(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 109815267 A
(43)申请公布日 2019.05.28

(21)申请号 201811567505.0

(22)申请日 2018.12.21

(71)申请人 天翼征信有限公司
地址 200080 上海市虹口区广纪路173号1001-1007室107C

(72)发明人 凤杰 陈建立 李留洋 侯金鑫 徐明成

(74)专利代理机构 上海光华专利事务所(普通合伙) 31219
代理人 徐秋平

(51)Int.Cl.
G06F 16/2458(2019.01)
G06F 16/2457(2019.01)

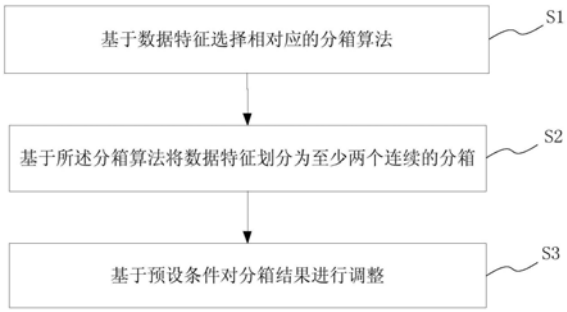
权利要求书1页 说明书5页 附图3页

(54)发明名称

数据建模中特征的分箱优化方法及系统、存储介质及终端

(57)摘要

本发明提供一种数据建模中特征的分箱优化方法及系统、存储介质及终端，包括以下步骤：基于数据特征选择相对应的分箱算法；基于所述分箱算法将数据特征划分为至少两个连续的分箱；基于预设条件对分箱结果进行调整。本发明的数据建模中特征的分箱优化方法及系统、存储介质及终端能够根据具体的业务场景，针对不同数据特征选择不同的分箱方法，再根据预选条件进行分箱调整，从而可以快速切换分箱方法，且分箱结果稳定、具有业务可解释性。



CN 109815267 A

1.一种数据建模中特征的分箱优化方法,其特征在于:包括以下步骤:

基于数据特征选择相对应的分箱算法;

基于所述分箱算法将数据特征划分为至少两个连续的分箱;

基于预设条件对分箱结果进行调整。

2.根据权利要求1所述的数据建模中特征的分箱优化方法,其特征在于:所述分箱算法包括等宽分箱、等深分箱、决策树分箱、卡方分箱中的一种或多种组合。

3.根据权利要求1所述的数据建模中特征的分箱优化方法,其特征在于:所述预设条件包括最大分箱个数、箱内样本数阈值、箱内正负样例占比、各分箱WOE值满足单调性中的一种或多种组合。

4.根据权利要求1所述的数据建模中特征的分箱优化方法,其特征在于:所述数据特征包括用户征信特征。

5.一种数据建模中特征的分箱优化系统,其特征在于:包括选择模块、分箱模块和调整模块;所述选择模块用于基于数据特征选择相对应的分箱算法;

所述分箱模块用于基于所述分箱算法将数据特征划分为至少两个连续的分箱;

所述调整模块用于基于预设条件对分箱结果进行调整。

6.根据权利要求5所述的数据建模中特征的分箱优化系统,其特征在于:所述分箱算法包括等宽分箱、等深分箱、决策树分箱、卡方分箱中的一种或多种组合。

7.根据权利要求5所述的数据建模中特征的分箱优化系统,其特征在于:所述预设条件包括最大分箱个数、箱内样本数阈值、箱内正负样例占比、各分箱WOE值满足单调性中的一种或多种组合。

8.根据权利要求5所述的数据建模中特征的分箱优化系统,其特征在于:所述数据特征包括用户征信特征。

9.一种存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1至4中任一所述的数据建模中特征的分箱优化方法。

10.一种终端,其特征在于,包括:处理器及存储器;

所述存储器用于存储计算机程序;

所述处理器用于执行所述存储器存储的计算机程序,以使所述终端执行权利要求1至4中任一所述的数据建模中特征的分箱优化方法。

数据建模中特征的分箱优化方法及系统、存储介质及终端

技术领域

[0001] 本发明涉及数据挖掘建模的技术领域，特别是涉及一种数据建模中特征的分箱优化方法及系统、存储介质及终端。

背景技术

[0002] 金融、互联网行业的高速发展，离不开大数据、人工智能技术作为支撑。为了快速、高效地实现数据变现、防控风险，数据挖掘建模能力愈发受到重视。

[0003] 现有技术中，一站式数据挖掘建模流程通常包含业务理解、特征预处理、特征工程、模型训练、模型部署上线和模型迭代等步骤。其中特征工程在整套流程中承前启后，一直是数据挖掘建模的重点及难点。为了高效地进行特征离散化、特征衍生、特征重要性评估等工作，特征分箱技术又在特征工程中发挥着举足轻重的作用。

[0004] 特征分箱技术包含以下两大类：

[0005] (1)无监督分箱

[0006] 常用的无监督分箱包括等宽分箱、等深分箱等。等宽分箱是将特征变量的取值范围分为若干个等宽的区间，每个区间为一个分箱。等深分箱是将特征变量的取值由小到大排序，根据分位数和分箱个数确定分箱切割点，两两切割点组成的区间即为一个分箱。

[0007] (2)有监督分箱。

[0008] 常用的有监督分箱包括决策树分箱、卡方分箱等。决策树分箱是通过单一特征训练决策树模型，根据分箱个数控制树的深度，遍历特征的所有取值，通过最小化所有叶子节点的总熵值获取特征分割点，所有分割点由小到大排序，两两切割点组成的区间即为一个分箱。卡方分箱是在初始分箱(如等宽分箱)的基础上，将具有最小卡方值的相邻区间合并，直到满足给定的停止条件，如最小卡方阈值或者最大分箱个数。

[0009] 然而，现有技术中的分箱技术一般基于业务及建模人员人工判断，存在方法单一、鲁棒性差、受人工主观因素影响较大、变量的分箱结果不具有业务可解释性等问题，从而在一定程度上影响到了模型整体效果。

发明内容

[0010] 鉴于以上所述现有技术的缺点，本发明的目的在于提供一种数据建模中特征的分箱优化方法及系统、存储介质及终端，能够根据具体的业务场景，针对不同数据特征选择不同的分箱方法，再根据预选条件进行分箱调整，从而可以快速切换分箱方法，且分箱结果稳定、具有业务可解释性。

[0011] 为实现上述目的及其他相关目的，本发明提供一种数据建模中特征的分箱优化方法，包括以下步骤：基于数据特征选择相对应的分箱算法；基于所述分箱算法将数据特征划分为至少两个连续的分箱；基于预设条件对分箱结果进行调整。

[0012] 于本发明一实施例中，所述分箱算法包括等宽分箱、等深分箱、决策树分箱、卡方分箱中的一种或多种组合。

[0013] 于本发明一实施例中，所述预设条件包括最大分箱个数、箱内样本数阈值、箱内正负样例占比、各分箱WOE值满足单调性中的一种或多种组合。

[0014] 于本发明一实施例中，所述数据特征包括用户征信特征。

[0015] 对应地，本发明提供一种数据建模中特征的分箱优化系统，包括选择模块、分箱模块和调整模块；

[0016] 所述选择模块用于基于数据特征选择相对应的分箱算法；

[0017] 所述分箱模块用于基于所述分箱算法将数据特征划分为至少两个连续的分箱；

[0018] 所述调整模块用于基于预设条件对分箱结果进行调整。

[0019] 于本发明一实施例中，所述分箱算法包括等宽分箱、等深分箱、决策树分箱、卡方分箱中的一种或多种组合。

[0020] 于本发明一实施例中，所述预设条件包括最大分箱个数、箱内样本数阈值、箱内正负样例占比、各分箱WOE值满足单调性中的一种或多种组合。

[0021] 于本发明一实施例中，所述数据特征包括用户征信特征。

[0022] 本发明提供一种存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现上述的数据建模中特征的分箱优化方法。

[0023] 最后，本发明提供一种终端，包括：处理器及存储器；

[0024] 所述存储器用于存储计算机程序；

[0025] 所述处理器用于执行所述存储器存储的计算机程序，以使所述终端执行上述的数据建模中特征的分箱优化方法。

[0026] 如上所述，本发明的数据建模中特征的分箱优化方法及系统、存储介质及终端，具有以下有益效果：

[0027] (1)针对不同数据特征可采用不同的分箱方法，并能够根据预选条件进行分箱调整；

[0028] (2)支持多类参数设置，鲁棒性更强；

[0029] (3)支持各分箱证据权重(Weight of Evidence，WOE)单调性检测，满足特征分箱结果的可解释性。

附图说明

[0030] 图1显示为本发明的数据建模中特征的分箱优化方法于一实施例中的流程图；

[0031] 图2显示为于一实施例中未调整单调性时用户近12个月非逾期还款订单数在总还款订单数的占比随等深分箱WOE走势图；

[0032] 图3显示为于一实施例中调整单调性时用户近12个月非逾期还款订单数在总还款订单数的占比随等深分箱WOE走势图；

[0033] 图4显示为于一实施例中未调整单调性时用户近12个月非逾期还款订单数在总还款订单数的占比随卡分分箱WOE走势图；

[0034] 图5显示为于一实施例中调整单调性时用户近12个月非逾期还款订单数在总还款订单数的占比随卡分分箱WOE走势图；

[0035] 图6显示为本发明的数据建模中特征的分箱优化系统于一实施例中的结构示意图；

[0036] 图7显示为本发明的终端于一实施例中的结构示意图。

[0037] 元件标号说明

[0038] 61 选择模块

[0039] 62 分箱模块

[0040] 63 调整模块

[0041] 71 处理器

[0042] 72 存储器

具体实施方式

[0043] 以下通过特定的具体实例说明本发明的实施方式，本领域技术人员可由本说明书所揭露的内容轻易地了解本发明的其他优点与功效。本发明还可以通过另外不同的具体实施方式加以实施或应用，本说明书中的各项细节也可以基于不同观点与应用，在没有背离本发明的精神下进行各种修饰或改变。

[0044] 需要说明的是，本实施例中所提供的图示仅以示意方式说明本发明的基本构想，遂图式中仅显示与本发明中有关的组件而非按照实际实施时的组件数目、形状及尺寸绘制，其实际实施时各组件的型态、数量及比例可为一种随意的改变，且其组件布局型态也可能更为复杂。

[0045] 本发明的数据建模中特征的分箱优化方法及系统、存储介质及终端能够根据具体的业务场景，针对不同数据特征选择不同的分箱方法，以选择适配的分箱算法；再根据预选条件进行分箱调整，以达到预期的分箱效果，从而能够快速切换分箱方法，且分箱结果稳定、具有业务可解释性。

[0046] 如图1所示，于一实施例中，本发明的数据建模中特征的分箱优化方法包括以下步骤：

[0047] 步骤S1、基于数据特征选择相对应的分箱算法。

[0048] 具体地，对于建模数据的数据特征进行分析，根据所述数据特征选择与之相适配的分箱算法。优选地，所述数据特征可以是用户征信特征，从而能够基于用户征信特征进行分箱以及数据建模。

[0049] 于本发明一实施例中，所述分箱算法包括等宽分箱、等深分箱、决策树分箱、卡方分箱中的一种或多种组合。

[0050] 步骤S2、基于所述分箱算法将数据特征划分为至少两个连续的分箱。

[0051] 具体地，确定分箱算法之后，基于所述分箱算法对所述数据特征进行分箱，从而获取至少两个连续的分箱。其中，分箱的个数须根据数据特征而确定。分箱个数较少，无法基于分箱结构进行数据建模；分箱个数较多，则数据处理复杂度过高。

[0052] 步骤S3、基于预设条件对分箱结果进行调整。

[0053] 具体地，分箱完毕后，基于用户需求还可以对分箱结果进行进一步调整，以满足个性化需求。

[0054] 于本发明一实施例中，所述预设条件包括最大分箱个数、箱内样本数阈值、箱内正负样例占比、各分箱WOE值满足单调性中的一种或多种组合。

[0055] 下面通过具体实施例来进一步阐述本发明的数据建模中特征的分箱优化方法。

[0056] 在该实施例中，需要对用户征信评价进行数据建模。其中，以连续特征“用户近12个月非逾期还款订单数在总还款订单数的占比”为例，从业务场景及特征变量分箱结果的可解释性出发，该特征应当与用户是好用户的可能性(概率)负相关。即用户近12个月非逾期还款订单数在总还款订单数的占比越低，用户是好用户的可能性越高；用户近12个月非逾期还款订单数在总还款订单数的占比越高，用户是好用户的可能性越低。因此，如图2和图4所示，首先可基于等深分箱、卡方分箱等分箱方法对该连续特征进行分箱；如图3和图5所示，然后基于各分箱WOE值满足单调性的预设条件对分箱结果进行调整，从而得到各分箱WOE值满足单调性的分箱结果。

[0057] 如图6所示，于一实施例中，本发明提的数据建模中特征的分箱优化系统包括选择模块61、分箱模块62和调整模块63。

[0058] 选择模块61用于基于数据特征选择相对应的分箱算法。

[0059] 具体地，对于建模数据的数据特征进行分析，根据所述数据特征选择与之相适配的分箱算法。优选地，所述数据特征可以是用户征信特征，从而能够基于用户征信特征进行分箱以及数据建模。

[0060] 于本发明一实施例中，所述分箱算法包括等宽分箱、等深分箱、决策树分箱、卡方分箱中的一种或多种组合。

[0061] 分箱模块62与选择模块61相连，用于基于所述分箱算法将数据特征划分为至少两个连续的分箱。

[0062] 具体地，确定分箱算法之后，基于所述分箱算法对所述数据特征进行分箱，从而获取至少两个连续的分箱。其中，分箱的个数须根据数据特征而确定。分箱个数较少，无法基于分箱结构进行数据建模；分箱个数较多，则数据处理复杂度过高。

[0063] 调整模块63与分箱模块62相连，用于基于预设条件对分箱结果进行调整。

[0064] 具体地，分箱完毕后，基于用户需求还可以对分箱结果进行进一步调整，以满足个性化需求。

[0065] 于本发明一实施例中，所述预设条件包括最大分箱个数、箱内样本数阈值、箱内正负样例占比、各分箱WOE值满足单调性中的一种或多种组合。

[0066] 需要说明的是，应理解以上装置的各个模块的划分仅仅是一种逻辑功能的划分，实际实现时可以全部或部分集成到一个物理实体上，也可以物理上分开。且这些模块可以全部以软件通过处理元件调用的形式实现；也可以全部以硬件的形式实现；还可以部分模块通过处理元件调用软件的形式实现，部分模块通过硬件的形式实现。例如，x模块可以为单独设立的处理元件，也可以集成在上述装置的某一个芯片中实现，此外，也可以以程序代码的形式存储于上述装置的存储器中，由上述装置的某一个处理元件调用并执行以上x模块的功能。其它模块的实现与之类似。此外这些模块全部或部分可以集成在一起，也可以独立实现。这里所述的处理元件可以是一种集成电路，具有信号的处理能力。在实现过程中，上述方法的各步骤或以上各个模块可以通过处理器元件中的硬件的集成逻辑电路或者软件形式的指令完成。

[0067] 例如，以上这些模块可以是被配置成实施以上方法的一个或多个集成电路，例如：一个或多个特定集成电路(Application Specific Integrated Circuit，简称ASIC)，或，一个或多个微处理器(Digital Singnal Processor，简称DSP)，或，一个或者多个现场可编

程门阵列(Field Programmable Gate Array,简称FPGA)等。再如,当以上某个模块通过处理元件调度程序代码的形式实现时,该处理元件可以是通用处理器,例如中央处理器(Central Processing Unit,简称CPU)或其它可以调用程序代码的处理器。再如,这些模块可以集成在一起,以片上系统(system-on-a-chip,简称SOC)的形式实现。

[0068] 本发明的存储介质上存储有计算机程序,所述计算机程序被处理器执行时实现上述的数据建模中特征的分箱优化方法。

[0069] 如图7所示,于一实施例中,本发明的终端包括:处理器71及存储器72。

[0070] 所述存储器72用于存储计算机程序。

[0071] 所述存储器72包括:ROM、RAM、磁碟、U盘、存储卡或者光盘等各种可以存储程序代码的介质。

[0072] 所述处理器71与所述存储器72相连,用于执行所述存储器72存储的计算机程序,以使所述终端执行上述的数据建模中特征的分箱优化方法。

[0073] 优选地,所述处理器71可以是通用处理器,包括中央处理器(Central Processing Unit,简称CPU)、网络处理器(Network Processor,简称NP)等;还可以是数字信号处理器(Digital Signal Processor,简称DSP)、专用集成电路(Application Specific Integrated Circuit,简称ASIC)、现场可编程门阵列(Field Programmable Gate Array,简称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。

[0074] 综上所述,本发明的数据建模中特征的分箱优化方法及系统、存储介质及终端针对不同数据特征可采用不同的分箱方法,并能够根据预选条件进行分箱调整;支持多类参数设置,鲁棒性更强;支持各分箱WOE单调性检测,满足特征分箱结果的可解释性。所以,本发明有效克服了现有技术中的种种缺点而具高度产业利用价值。

[0075] 上述实施例仅例示性说明本发明的原理及其功效,而非用于限制本发明。任何熟悉此技术的人士皆可在不违背本发明的精神及范畴下,对上述实施例进行修饰或改变。因此,举凡所属技术领域中具有通常知识者在未脱离本发明所揭示的精神与技术思想下所完成的一切等效修饰或改变,仍应由本发明的权利要求所涵盖。

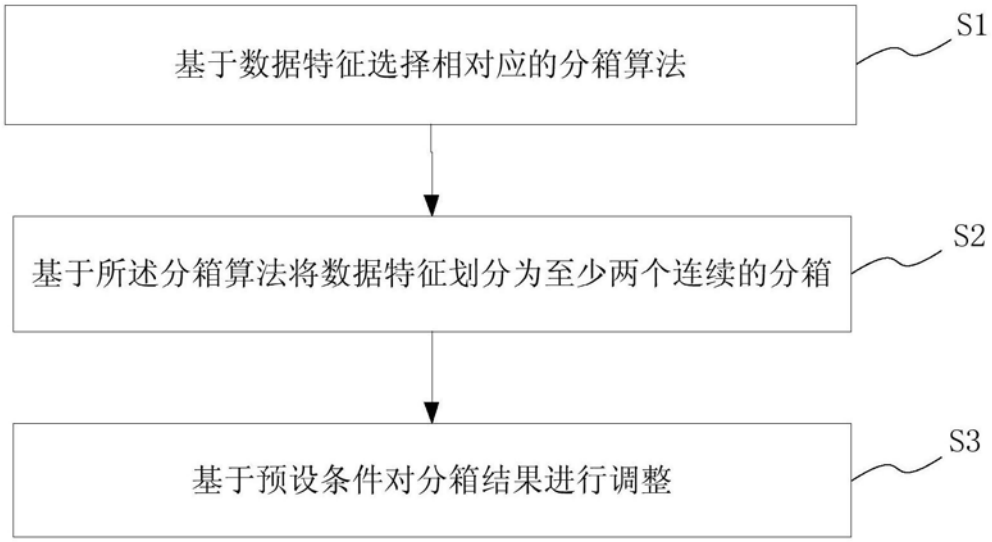

基于数据特征选择相对应的分箱算法
S1
基于所述分箱算法将数据特征划分为至少两个连续的分箱
S2
基于预设条件对分箱结果进行调整
S3

图1

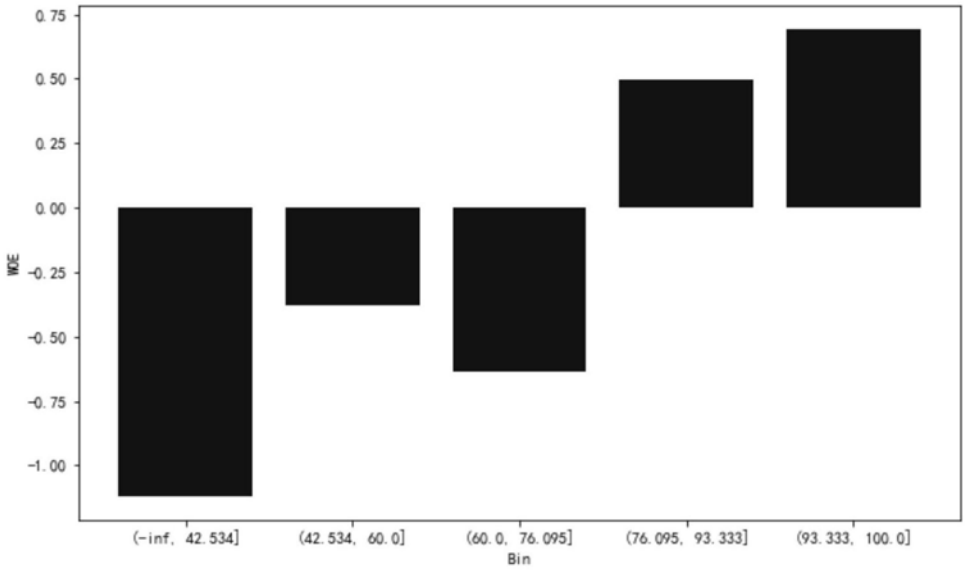

0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
WOE
(-inf, 42.534]
(42.534, 60.0]
(60.0, 76.095]
(76.095, 93.333]
(93.333, 100.0]
Bin

图2

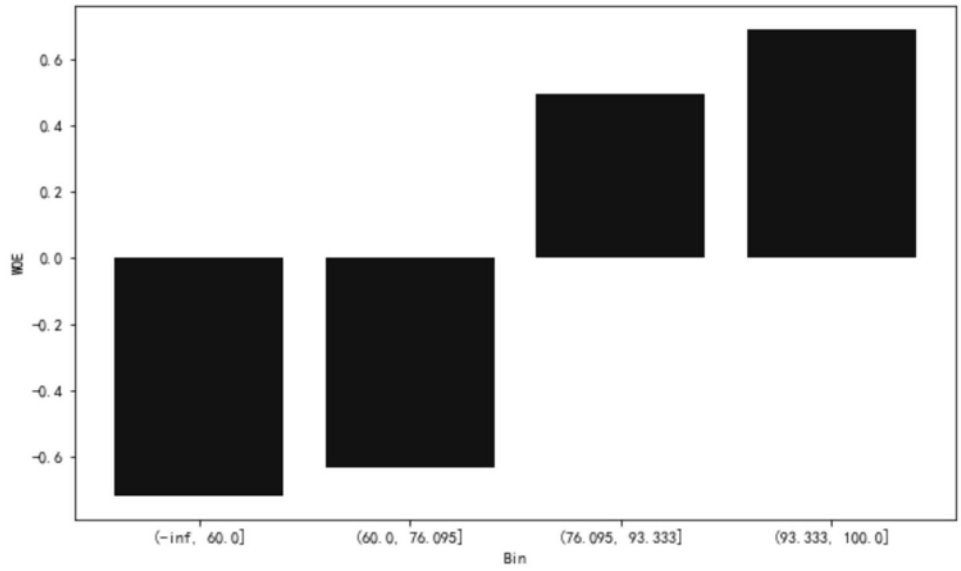

0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
WOE
(-inf, 60.0]
(60.0, 76.095]
(76.095, 93.333]
(93.333, 100.0]
Bin

图3

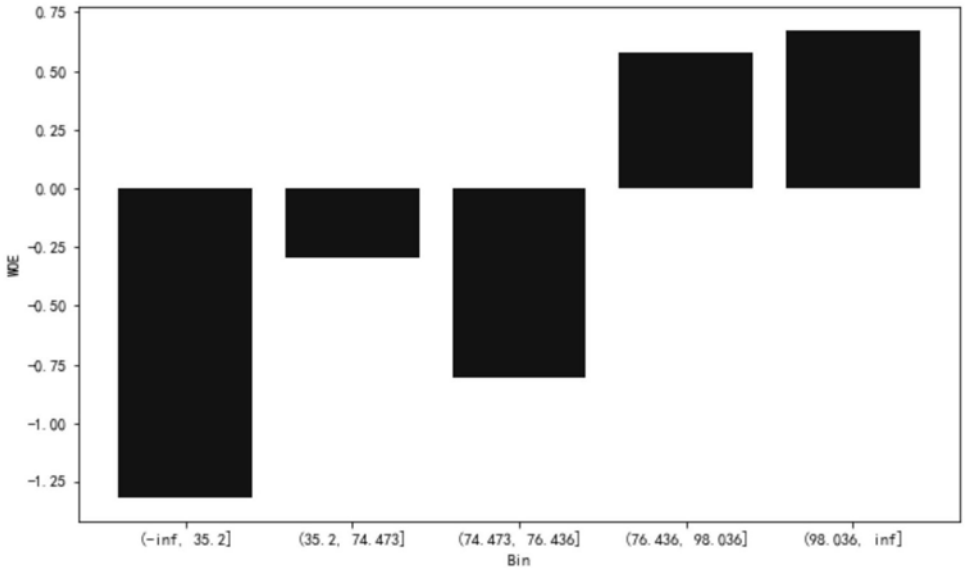

0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
-1.25
WOE
(-inf, 35.2]
(35.2, 74.473]
(74.473, 76.436]
(76.436, 98.036]
(98.036, inf]
Bin

图4

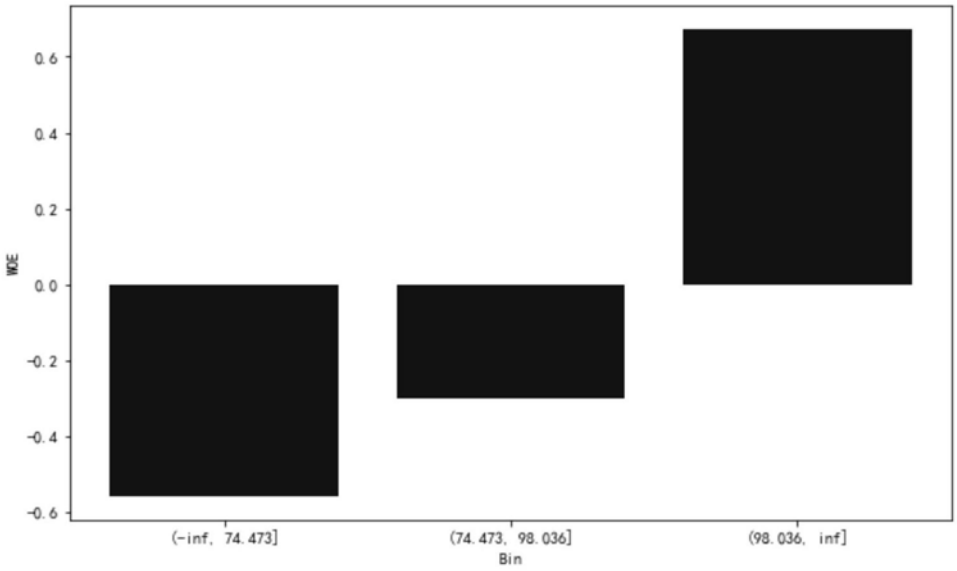

0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
WOE
(-inf, 74.473]
(74.473, 98.036]
(98.036, inf]
Bin

图5

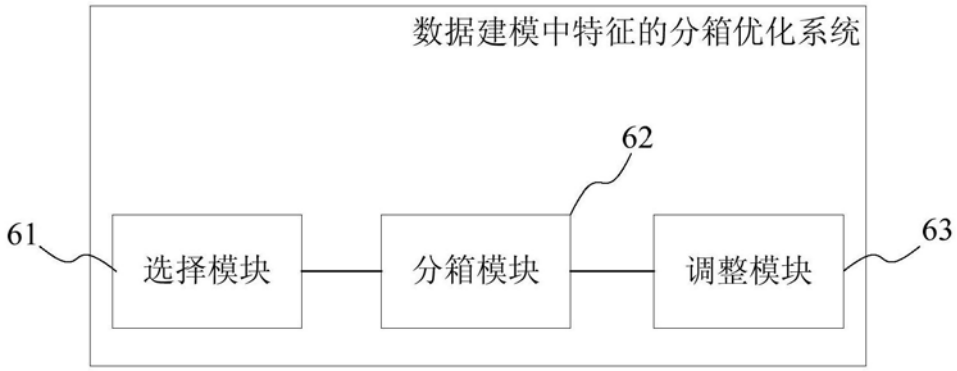

数据建模中特征的分箱优化系统
62
61
选择模块
分箱模块
调整模块
63

图6

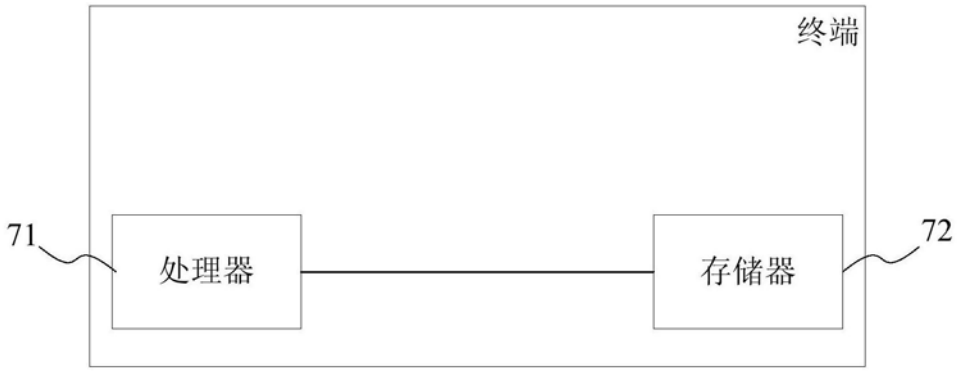

终端
71
处理器
存储器
72

图7