

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 7 月 16 日 (16.07.2020)



(10) 国际公布号
WO 2020/143233 A1

(51) 国际专利分类号:
G06F 17/50 (2006.01)

(21) 国际申请号: PCT/CN2019/103489

(22) 国际申请日: 2019 年 8 月 30 日 (30.08.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910012412.X 2019 年 1 月 7 日 (07.01.2019) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。

(72) 发明人: 季洁璐(JI, Jielu); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。 何友鑫(HE, Youxin); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。 彭琛(PENG, Chen); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。 汪伟(WANG, Wei); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。

(74) 代理人: 广州华进联合专利商标代理有限公司(ADVANCE CHINA IP LAW OFFICE); 中国广东省广州市天河区珠江东路 6 号 4501 房(部位: 自编 01-03 和 08-12 单元)(仅限办公用途), Guangdong 510623 (CN)。

(54) Title: METHOD AND DEVICE FOR BUILDING SCORECARD MODEL, COMPUTER APPARATUS AND STORAGE MEDIUM

(54) 发明名称: 评分卡模型的建立方法、装置、计算机设备和存储介质

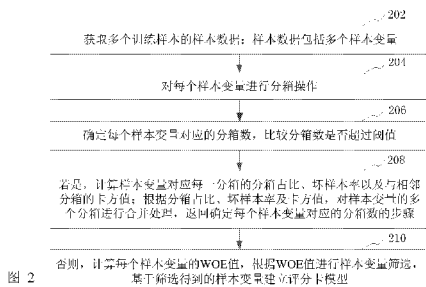


图 2

202 Acquire sample data of multiple training samples, the sample data comprising multiple sample variables
204 Perform binning with respect to each sample variable
206 Determine the number of bins corresponding to each sample variable, and compare the same to determine whether the number of bins exceeds a threshold
208 If so, calculate to obtain a ratio of the total number of samples corresponding to a given variable in a respective bin to the total number of samples, a bad sample rate, and a chi-squared value based on an adjacent bin, combine multiple bins associated with a given sample variable according to the ratio, the bad sample rate and the chi-squared value, and return to the step of determining the number of bins corresponding to each sample variable
210 If not, calculate to obtain WOE values of the respective sample variables, perform filtering of the sample variables according to the WOE values, and establish a scorecard model on the basis of the sample variables obtained from filtering

(57) Abstract: A method for building a scorecard model comprises: acquiring sample data of multiple training samples, the sample data comprising multiple sample variables; performing binning with respect to each sample variable; determining the number of bins corresponding to each sample variable, and comparing the same to determine whether the number of bins exceeds a threshold; and if so, calculating to obtain a ratio of the total number of samples corresponding to a given variable in a respective bin to the total number of samples, a bad sample rate, and a chi-squared value based on an adjacent bin, combining multiple bins associated with a given sample variable according to the ratio, the bad sample rate and the chi-squared value, and returning to the step of determining the number of bins corresponding to each sample variable; or if not, calculating to obtain WOE values of the respective sample variables, performing filtering of the sample variables according to the WOE values, and establishing a scorecard model on the basis of the sample variables obtained from filtering.

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 一种评分卡模型的建立方法, 包括: 获取多个训练样本的样本数据; 样本数据包括多个样本变量; 对每个样本变量进行分箱操作; 确定每个样本变量对应的分箱数, 比较分箱数是否超过阈值; 若是, 计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱的卡方值; 根据分箱占比、坏样本率及卡方值, 对样本变量的多个分箱进行合并处理, 返回确定每个样本变量对应的分箱数的步骤; 否则, 计算每个样本变量的WOE值, 根据WOE值进行样本变量筛选, 基于筛选得到的样本变量建立评分卡模型。

评分卡模型的建立方法、装置、计算机设备和存储介质

本申请要求于 2019 年 01 月 07 日提交中国专利局，申请号为 201910012412X，申请名称为“评分卡模型的建立方法、装置、计算机设备和存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5

技术领域

本申请涉及一种评分卡模型的建立方法、装置、计算机设备和存储介质。

背景技术

10 通常在建立分类模型时，需要对连续变量离散化，特征离散化可以使模型更稳定，降低模型过拟合的风险。比如，在建立评分卡模型时用逻辑回归模型作为基模型就需要对连续变量进行离散化，而离散化通常采用分箱法。而传统的数据分箱方法存在分箱数过多等现象，使得模型训练效率降低，且会影响模型输出的精准度。

15 发明内容

根据本申请公开的各种实施例，提供一种评分卡模型的建立方法、装置、计算机设备和存储介质。

一种评分卡模型的建立方法，由计算机设备执行，所述方法包括：获取多个训练样本的样本数据；所述样本数据包括多个样本变量；对每个样本变量进行分箱操作；确定每个
20 样本变量对应的分箱数，比较所述分箱数是否超过阈值；若是，计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱的卡方值；根据所述分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理，返回所述确定每个样本变量对应的分箱数的步骤；及否则，计算每个样本变量的 WOE 值，根据所述 WOE 值进行样本变量筛选，基于筛选得到的样本变量建立评分卡模型。

25 在其中一个实施例中，所述对每个样本变量进行分箱操作，包括：识别所述训练样本的关联样本，爬取所述关联样本的关联数据；所述关联数据包括多个关联变量；接收终端发送的模型配置信息，在所述模型配置信息中提取衍生因子，获取训练样本对应每个衍生因子的衍生变量；及对每个样本变量、关联变量和衍生变量进行分箱操作。

在其中一个实施例中，所述根据分箱占比、坏样本率及卡方值，对样本变量的多个分
30 箱进行合并处理，包括：根据所述坏样本率确定多个分箱的单调特性；识别不符合所述单调特性、坏样本率为预设值、分箱占比最小或者卡方值最小的分箱，分别记作待合并分箱；及将所述待合并分箱与前一相邻分箱或后一相邻分箱进行合并。

在其中一个实施例中，所述根据所述坏样本率确定多个分箱的单调特性，包括：统计坏样本率呈单调趋势的分箱数；确定分箱数最大的单调趋势；及根据所述分箱数最大的单

调趋势确定相应样本变量的单调特性。

在其中一个实施例中，所述将所述待合并分箱与前一相邻分箱或后一相邻分箱进行合并，包括： 计算所述待合并分箱与前一相邻分箱的卡方值，记作第一卡方值；计算所述待合并分箱与后一相邻分箱的卡方值，记作第二卡方值；比较所述第一卡方值是否等于所述第二卡方值；若是，将待合并分箱与分箱占比小的前一相邻分箱或后一相邻分箱合并；及否则，将待合并分箱与卡方值小的前一相邻分箱或后一相邻分箱进行合并。

在其中一个实施例中，所述基于筛选得到的样本变量建立评分卡模型，包括：在所述训练样本中筛选目标样本，提取所述目标样本的样本特征；对所述样本特征进行强化学习，得到更多的衍生样本；利用所述训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算所述评分卡模型的准确度，比较所述准确度是否达到阈值；及若否，基于所述衍生样本生成再生样本；将所述再生样本作为当前的衍生样本，返回所述利用所述训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算所述评分卡模型的准确度，比较所述准确度是否达到阈值的步骤，直至所述准确度达到阈值。

一种评分卡模型的建立装置，所述装置包括：数据分箱模块，用于获取多个训练样本的样本数据；所述样本数据包括多个样本变量；对每个样本变量进行分箱操作；分箱合并模块，用于确定每个样本变量对应的分箱数，比较所述分箱数是否超过阈值；若是，计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱的卡方值；根据所述分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理，返回所述确定每个样本变量对应的分箱数的步骤；及模型建立模块，用于当样本变量对应的分箱数小于或等于所述阈值时，计算每个样本变量的 WOE 值，根据所述 WOE 值进行样本变量筛选，基于筛选得到的样本变量建立评分卡模型。

在其中一个实施例中，所述数据分箱模块还用于识别所述训练样本的关联样本，爬取所述关联样本的关联数据；所述关联数据包括多个关联变量； 接收终端发送的模型配置信息，在模型配置信息中提取衍生因子，获取训练样本对应每个衍生因子的衍生变量；及对每个样本变量、关联变量和衍生变量进行分箱操作。

一种计算机设备，包括存储器和一个或多个处理器，存储器中存储有计算机可读指令，计算机可读指令被处理器执行时实现本申请任意一个实施例中提供的风险预测处理方法的步骤。

一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器实现本申请任意一个实施例中提供的风险预测处理方法的步骤。

本申请的一个或多个实施例的细节在下面的附图和描述中提出。本申请的其它特征和优点将从说明书、附图以及权利要求书变得明显。

附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例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申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

5 图 1 为根据一个或多个实施例中评分卡模型的建立方法的应用场景图。

图 2 为根据一个或多个实施例中评分卡模型的建立方法的流程示意图。

图 3 为根据一个或多个实施例中分箱合并的步骤的流程示意图。

图 4 为根据一个或多个实施例中评分卡模型的建立装置的结构框图。

图 5 为根据一个或多个实施例中计算机设备的框图。

10

具体实施方式

为了使本申请的技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

15 本申请提供的评分卡模型的建立方法，可以应用于如图 1 所示的应用环境中。终端 102 与服务器 104 通过网络进行通信。终端 102 可以但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备，服务器 104 可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

当用户需要对目标对象进行评分时，可以通过终端 102 向服务器 104 发送评分请求。
20 服务器 104 根据评分请求获取监控对象的监控数据，调用评分卡模型对监控数据进行处理。其中，评分卡模型可以是服务器 104 基于筛选得到的样本变量训练得到的。具体的，当需要建立评分卡模型时，用户可以通过终端 102 向服务器 104 发送模型建立请求。服务器 104 根据模型建立请求获取多个训练样本的样本数据。样本数据包括多个样本变量。服务器 104 按照无监督分箱方法或者有监督分箱方法对每个样本变量进行分箱操作。服务器
25 104 确定每个样本变量对应的分箱数，比较分箱数是否超过阈值。如果一个样本变量的分箱数超过阈值，则服务器 104 计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱的卡方值，并根据坏样本率确定多个分箱的单调特性。服务器 104 识别不符合单调特性、坏样本率为预设值、分箱占比最小或者卡方值最小的分箱，分别记作待合并分箱。服务器 104 将待合并分箱与前一相邻分箱或后一相邻分箱进行合并，合并后再次判断该样本变量的分箱数是否超过阈值。若仍超过阈值，则服务器 104 按照上述方式样本变量的多个分箱再次进行合并处理，直至该样本变量的分箱数小于或等于阈值。当样本变量的分箱数小于或等于阈值时，服务器 104 计算每个样本变量的 WOE 值，根据 WOE 值进行样本变量筛选，基于筛选得到的样本变量建立评分卡模型。上述评分卡模型建立过程，基于改进的卡方分箱方式对训练样本进行样本均衡处理，基于样本均衡得到的多个样本集训练建
30 立评分卡模型，可以提高模型精度，也可以减少分箱数，进而可以提高模型训练效率。
35

在其中一个实施例中，如图 2 所示，提供了一种评分卡模型的建立方法，以该方法应用于图 1 中的服务器为例进行说明，包括以下步骤：

步骤 202，获取多个训练样本的样本数据；样本数据包括多个样本变量。

- 5 每个训练样本为虚拟资源的一个监控对象，如已经曝出具有信用风险的企业或个人，也可以是尚未被曝出具有信用风险的企业或个人。虚拟资源可以是股票、债券等。监控对象作为风险监控对象，样本数据包括监控对象的多个维度的监控数据，如训练样本对应的样本数据包括财务变量 A1，财务变量 A2，司法变量 B1、舆情变量 C1、行业变量 D1；训练样本 2 对应的样本数据包括财务变量 A1，财务变量 A2，财务变量 A3、舆情变量 C2、
10 行业变量 D2。

步骤 204，对每个样本变量进行分箱操作。

确定每种样本变量的变量类型。变量类型包括定性变量和定量变量。

如下表 1 所示，若变量类型为定性变量，根据定性变量对应的属性值进行初始分箱。

表 1

财务变量 A1	响应（坏样本）	未响应（好样本）	合计	坏样本率
很好	4000	16000	20000	20%
较好	3000	27000	30000	10%
一般	3000	12000	15000	20%
较差	1500	8500	10000	15%
很差	1000	5000	5000	10%

- 15 如下表 2 所示，若变量类型为定量变量，若变量类型为定量变量，基于无监督分箱方法或有监督分箱方法对定量变量进行初始分箱。无监督分箱方法可以是等距分箱、等宽分箱等。有监督分箱方法可以是卡方分箱等。

表 2

财务变量 A2	响应（坏样本）	未响应（好样本）	合计	坏样本率
<100 元	2500	47500	50000	5%
[100,200]	3000	27000	30000	10%
[200,500]	3000	12000	15000	20%
[500,700]	1500	8500	10000	15%
[700,900]	2000	8000	10000	20%
≥900 元	1000	4000	5000	20%

步骤 206，确定每个样本变量对应的分箱数，比较分箱数是否超过阈值。

- 20 分箱数的阈值可以根据经验自由设定，如 5 等。需要说明的是，不同样本变量对应的分箱数阈值可以不同。

步骤 208，若是，计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱

的卡方值；根据分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理，返回确定每个样本变量对应的分箱数的步骤。

分箱占比是指变量值落入当前分箱的样本变量的数量与包含该样本变量的全部训练样本的数量的比值，例如，财务变量 A2 在[500, 700]分箱的分箱占比为 10000/120000。

- 5 坏样本率是指当前分箱内坏样本的数量与当前分箱内全部样本变量的数量的比值。例如，财务变量 A2 在[500, 700]分箱的坏样本率为 2500/50000。与相邻分箱的卡方值是非参数检验中的一个统计量，用于检验相邻分箱的数据相关性。

- 10 服务器识别是否存在坏样本率为 0 或者 ∞ 的分箱，将坏样本率为 0 或者 ∞ 的分箱标记为待合并分箱。服务器识别卡方值最小的一对相邻分箱，将卡方值最小的相邻分箱标记为待合并分箱。服务器识别分箱占比最小的一个分箱，将分箱占比最小的分箱标记为待合并分箱。服务器将待合并分箱与前一相邻分箱或后一相邻分箱进行合并。待合并处理完毕，服务器重新判断该样本变量的分箱数是否依然超过阈值。若是，继续按照上述方式对样本变量对应的多个分箱进行合并处理，直至样本变量对应的分箱数小于或等于阈值。

- 15 步骤 210，否则，计算每个样本变量的 WOE 值，根据 WOE 值进行样本变量筛选，基于筛选得到的样本变量建立评分卡模型。

基于改进的分箱方法对训练样本进行样本均衡处理，基于样本均衡得到的多个样本集训练建立评分卡模型，可以提高模型精度，也可以减少分箱数，进而可以提高模型训练效率。

- 20 本实施例中，获取多个训练样本的样本数据，并对样本数据包含的多个样本变量进行分箱操作，可以确定每个样本变量对应的分箱数；当分箱数超过阈值时，可以计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱的卡方值，并根据分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理；当分箱数小于或等于阈值时，可以计算每个样本变量的 WOE 值，根据 WOE 值进行样本变量筛选；基于筛选得到的样本变量，可以建立评分卡模型。由于基于分箱占比、坏样本率及卡方值多个分箱条件进行分箱，
25 可以限定每个样本变量分箱的数量，也可以使得稀少样本均匀分布在不同分箱，实现样本均衡，从而根据这种分箱机制进行样本变量的筛选可以提高模型训练效率和精准度。

- 30 在其中一个实施例中，对每个样本变量进行分箱操作，包括：识别训练样本的关联样本，爬取关联样本的关联数据；关联数据包括多个关联变量；接收终端发送的模型配置信息，在模型配置信息中提取衍生因子，获取训练样本对应每个衍生因子的衍生变量；
35 对每个样本变量、关联变量和衍生变量进行分箱操作。

服务器调用预设的风险传导模型。风险传导模型包括关系提取模型和传导预测模型。服务器在指定网站爬取监控对象的社交关系数据，将社交关系数据输入关系提取模型，确定训练样本对应的一个或多个关联样本，并基于确定的关联样本生成训练样本对应的知识图谱。关联样本可以是与监控对象存在投资关系、供应关系或其他关系的关联对象。知识
35 图谱包括监控对象节点和多个关联对象节点。

关系提取模型包括亲密度测算子模型。利用亲密度测算子模型计算每个关联样本与训练样本的亲密度。亲密度测算子模型可以是：

$$Q(v, w) = \frac{|N(v) \cap N(w)| + 1}{|N(v) \cup N(w)|}$$

其中， $Q(v, w)$ 为关联样本节点 w 与训练样本节点 v 之间的亲密度； $N(v)$ 表示训练样本节点 v 的邻接节点集合；训练样本节点 v 和关联样本节点 w 的共同邻接节点数为 $|N(v) \cap N(w)|$ ；训练样本节点 v 和关联样本节点 w 均不邻接的节点数为 $|N(v) \cup N(w)|$ 。在另一个实施例中，亲密度根据关联关系，可以是投资比例、出质比例、出资比例等。

服务器在指定网站爬取关联样本的关联数据。服务器将关联数据以及对应的亲密度输入传导预测模型，计算关联样本的传导风险评分，将传导风险评分标记为关联变量。在另一个实施例中，服务器将多个关联对象分别对应的传导风险评分中值最高的传导风险评分作为一个关联变量，或者将多个关联对象分别对应的传导风险评分的平均值作为一个关联变量，对此不作限制。

虚拟资源获取平台本身提供的评分卡模型只是提供一个模型框架。若用户对用于虚拟资源风险分析的评分卡模型不满意，可以通过终端向服务器发送模型配置请求，进而根据自己的行业经验对评分卡模型进行变更。具体的，服务器根据模型配置请求携带的模型标识，获取对应的评分卡模型。服务器识别评分卡模型中的可编辑元素，利用空白单元对可编辑元素进行替换，并将可编辑元素填充至空白单元，得到模型编辑页面，将模型编辑页面返回至终端。模型编辑页面允许用户在模型框架基础上根据行业经验进行自由编辑，实现模型定制化。例如，允许用户修改变量权重、更改变量值等。另外，模型编辑页面还包括“新增指标”按钮，以支持用户加入全新变量。服务器根据模型配置请求向终端返回模型编辑页面。编辑信息包括变更后的评分卡模型。

当用户增加全新变量时，在用户终端录入采用自然语言或公式等的形式录入新增变量的变量名称、变量类型和变量值。其中变量类型可以是财务类、舆情类等。例如，用户新增指标“高管是否有司法处罚”。服务器定期进行全网筛查，将用户新增的变量（记作衍生变量）添加至变量库中，以便后续该用户或其他用户再次利用。在另一个实施例中，不同用户对评分卡模型的定制逻辑可能不同，为了保护用户的定制逻辑，可以对不同用户对评分卡模型的定制化操作进行数据隔离。

若新增因子为公式，则服务器对公式进行拆分，在监控数据中获取对应的多个分项变量，按照公式逻辑对多个分项变量进行预设逻辑运算即可得到对应的变量值。若新增变量为自然语言，还允许用户针对新增变量配置对应的变量值获取逻辑，基于配置的获取逻辑自动获取变量值。例如，根据新增变量“高管是否有司法处罚”，首先在指定网站爬取有关高管司法处罚的信息，对涉世高管进行主体（是否为期望监控的高管）识别；若主体一致，对爬取到的信息进行舆情分析，得到对应的变量值。容易理解，新增变量值也可以是人为录入的。

服务器按照上述方式对样本变量、关联变量和衍生变量进行分箱操作。

本实施例中，由于除了在训练样本自身的样本数据中提取样本变量，还识别训练样本的关联样本，并在关联样本对应风险数据中提取关联变量，将关联变量纳入风险测算考虑范围，可以拓展风险预测维度，进而可以提高风险预测准确性。此外，向用户提供通用的评分卡模型，降低风险预测门槛的同时，允许用户根据自己的行业经验对评分卡模型进行变更，可以实现模型定制化，从而使得该虚拟资源平台适用于任何有无行业背景的用户。

在其中一个实施例中，如图3所示，根据分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理，即分箱合并的步骤，包括：

步骤302，根据坏样本率确定多个分箱的单调特性。

单调特性是指坏样本率呈连续递增变化趋势或连续递减变化趋势的特性。

在其中一个实施例中，根据坏样本率确定多个分箱的单调特性，包括：统计坏样本率呈单调趋势的分箱数；确定分箱数最大的单调趋势；根据分箱数最大的单调趋势确定相应样本变量的单调特性。

例如，在上述表2中，财务变量A2对应的所有分箱有6个，这6个分箱的坏样本率，第1至3分箱呈连续递增变化，第3至4分箱呈连续递减变化，第4至5分箱呈连续递增变化，第5至6分箱连续不变。通过统计，坏样本率呈连续递增变化趋势的分箱数为4，呈连续递减变化趋势的分箱数为1，从而财务变量A2的单调特性为连续递增变化趋势。

步骤304，识别不符合单调特性、坏样本率为预设值、分箱占比最小或者卡方值最小的分箱，分别记作待合并分箱。

服务器识别样本变量是否存在坏样本率为0或者 ∞ 的分箱，将坏样本率为0或者 ∞ 的分箱标记为待合并分箱。服务器识别是否存在不符合单调特性的分箱，将不符合单调特性的分箱标记为待合并分箱。服务器识别卡方值最小的一对相邻分箱，将卡方值最小的相邻分箱标记为待合并分箱。服务器识别分箱占比最小的一个分箱，将分箱占比最小的分箱标记为待合并分箱。

步骤306，将待合并分箱与前一相邻分箱或后一相邻分箱进行合并。

在其中一个实施例中，将待合并分箱与前一相邻分箱或后一相邻分箱进行合并，包括：计算待合并分箱与前一相邻分箱的卡方值，记作第一卡方值；计算待合并分箱与后一相邻分箱的卡方值，记作第二卡方值；比较第一卡方值是否等于第二卡方值；若是，将待合并分箱与分箱占比小的前一相邻分箱或后一相邻分箱合并；否则，将待合并分箱与卡方值小的前一相邻分箱或后一相邻分箱进行合并。

服务器获取坏样本率为0或者 ∞ 的分箱或者不符合单调特性的分箱，记作第一待合并分箱。若第一待合并分箱为首个分箱，将第一待合并分箱与后一相邻分箱合并。若第一待合并分箱为最后一个分箱，将第一待合并分箱与前一相邻分箱合并。

若第一待合并分箱为该样本变量中间的一个分箱，服务器获取第一待合并分箱与前一相邻分箱的卡方值（记作第一卡方值）以及与后一相邻分箱的卡方值（记作第二卡方值），并比较第一卡方值和第二卡方值。若第一卡方值大于第二卡方值，将第一待合并分箱与后

一相邻分箱进行合并。若第一卡方值小于第二卡方值，将第一待合并分箱与前一相邻分箱进行合并。若第一卡方值等于第二卡方值，将待合并分箱与分箱占比小的前一相邻分箱或后一相邻分箱进行合并。

本实施例中，基于分箱占比、坏样本率、相邻分箱的卡方值以及单调特性多个分箱条件进行分箱，可以限定每个样本变量分箱的数量，也可以使得稀少样本均匀分布在不同分箱，实现样本均衡，基于这种分箱机制进行样本变量的筛选可以提高模型训练效率和精准度。

在其中一个实施例中，基于筛选得到的样本变量建立评分卡模型，包括：在训练样本中筛选目标样本，提取目标样本的样本特征；对样本特征进行强化学习，得到更多的衍生样本；利用训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算评分卡模型的准确度，比较准确度是否达到阈值；若否，基于衍生样本生成再生样本；将再生样本作为当前的衍生样本，返回利用训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算评分卡模型的准确度，比较准确度是否达到阈值的步骤，直至准确度达到阈值。

训练样本包括好样本和坏样本。同一监控对象往往并非一直具有风险行为，有可能一段时间（记作白色时段）不具有风险行为，一段时间（记作黑色时段）具有风险行为。例如，企业 A 被曝出在 2017.08~2017.11 发生财务造假风险行为，在被监管部门要求整改后企业 A 对该段时间的财务数据进行修正，则 2017.08~2017.11 被曝出的财务数据可以作为坏样本，对应修改后的监控数据可以作为好样本。训练样本还包括灰色样本。实际上，大部分监控对象处于白色时段和黑色时段之间的灰色时段。灰色时段是指有可能存在风险行为但未被曝出的时段。好样本与坏样本的数量有限，而灰色样本则比较多。

训练样本具有对应的分类标签。服务器获取训练样本，将训练样本输入基分类器，得到模型分类结果。服务器比较模型分类结果与相应分类标签是否一致。若否，服务器将已有样本标记为目标样本。目标样本是指实际存在风险行为但未被评分卡模型识别出来的坏样本。样本特征包括坏样本的正常指标以及一种或多种异常指标。

服务器提取目标样本的样本特征。具体的，服务器获取目标样本对应的样本数据；对样本数据进行预处理，得到多个样本指标。服务器根据证监会等监管部门公布的处罚文件将其中一个或多个样本指标标记为异常指标，进而确定样本指标的指标类型。指标类型包括正常指标和异常指标。

服务器对样本特征进行强化学习，得到更多的衍生样本。具体的，服务器预存储了多个样本指标对应的强化学习规则。强化学习规则包括多种样本指标分别对应的增大幅度或减小幅度。换言之，强化学习规则包括对该样本指标值进行增大或缩小的一级幅度、二级幅度等。其中，一级幅度是指第一次根据强化学习规则对样本指标进行强化学习时对样本指标值增大的幅度或者缩小的幅度；二级幅度是指第二次根据强化学习规则对同一样本指标进行强化学习时对样本指标值增大的幅度或者缩小的幅度，如此类推。服务器根据增大幅度对异常指标进行增大处理；或根据减小幅度对异常指标进行减小处理。服务器对强化

处理后的多个异常指标进行组合，得到多种指标组合。服务器基于正常指标及强化处理后的每种指标组合生成一种衍生样本。

服务器利用训练样本和衍生样本对基础模型进行半监督训练，得到评分卡模型。具体的，基分类器可以是梯度促进模型（Gradient Boosting Decision Tree, GBDT）。容易理解，基础模型也可以是其他模型，对此不做限制。好样本与坏样本具有明确的是否造假的标注（记作分类标签），而灰色样本则没有明确的分类标签。换言之，有标注的样本数据稀少，而无标注的样本数据很多。基于有限的好样本、坏样本以及大量的灰色样本对基础模型进行半监督训练，使得训练得到的评分卡模型更加符合实际情况，从而可以提高模型监控精度。

本实施例中，由于信用风险情形太多，初始的评分卡模型难以识别出在特定情况下的信用风险，通过强化学习使得评分卡模型能够捕捉到更多情况下造成的信用风险，进而逐步识别出越来越多情形下的信用风险，提高模型精度。此外，同时将大量的灰色时段的样本数据作为样本对模型进行半监督训练，可以使模型尽可能多的学习到大部分企业在常规状态下的风险行为特征，从而可以进一步提高模型精度，可以提早发现尚处于灰色时段但有风险行为征兆的企业。

应该理解的是，虽然图 2 和图 3 的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图 2 和图 3 中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

在其中一个实施例中，如图 4 所示，提供了一种评分卡模型的建立装置，包括：数据分箱模块 402、分箱合并模块 404 和模型建立模块 406，其中：

数据分箱模块 402，用于获取多个训练样本的样本数据；样本数据包括多个样本变量；对每个样本变量进行分箱操作；

分箱合并模块 404，用于确定每个样本变量对应的分箱数，比较分箱数是否超过阈值；若是，计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱的卡方值；根据分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理，返回确定每个样本变量对应的分箱数的步骤；

模型建立模块 406，用于当样本变量对应的分箱数小于或等于阈值时，计算每个样本变量的 WOE 值，根据 WOE 值进行样本变量筛选，基于筛选得到的样本变量建立评分卡模型。

在其中一个实施例中，数据分箱模块 402 还用于识别训练样本的关联样本，爬取关联

样本的关联数据；关联数据包括多个关联变量；接收终端发送的模型配置信息，在模型配置信息中提取衍生因子，获取训练样本对应每个衍生因子的衍生变量；对每个样本变量、关联变量和衍生变量进行分箱操作。

在其中一个实施例中，分箱合并模块 404 还用于根据坏样本率确定多个分箱的单调特性；识别不符合单调特性、坏样本率为预设值、分箱占比最小或者卡方值最小的分箱，分别记作待合并分箱；将待合并分箱与前一相邻分箱或后一相邻分箱进行合并。

在其中一个实施例中，分箱合并模块 404 还用于统计坏样本率呈单调趋势的分箱数；确定分箱数最大的单调趋势；根据分箱数最大的单调趋势确定相应样本变量的单调特性。

在其中一个实施例中，分箱合并模块 404 还用于计算待合并分箱与前一相邻分箱的卡方值，记作第一卡方值；计算待合并分箱与后一相邻分箱的卡方值，记作第二卡方值；比较第一卡方值是否等于第二卡方值；若是，将待合并分箱与分箱占比小的前一相邻分箱或后一相邻分箱合并；否则，将待合并分箱与卡方值小的前一相邻分箱或后一相邻分箱进行合并。

在其中一个实施例中，模型建立模块 406 还用于在训练样本中筛选目标样本，提取目标样本的样本特征；对样本特征进行强化学习，得到更多的衍生样本；利用训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算评分卡模型的准确度，比较准确度是否达到阈值；若否，基于衍生样本生成再生样本；将再生样本作为当前的衍生样本，返回利用训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算评分卡模型的准确度，比较准确度是否达到阈值的步骤，直至准确度达到阈值。

关于评分卡模型的建立装置的具体限定可以参见上文中对于评分卡模型的建立方法的限定，在此不再赘述。上述评分卡模型的建立装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

在其中一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图 5 所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机可读指令和数据库。该内存储器为非易失性存储介质中的操作系统和计算机可读指令的运行提供环境。该计算机设备的数据库用于存储训练样本的样本数据和关联样本的关联数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机可读指令被处理器执行时以实现一种评分卡模型的建立方法。

本领域技术人员可以理解，图 5 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可

以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

一个或多个存储有计算机可读指令的非易失性存储介质，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器实现本申请任意一个实施例中提供的评分卡模型的建立方法的步骤。

- 5 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通
过计算机可读指令来指令相关的硬件来完成，的计算机可读指令可存储于一非易失性计算
机可读取存储介质中，该计算机可读指令在执行时，可包括如上述各方法的实施例的流程。
其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引
用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器（ROM）、
10 可编程 ROM（PROM）、电可编程 ROM（EPROM）、电可擦除可编程 ROM（EEPROM）
或闪存。易失性存储器可包括随机存取存储器（RAM）或者外部高速缓冲存储器。作为
说明而非局限，RAM 以多种形式可得，诸如静态 RAM（SRAM）、动态 RAM（DRAM）、
同步 DRAM（SDRAM）、双数据率 SDRAM（DDRSDRAM）、增强型 SDRAM（ESDRAM）、
同步链路（Synchlink） DRAM（SLDRAM）、存储器总线（Rambus）直接 RAM（RDRAM）、
15 直接存储器总线动态 RAM（DRDRAM）、以及存储器总线动态 RAM（RDRAM）等。

以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

- 20 以上实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此
而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。
因此，本申请专利的保护范围应以所附权利要求为准。

权利要求书

1、一种评分卡模型的建立方法，由计算机设备执行，包括：

获取多个训练样本的样本数据；所述样本数据包括多个样本变量；

对每个样本变量进行分箱操作；

确定每个样本变量对应的分箱数，比较所述分箱数是否超过阈值；

5 若是，计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱的卡方值；
根据所述分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理，返回
所述确定每个样本变量对应的分箱数的步骤；及

否则，计算每个样本变量的 WOE 值，根据所述 WOE 值进行样本变量筛选，基于筛选得到的样本变量建立评分卡模型。

10 2、根据权利要求 1 所述的方法，其特征在于，所述对每个样本变量进行分箱操作，
包括：

识别所述训练样本的关联样本，爬取所述关联样本的关联数据；所述关联数据包括多个关联变量；

15 接收终端发送的模型配置信息，在所述模型配置信息中提取衍生因子，获取训练样本
对应每个衍生因子的衍生变量；及

对每个样本变量、关联变量和衍生变量进行分箱操作。

3、根据权利要求 1 所述的方法，其特征在于，所述根据分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理，包括：

根据所述坏样本率确定多个分箱的单调特性；

20 识别不符合所述单调特性的分箱，记作待合并分箱；

识别坏样本率不等于预设值的分箱，记作待合并分箱；

识别分箱占比不等于最小值的分箱，记作待合并分箱；

识别卡方值不等于最小值的分箱，记作待合并分箱；及

将所述待合并分箱与前一相邻分箱或后一相邻分箱进行合并。

25 4、根据权利要求 3 所述的方法，其特征在于，所述根据所述坏样本率确定多个分箱
的单调特性，包括：

统计所述坏样本率呈单调趋势的分箱数；

确定所述分箱数最大的单调趋势；及

根据所述分箱数最大的单调趋势确定相应样本变量的单调特性。

30 5、根据权利要求 3 所述的方法，其特征在于，所述将所述待合并分箱与前一相邻分
箱或后一相邻分箱进行合并，包括：

计算所述待合并分箱与前一相邻分箱的卡方值，记作第一卡方值；

计算所述待合并分箱与后一相邻分箱的卡方值，记作第二卡方值；

比较所述第一卡方值是否等于所述第二卡方值；

若是，将待合并分箱与分箱占比小的前一相邻分箱或后一相邻分箱合并；
否则，将待合并分箱与卡方值小的前一相邻分箱或后一相邻分箱进行合并。

6、根据权利要求 1 所述的方法，其特征在于，所述基于筛选得到的样本变量建立评分卡模型，包括：

- 5 在所述训练样本中筛选目标样本，提取所述目标样本的样本特征；
 对所述样本特征进行强化学习，得到更多的衍生样本；
 利用所述训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算所述评分卡模型的准确度，比较所述准确度是否达到阈值；
 若否，基于所述衍生样本生成再生样本；及
- 10 将所述再生样本作为当前的衍生样本，返回所述利用所述训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算所述评分卡模型的准确度，比较所述准确度是否达到阈值的步骤，直至所述准确度达到阈值。

7、一种评分卡模型的建立装置，所述装置包括：

- 数据分箱模块，用于获取多个训练样本的样本数据；所述样本数据包括多个样本变量；
- 15 对每个样本变量进行分箱操作；
 分箱合并模块，用于确定每个样本变量对应的分箱数，比较所述分箱数是否超过阈值；
 若是，计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱的卡方值；根据所述分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理，返回所述确定每个样本变量对应的分箱数的步骤；
- 20 模型建立模块，用于当样本变量对应的分箱数小于或等于所述阈值时，计算每个样本变量的 WOE 值，根据所述 WOE 值进行样本变量筛选，基于筛选得到的样本变量建立评分卡模型。

- 8、根据权利要求 7 所述的装置，其特征在于，所述数据分箱模块还用于识别所述训练样本的关联样本，爬取所述关联样本的关联数据；所述关联数据包括多个关联变量；接收终端发送的模型配置信息，在所述模型配置信息中提取衍生因子，获取训练样本对应每个衍生因子的衍生变量；对每个样本变量、关联变量和衍生变量进行分箱操作。
- 25

9、一种计算机设备，包括存储器及一个或多个处理器，所述存储器中储存有计算机可读指令，所述计算机可读指令被所述一个或多个处理器执行时，使得所述一个或多个处理器执行以下步骤：

- 30 获取多个训练样本的样本数据；所述样本数据包括多个样本变量；
 对每个样本变量进行分箱操作；
 确定每个样本变量对应的分箱数，比较所述分箱数是否超过阈值；
 若是，计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱的卡方值；
 根据所述分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理，返回
- 35 所述确定每个样本变量对应的分箱数的步骤；及

否则，计算每个样本变量的 WOE 值，根据所述 WOE 值进行样本变量筛选，基于筛选得到的样本变量建立评分卡模型。

10、根据权利要求 9 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

- 5 识别所述训练样本的关联样本，爬取所述关联样本的关联数据；所述关联数据包括多个关联变量；

接收终端发送的模型配置信息，在所述模型配置信息中提取衍生因子，获取训练样本对应每个衍生因子的衍生变量；及

对每个样本变量、关联变量和衍生变量进行分箱操作。

- 10 11、根据权利要求 9 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

根据所述坏样本率确定多个分箱的单调特性；

识别不符合所述单调特性的分箱，记作待合并分箱；

识别坏样本率不等于预设值的分箱，记作待合并分箱；

- 15 识别分箱占比不等于最小值的分箱，记作待合并分箱；

识别卡方值不等于最小值的分箱，记作待合并分箱；及

将所述待合并分箱与前一相邻分箱或后一相邻分箱进行合并。

12、根据权利要求 11 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

- 20 统计坏样本率呈单调趋势的分箱数；

确定分箱数最大的单调趋势；及

根据所述分箱数最大的单调趋势确定相应样本变量的单调特性。

13、根据权利要求 11 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

- 25 计算所述待合并分箱与前一相邻分箱的卡方值，记作第一卡方值；

计算所述待合并分箱与后一相邻分箱的卡方值，记作第二卡方值；

比较所述第一卡方值是否等于所述第二卡方值；

若是，将待合并分箱与分箱占比小的前一相邻分箱或后一相邻分箱合并；

否则，将待合并分箱与卡方值小的前一相邻分箱或后一相邻分箱进行合并。

- 30 14、根据权利要求 9 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

在所述训练样本中筛选目标样本，提取所述目标样本的样本特征；

对所述样本特征进行强化学习，得到更多的衍生样本；

利用所述训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算所述评分

- 35 卡模型的准确度，比较所述准确度是否达到阈值；

若否，基于所述衍生样本生成再生样本；及

将所述再生样本作为当前的衍生样本，返回所述利用所述训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算所述评分卡模型的准确度，比较所述准确度是否达到阈值的步骤，直至所述准确度达到阈值。

5 15、一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行以下步骤：

获取多个训练样本的样本数据；所述样本数据包括多个样本变量；

对每个样本变量进行分箱操作；

确定每个样本变量对应的分箱数，比较所述分箱数是否超过阈值；

10 若是，计算样本变量对应每一分箱的分箱占比、坏样本率以及与相邻分箱的卡方值；
根据所述分箱占比、坏样本率及卡方值，对样本变量的多个分箱进行合并处理，返回所述确定每个样本变量对应的分箱数的步骤；及

否则，计算每个样本变量的 WOE 值，根据所述 WOE 值进行样本变量筛选，基于筛选得到的样本变量建立评分卡模型。

15 16、根据权利要求 15 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

识别所述训练样本的关联样本，爬取所述关联样本的关联数据；所述关联数据包括多个关联变量；

20 接收终端发送的模型配置信息，在所述模型配置信息中提取衍生因子，获取训练样本对应每个衍生因子的衍生变量；及

对每个样本变量、关联变量和衍生变量进行分箱操作。

17、根据权利要求 15 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

根据所述坏样本率确定多个分箱的单调特性；

25 识别不符合所述单调特性的分箱，记作待合并分箱；

识别坏样本率不等于预设值的分箱，记作待合并分箱；

识别分箱占比不等于最小值的分箱，记作待合并分箱；

识别卡方值不等于最小值的分箱，记作待合并分箱；及

将所述待合并分箱与前一相邻分箱或后一相邻分箱进行合并。

30 18、根据权利要求 17 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

统计坏样本率呈单调趋势的分箱数；

确定分箱数最大的单调趋势；及

根据所述分箱数最大的单调趋势确定相应样本变量的单调特性。

35 19、根据权利要求 17 所述的存储介质，其特征在于，所述计算机可读指令被所述处

理器执行时还执行以下步骤：

计算所述待合并分箱与前一相邻分箱的卡方值，记作第一卡方值；

计算所述待合并分箱与后一相邻分箱的卡方值，记作第二卡方值；

比较所述第一卡方值是否等于所述第二卡方值；

- 5 若是，将待合并分箱与分箱占比小的前一相邻分箱或后一相邻分箱合并；
否则，将待合并分箱与卡方值小的前一相邻分箱或后一相邻分箱进行合并。

20、根据权利要求 15 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

在所述训练样本中筛选目标样本，提取所述目标样本的样本特征；

- 10 对所述样本特征进行强化学习，得到更多的衍生样本；

利用所述训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算所述评分卡模型的准确度，比较所述准确度是否达到阈值；

若否，基于所述衍生样本生成再生样本；及

- 15 将所述再生样本作为当前的衍生样本，返回所述利用所述训练样本和衍生样本对基础模型进行训练，得到评分卡模型，计算所述评分卡模型的准确度，比较所述准确度是否达到阈值的步骤，直至所述准确度达到阈值。

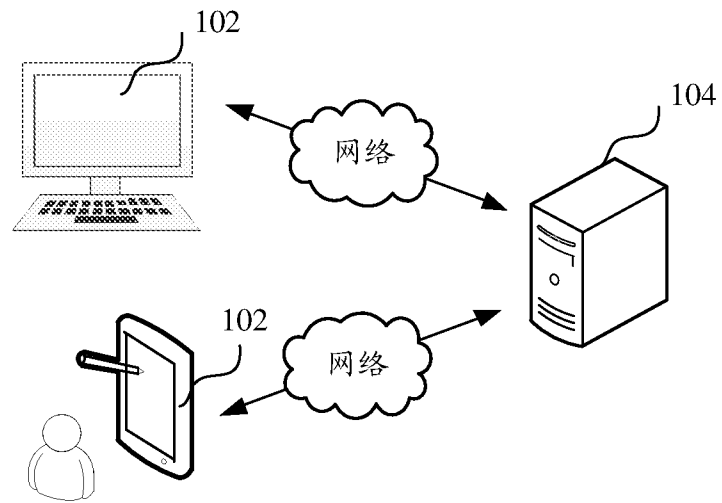


图 1

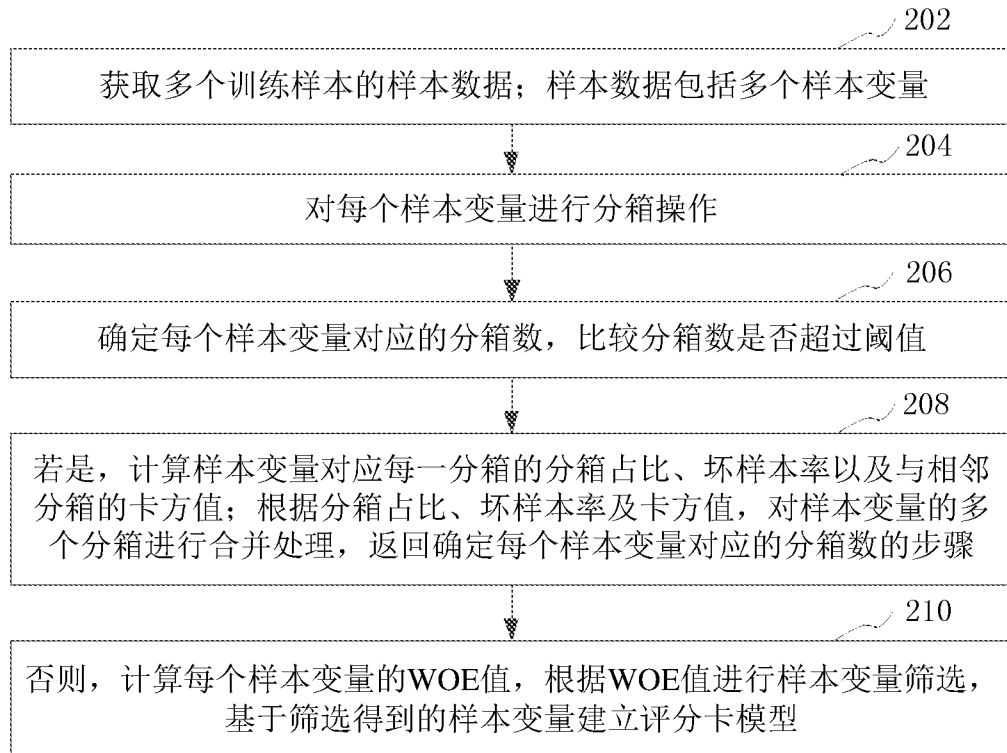


图 2

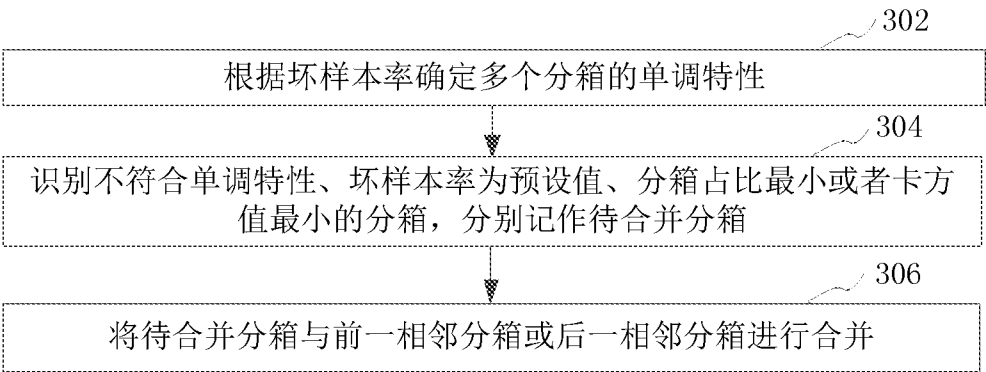


图 3

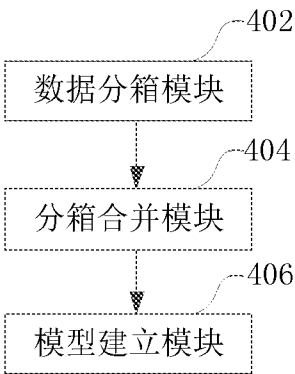


图 4

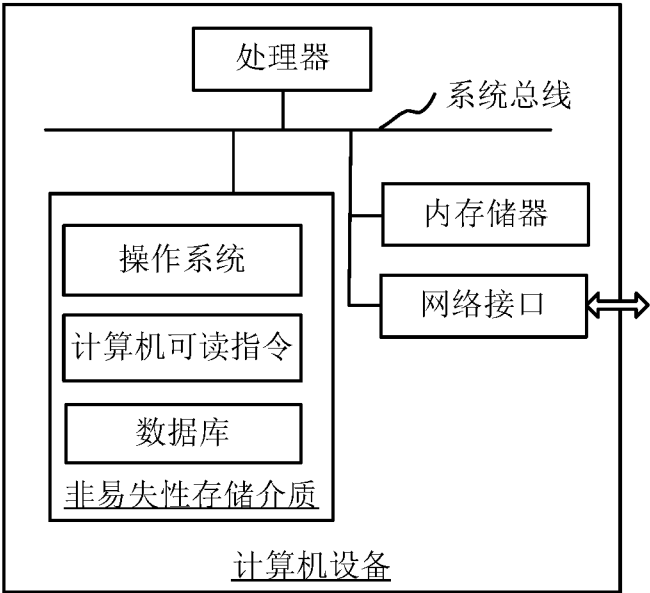


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/103489

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/50(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI, CNPAT, WPI, EPODOC: 分箱, 装箱, 分格, 数目, 数量, 多, 预设值, 一定值, 阈值, 门限, 样本, 变量, 占比, 率, 卡方值, 合并, WOE, 模型, binning, number, multiple, threshold, sample, variable, ratio, chi-square, merge, model

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109598095 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 09 April 2019 (2019-04-09) claims 1-10	1-20
A	CN 108959187 A (PING AN LIFE INSURANCE COMPANY OF CHINA) 07 December 2018 (2018-12-07) description, paragraphs [0031]-[0070]	1-20
A	WO 2015163720 A1 (INDUSTRY ACADEMIA COOPERATION FOUNDATION OF SEJONG UNIVERSITY et al.) 29 October 2015 (2015-10-29) entire document	1-20
A	CN 108876076 A (CHINA MOBILE GROUP GUANGDONG COMPANY LIMITED et al.) 23 November 2018 (2018-11-23) entire document	1-20
A	CN 108334954 A (PING AN LIFE INSURANCE COMPANY OF CHINA) 27 July 2018 (2018-07-27) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

23 October 2019

Date of mailing of the international search report

25 November 2019

Name and mailing address of the ISA/CN

China National Intellectual Property Administration
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/103489

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)		Publication date (day/month/year)
CN	109598095	A	09 April 2019	None		
CN	108959187	A	07 December 2018	None		
WO	2015163720	A1	29 October 2015	KR	101560283	B1 15 October 2015
CN	108876076	A	23 November 2018	None		
CN	108334954	A	27 July 2018	None		

国际检索报告

国际申请号

PCT/CN2019/103489

A. 主题的分类 G06F 17/50 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																				
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNKI, CNPAT, WPI, EPDOC: 分箱, 装箱, 分格, 数目, 数量, 多, 预设值, 一定值, 阈值, 门限, 样本, 变量, 占比, 率, 卡方值, 合并, WOE, 模型, binning, number, multiple, threshold, sample, variable, ratio, chi-square, merge, model																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 109598095 A (平安科技深圳有限公司) 2019年 4月 9日 (2019 - 04 - 09) 权利要求1-10</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 108959187 A (中国平安人寿保险股份有限公司) 2018年 12月 7日 (2018 - 12 - 07) 说明书第[0031]-[0070]段</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>WO 2015163720 A1 (INDUSTRY ACADEMIA COOPERATION FOUNDATION OF SEJONG UNIVERSITY 等) 2015年 10月 29日 (2015 - 10 - 29) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 108876076 A (中国移动通信集团广东有限公司等) 2018年 11月 23日 (2018 - 11 - 23) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 108334954 A (中国平安人寿保险股份有限公司) 2018年 7月 27日 (2018 - 07 - 27) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 109598095 A (平安科技深圳有限公司) 2019年 4月 9日 (2019 - 04 - 09) 权利要求1-10	1-20	A	CN 108959187 A (中国平安人寿保险股份有限公司) 2018年 12月 7日 (2018 - 12 - 07) 说明书第[0031]-[0070]段	1-20	A	WO 2015163720 A1 (INDUSTRY ACADEMIA COOPERATION FOUNDATION OF SEJONG UNIVERSITY 等) 2015年 10月 29日 (2015 - 10 - 29) 全文	1-20	A	CN 108876076 A (中国移动通信集团广东有限公司等) 2018年 11月 23日 (2018 - 11 - 23) 全文	1-20	A	CN 108334954 A (中国平安人寿保险股份有限公司) 2018年 7月 27日 (2018 - 07 - 27) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
PX	CN 109598095 A (平安科技深圳有限公司) 2019年 4月 9日 (2019 - 04 - 09) 权利要求1-10	1-20																		
A	CN 108959187 A (中国平安人寿保险股份有限公司) 2018年 12月 7日 (2018 - 12 - 07) 说明书第[0031]-[0070]段	1-20																		
A	WO 2015163720 A1 (INDUSTRY ACADEMIA COOPERATION FOUNDATION OF SEJONG UNIVERSITY 等) 2015年 10月 29日 (2015 - 10 - 29) 全文	1-20																		
A	CN 108876076 A (中国移动通信集团广东有限公司等) 2018年 11月 23日 (2018 - 11 - 23) 全文	1-20																		
A	CN 108334954 A (中国平安人寿保险股份有限公司) 2018年 7月 27日 (2018 - 07 - 27) 全文	1-20																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																				
国际检索实际完成的日期 2019年 10月 23日		国际检索报告邮寄日期 2019年 11月 25日																		
ISA/CN的名称和邮寄地址 中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		授权官员 王继梅 电话号码 86-(10)-53961802																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/103489

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	109598095	A	2019年 4月 9日	无	
CN	108959187	A	2018年 12月 7日	无	
WO	2015163720	A1	2015年 10月 29日	KR 101560283 B1	2015年 10月 15日
CN	108876076	A	2018年 11月 23日	无	
CN	108334954	A	2018年 7月 27日	无	