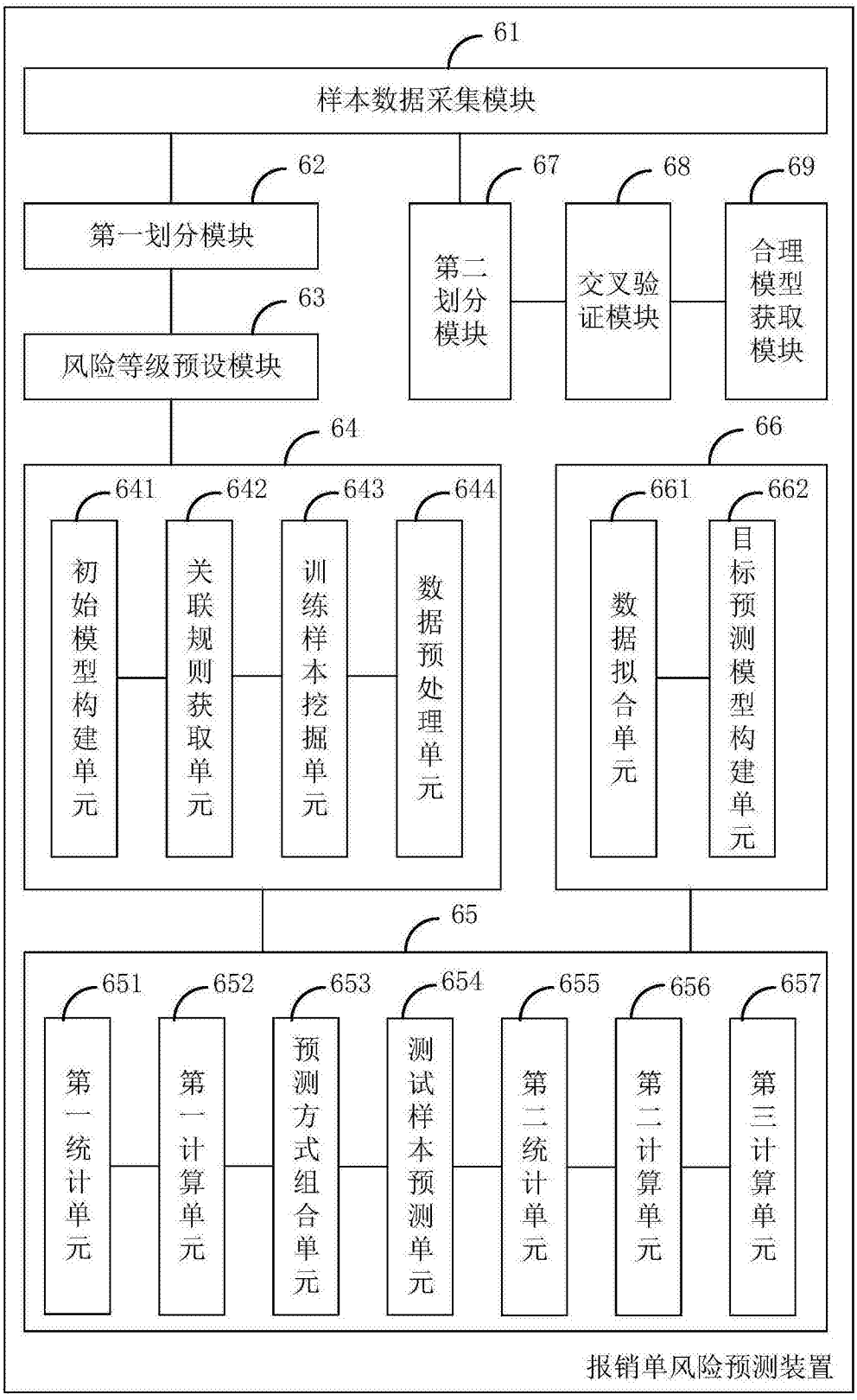

61
样本数据采集模块
62
第一划分模块
63
风险等级预设模块
67
第二划分模块
68
交叉验证模块
69
合理模型获取模块
64
641
初始模型构建单元
642
关联规则获取单元
643
训练样本挖掘单元
644
数据预处理单元
66
661
数据拟合单元
662
目标预测模型构建单元
65
651
第一统计单元
652
第一计算单元
653
预测方式组合单元
654
测试样本预测单元
655
第二统计单元
656
第二计算单元
657
第三计算单元
报销单风险预测装置

图6

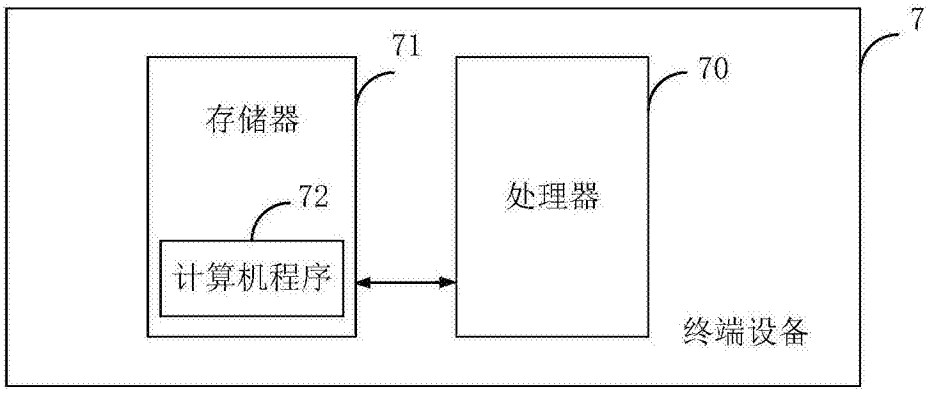


图7