(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 4 月 16 日 (16.04.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/073714 A1

(51) 国际专利分类号:
G06K 9/62 (2006.01)

(21) 国际申请号: PCT/CN2019/097090

(22) 国际申请日: 2019 年 7 月 22 日 (22.07.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811191003.2 2018年10月12日 (12.10.2018) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (**ALIBABA GROUP HOLDING LIMITED**) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人: 张雅淋 (**ZHANG, Yalin**); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 周俊 (**ZHOU, Jun**); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (**BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION**); 中国北京市海淀区上地三街 9 号嘉华大厦 B 座 409, Beijing 100085 (CN)。

(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,

(54) **Title:** TRAINING SAMPLE OBTAINING METHOD, ACCOUNT PREDICTION METHOD, AND CORRESPONDING DEVICES

(54) 发明名称: 训练样本获取方法, 账户预测方法及对应装置

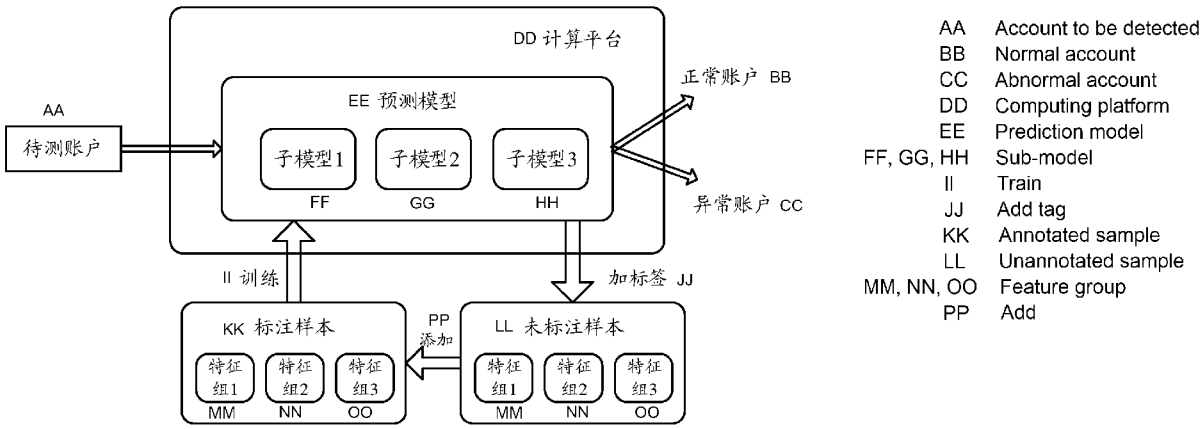

图 1

(57) **Abstract:** Embodiments of the description provide a training sample obtaining method, an account prediction method, and corresponding devices. The training sample obtaining method comprises: first, obtaining an annotated sample set, a sample feature being divided into n groups of features so as to form n sub-annotated sample sets, and obtaining n sub-models through training by using the n sub-annotated sample sets; and then obtaining an unannotated first sample, which comprises corresponding n groups of features; respectively inputting (n-1) groups of features of the first sample except for an i-th groups of features into (n-1) sub-models in the n sub-models except for an i-th sub-model to obtain (n-1) scores. Next, obtaining a first comprehensive score on the basis of the (n-1) scores, and when the first comprehensive score meets a predetermined condition, adding a first tag to the i-th groups of features of the first sample so as to form a first sub-annotated sample. In this way, the first sub-annotated sample can be added to an i-th sub-annotated sample set to update the annotated sample set.

WO 2020/073714 A1

[见续页]

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：本说明书实施例提供一种训练样本获取方法，账户预测方法以及对应装置。训练样本获取方法包括，首先获取标注样本集，其中的样本特征被划分为n组特征，由此形成n个子标注样本集。利用这n个子标注样本集训练得到n个子模型。然后，获取未标注的第一样本，其包括对应的n组特征。将第一样本的除第i组特征外的(n-1)组特征，分别对应输入n个子模型中除第i子模型外的(n-1)个子模型，分别得到(n-1)个打分。接着，基于这(n-1)个打分，得到第一综合分，并且在该第一综合分满足预定条件的情况下，为第一样本的第i组特征添加第一标签，由此形成第一子标注样本。于是，可以将该第一子标注样本添加到第i个子标注样本集，以更新标注样本集。

训练样本获取方法，账户预测方法及对应装置

技术领域

[01] 本说明书一个或多个实施例涉及机器学习领域，尤其涉及利用机器学习模型进行非法账户预测，以及为该机器学习模型获取训练样本的方法和装置。

背景技术

[02] 随着互联网的发展，移动支付的普及，基于支付宝等 app 的移动支付手段越受青睐。然而，与此同时，相关的问题接踵而至，对于移动支付平台而言，一个重要的威胁来自于非法账户的存在以及其恶性发展。非法用户注册大量的备用账户，并通过非法手段进行花呗套现等行为，这对于移动支付平台而言是极大的威胁。对于非法账户的检测以及对其潜在非法行为的禁止，对于构建更为安全稳定的移动支付平台，减少相关平台的经济损失，具有重要意义。

[03] 目前业界对于非法账户检测的系统，几乎都是依赖于规则实现，如此的系统只能覆盖极少的非法账户类型，对于潜在非法账户，很难做到及时的发现。

[04] 而基于机器学习的方法，大部分都采用监督学习的方式，即利用完全标注的数据来学习，此时需要花费极大的时间和精力来进行数据标注，在互联网的场景下很难做到。

[05] 因此，希望能有改进的方案，更加有效地对非法账户进行预测。

发明内容

[06] 本说明书一个或多个实施例描述了训练样本获取方法，账户预测方法及对应装置，从而基于较少的人工标注样本，通过子模型协同训练的方式，扩充训练样本集，得到可靠的预测模型，用以进行异常账户的预测。

[07] 根据第一方面，提供了一种获取训练样本的方法，包括:

[08] 获取标注样本集，所述标注样本集包括 M 个标注样本，每个标注样本包括与账户信息相关联的样本特征，以及该账户是否为异常账户的样本标签，其中所述样本特征按照预定分组规则被划分为 n 组特征，其中 n 为大于 2 的自然数;

[09] 形成 n 个子标注样本集，其中第 i 个子标注样本集包括 M 个子标注样本，每个子标注样本包括所述 n 组特征中的第 i 组特征作为子样本特征，以及所述样本标签作为其

子样本标签;

[10] 分别利用所述 n 个子标注样本集训练得到 n 个子模型，其中第 i 子模型用于基于第 i 组特征预测对应账户为异常账户的概率;

[11] 获取多个未标注样本，每个未标注样本包括按照所述预定分组规则进行划分的 n 组特征，所述多个未标注样本包括第一样本;

[12] 将所述第一样本的 n 组特征中除第 i 组特征外的（n-1）组特征，分别对应输入所述 n 个子模型中除第 i 子模型外的（n-1）个子模型，分别得到所述（n-1）个子模型对该第一样本的（n-1）个打分，所述打分表示与该第一样本对应的账户为异常账户的概率;

[13] 基于所述（n-1）个打分，得到针对第 i 组特征的第一综合分;

[14] 在所述第一综合分满足预定条件的情况下，为所述第一样本的第 i 组特征添加第一标签，所述第 i 组特征和所述第一标签形成第一子标注样本;

[15] 将所述第一子标注样本添加到所述第 i 个子标注样本集，以更新所述第 i 个子标注样本集。

[16] 在一个实施例中，所述 n 组特征包括以下特征组中的多个：账户对应的用户的基本属性特征；用户的历史行为特征；用户的关联关系特征；用户的交互特征。

[17] 根据一种实施方式，通过以下方式得到上述第一综合分:

[18] 对所述（n-1）个打分求和，将和值作为所述第一综合分；或者

[19] 对所述（n-1）个打分求平均，将平均值作为所述第一综合分。

[20] 根据一种可能的设计，在所述第一综合分高于第一阈值的情况下，为所述第一样本的第 i 组特征添加异常账户的标签；在所述第一综合分低于第二阈值的情况下，为所述第一样本的第 i 组特征添加正常账户的标签，所述第二阈值小于所述第一阈值。

[21] 根据另一种可能的设计，还针对所述多个未标注样本，对应得到针对第 i 组特征的多个综合分;

[22] 如果所述第一综合分在所述多个综合分的从大到小排序中位于前端的第一数目之内，为所述第一样本的第 i 组特征添加异常账户的标签;

[23] 如果所述第一综合分在所述多个综合分的从大到小排序中位于后端的第二数目之

内，为所述第一样本的第 i 组特征添加正常账户的标签。

[24] 根据一种实施方式，方法还包括，用更新后的第 i 个子标注样本集，再次训练所述第 i 子模型。

[25] 根据第二方面，提供一种账户预测方法，包括：

[26] 获取待测账户的账户特征；

[27] 按照预定分类规则将所述账户特征划分为 n 组特征；

[28] 将所述 n 组特征分别输入 n 个子模型，得到所述 n 个子模型对所述待测账户异常概率的 n 个打分，所述 n 个子模型利用权利要求 1 的方法所获取的训练样本训练得到；

[29] 根据所述 n 个打分，确定所述待测账户的总得分；

[30] 根据所述总得分，确定所述待测账户的预测结果。

[31] 在一个实施例中，通过以下方式确定所述账户的总得分：

[32] 对所述 n 个打分求和，将和值作为总得分；或者

[33] 对所述 n 个打分求平均，将均值作为总得分。

[34] 根据一种可能的设计，如下确定所述待测账户的预测结果：

[35] 在所述总得分大于预定阈值的情况下，确定所述待测账户为异常账户。

[36] 根据另一种可能的设计，根据所述总得分，确定所述待测账户为异常账户的概率值，将该概率值作为预测结果。

[37] 根据第三方面，提供一种获取训练样本的装置，包括：

[38] 标注样本获取单元，配置为获取标注样本集，所述标注样本集包括 M 个标注样本，每个标注样本包括与账户信息相关联的样本特征，以及该账户是否为异常账户的样本标签，其中所述样本特征按照预定分组规则被划分为 n 组特征，其中 n 为大于 2 的自然数；

[39] 子样本集形成单元，配置为形成 n 个子标注样本集，其中第 i 个子标注样本集包括 M 个子标注样本，每个子标注样本包括所述 n 组特征中的第 i 组特征作为子样本特征，以及所述样本标签作为其子样本标签；

[40] 子模型训练单元，配置为分别利用所述 n 个子标注样本集训练得到 n 个子模型，其中第 i 子模型用于基于第 i 组特征预测对应账户为异常账户的概率；

[41] 未标注样本获取单元，配置为获取多个未标注样本，每个未标注样本包括按照所述预定分组规则进行划分的 n 组特征，所述多个未标注样本包括第一样本；

[42] 打分获取单元，配置为将所述第一样本的 n 组特征中除第 i 组特征外的（n-1）组特征，分别对应输入所述 n 个子模型中除第 i 子模型外的（n-1）个子模型，分别得到所述（n-1）个子模型对该第一样本的（n-1）个打分，所述打分表示与该第一样本对应的账户为异常账户的概率；

[43] 综合分获取单元，配置为基于所述（n-1）个打分，得到针对第 i 组特征的第一综合分；

[44] 标签添加单元，配置为在所述第一综合分满足预定条件的情况下，为所述第一样本的第 i 组特征添加第一标签，所述第 i 组特征和所述第一标签形成第一子标注样本；

[45] 样本添加单元，配置为将所述第一子标注样本添加到所述第 i 个子标注样本集，以更新所述第 i 个子标注样本集。

[46] 根据第四方面，提供一种账户预测装置，包括：

[47] 特征获取单元，配置为获取待测账户的账户特征；

[48] 特征分组单元，配置为按照预定分类规则将所述账户特征划分为 n 组特征；

[49] 打分获取单元，配置为将所述 n 组特征分别输入 n 个子模型，得到所述 n 个子模型对所述待测账户异常概率的 n 个打分，所述 n 个子模型利用权利要求 11 的装置所获取的训练样本训练得到；

[50] 总分确定单元，配置为根据所述 n 个打分，确定所述待测账户的总得分；

[51] 结果确定单元，配置为根据所述总得分，确定所述待测账户的预测结果。

[52] 根据第五方面，提供了一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行第一方面和第二方面的方法。

[53] 根据第六方面，提供了一种计算设备，包括存储器和处理器，其特征在于，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现第一方面和第二方面的方法。

[54] 通过本说明书实施例提供的方法和装置，采用半监督和多个子模型协同训练的方式，基于较少的人工标注数据，训练出多个可靠的子模型。在对待测账户进行预测时，利用如此训练的多个子模型分别进行预测，然后对结果进行综合，从而得到可靠的预测

结果。

附图说明

[55] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例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的前提下，还可以根据这些附图获得其它的附图。

[56] 图 1 为本说明书披露的一个实施例的实施场景示意图；

[57] 图 2 示出根据一个实施例的获取训练样本的方法流程图；

[58] 图 3 示出根据一个实施例形成的子标注样本集；

[59] 图 4 示出根据一个实施例的协同训练过程的示意图；

[60] 图 5 示出根据一个实施例的账户预测方法的流程图；

[61] 图 6 示出基于图 4 训练得到的模型进行账户预测的过程；

[62] 图 7 示出根据一个实施例的获取训练样本的装置的示意性框图；

[63] 图 8 示出根据一个实施例的账户预测装置的示意性框图。

具体实施方式

[64] 下面结合附图，对本说明书提供的方案进行描述。

[65] 如前所述，已经存在一些基于规则的方法来判断一个账户是否存在异常或是否是非法账户。然而这样的方案，很难做到对大量的非法账户的行为模式的覆盖。因此，仍然希望基于机器学习的方式来构建非法账户的检测系统，从而更加全面地对异常账户进行检测。然而，在互联网场景下，常规的机器学习方式存在一些困难，使其效果不够理想。

[66] 本案的发明人经过研究和分析提出，常规机器学习效果不够理想的原因至少有以下几点。一方面，监督学习需要大量的标注样本，标注样本越多，学习效果越好。但是，在非法账户预测的问题上，样本标注需要花费极大的时间和精力，因为要鉴别一个账户是不是真的非法账户，需要消耗巨大的人力，因此只有很少的账户是被标注出来非法或者合法的标记，大量的账户都是没有任何标记信息的。这使得可供监督学习的标注样本

数量不足，影响学习效果。另一方面，为了使得机器学习更加全面，往往对用户相关的大量特征进行采集。这就使得用来描述一个用户的特征向量变得非常大（如 5000+维），其中必然存在大量的信息冗余，而对于机器学习系统而言，这样的特征向量更是对于系统的效率带来极大的挑战。但是，如果简单地抛弃这一部分数据，又有可能造成信息损失，影响学习效果。因此，关于特征数据的采集，也存在两难问题。

[67] 基于以上的观察和分析，本案发明人创新性提出一种全新解决方案，采用半监督和多模型协同训练的方式提高机器学习的性能。具体来说，在说明书实施例的方案中，将用户特征进行分组，基于协同训练的方式对每组特征分别使用，来对无标记样本逐步打标，扩充标记样本集，来构建更为鲁棒的系统。

[68] 图 1 为本说明书披露的一个实施例的实施场景示意图。该实施场景可以划分为模型训练阶段，和模型预测阶段。在模型训练阶段，计算平台首先获取有标记的账户样本数据，这部分样本数据可能数量并不多。对于这些样本数据，将其特征划分为多个特征组（图 1 中示意性示出 3 个组），利用每个特征组和样本标签，训练出一个对应的子模型，图 1 中示意性示出 3 个子模型。然后，利用这 3 个子模型协同训练的方式，逐步给无标记数据打标。具体地，对于一条无标记样本数据，也对应的将样本特征划分为多个特征组（例如 3 个）。对于任意一个子模型 Mi 和对应的特征组 Fi，将其余的特征组分别输入对应的其余子模型，根据其余子模型的输出结果，为该条样本数据赋予标签，于是，特征组 Fi 和赋予的标签就可以作为用于训练子模型 Mi 的新的训练样本。如此，逐步给无标记样本数据赋予标签，扩充训练样本集。然后，可以利用不断扩充的训练样本集继续对各个子模型进行训练，提升其性能。

[69] 在模型预测阶段，接收到待测的账户数据。将该账户数据的样本特征也划分为多个特征组（例如 3 个），将这多个特征组分别对应输入训练得到的多个子模型，分别获得其输出结果。综合多个子模型的输出结果，判断待测账户是否为异常账户。

[70] 下面描述以上实施场景中的具体实施过程。

[71] 图 2 示出根据一个实施例的获取训练样本的方法流程图。该方法可以通过任何具有计算、处理能力的装置、设备、平台、设备集群等来执行，例如图 1 的计算平台。如图 2 所示，该方法至少包括：步骤 21，获取标注样本集，其中每个标注样本的样本特征被划分为 n 组特征；步骤 22，形成 n 个子标注样本集，其中第 i 个子标注样本集包括 n 组特征中的第 i 组特征，以及样本标签；步骤 23，利用所述 n 个子标注样本集训练得到 n 个子模型；步骤 24，获取多个未标注样本，其中包括第一样本，第一样本包括对应的

n 组特征；步骤 25，将第一样本的 n 组特征中除第 i 组特征外的（n-1）组特征，分别对应输入所述 n 个子模型中除第 i 子模型外的（n-1）个子模型，分别得到所述（n-1）个子模型对该第一样本的（n-1）个打分；步骤 26，基于所述（n-1）个打分，得到第一综合分；步骤 27，在第一综合分满足预定条件的情况下，为第一样本的第 i 组特征添加第一标签，所述第 i 组特征和所述第一标签形成第一子标注样本；步骤 28，将所述第一子标注样本添加到所述第 i 个子标注样本集，以更新所述第 i 个子标注样本集。下面描述上述各个步骤的执行方式。

[72] 首先，在步骤 21，获取标注样本集。一般地，标注样本集包括多个标注样本（下面为了描述方便，记为 M 个），每个标注样本包括与账户信息相关联的样本特征，以及该账户是否为异常账户的样本标签。

[73] 如前所述，为了使得模型训练的更加全面，样本特征往往包含与账户对应的用户的各个方面的全面特征，由此形成上千甚至几千维的特征向量。在本说明书的实施例中，按照预定分组规则，将样本特征划分为 n 个特征组，其中 n 为大于 2 的自然数。

[74] 在一个具体例子中，将账户样本的样本特征划分为 3 个特征组，例如，第一特征组包括与用户基本属性信息相关的特征，例如用户的性别、年龄、学历、注册时长等；第二特征组包括与用户的历史行为相关的特征，例如，用户最近一周的浏览记录，消费记录，交易记录等等；第三特征组包括与用户的社交关系相关的特征，例如与用户存在交互的好友的基本信息（年龄、性别、学历等）。如果一个样本的总体特征为 1500 维的向量，那么可以通过以上分组，形成 3 个约 500 维的分组特征向量。

[75] 可以理解，以上仅仅是一个示例。在其他实施例中，还可以从不同角度，对样本特征进行不同的分组，例如划分为不同数目的特征组，或者不同内容的特征组。例如，还可以从样本特征中提取出与用户和好友之间的交互行为相关的特征，形成交互特征组；或者专门提取出与用户的信贷行为相关的特征，形成信贷特征组，等等。本说明书对于样本特征的分组方式不作限定。

[76] 接着，在步骤 22，与划分的 n 个特征组相对应的，形成 n 个子标注样本集，其中第 i 个子标注样本集包括 M 个子标注样本，每个子标注样本包括 n 组特征中的第 i 组特征作为子样本特征，以及样本标签作为其子样本标签，其中 i 为 1 到 n 中任意的值。

[77] 下面延用 3 个特征组的例子，结合图 3 描述子标注样本集的形成。

[78] 图 3 示出根据一个实施例形成的子标注样本集。在图 3 的示例中，假定标注样本

集包含 100 个（即 M=100）标注样本，每个标注样本 Y_j 包含样本特征 F_j 和样本标签 L_j，而样本特征 F_j 都被划分为 3 个特征组，即 $F_j=G_{j1}+G_{j2}+G_{j3}$。例如，第 1 条样本的样本特征 F_1 划分为 3 个特征组 G_{11}，G_{12} 和 G_{13}，样本标签为 L_1；第 2 条样本的样本特征 F_2 被划分为 G_{21}，G_{22}，G_{23}，样本标签为 L_2，等等。

[79] 根据一个实施例，提取各个标注样本的第 i 组特征和样本标签，形成各个子标注样本，各个子标注样本构成第 i 个子标注样本集。例如，子标注样本集 1 对应于特征组 1，其中包括 100 个子标注样本，每个子标注样本包括特征组 1 中的特征作为子样本特征，以及样本标签作为其子样本标签。例如，提取第 1 条样本的特征组 1 中的特征，即 G_{11}，以及对应标签 L_1，形成基于特征组 1 的一条子标注样本；提取第 2 条样本的特征组 1 中的特征，即 G_{21}，以及对应标签 L_2，形成基于特征组 1 的另一条子标注样本，等等，如此，可以形成基于特征组 1 的 100 个子标注样本，从而构成第 1 子标注样本集。

[80] 类似地，子标注样本集 2 对应于特征组 2，其中的每个子标注样本包括特征组 2 中的特征作为子样本特征，以及样本标签作为其子样本标签；子标注样本集 3 对应于特征组 3，其中的每个子标注样本包括特征组 3 中的特征作为子样本特征，以及样本标签作为其子样本标签。

[81] 更多特征组的情况可以以此类推，不再赘述。

[82] 在形成了 n 个子标注样本集的基础上，在步骤 23，分别利用所述 n 个子标注样本集进行模型训练，得到对应的 n 个子模型，其中第 i 子模型用于基于第 i 组特征预测对应账户为异常账户的概率。

[83] 可以理解，对于 n 个子标注样本集中任意的第 i 子标注样本集，由于具有样本标签，因此可以采用各种监督学习的方式进行模型训练，得到对应的子模型 Mi。由于第 i 子标注样本集基于第 i 组特征形成，而样本标签用于示出对应账户是否为异常账户，那么相应地，基于第 i 子标注样本集训练得到的第 i 子模型 Mi 用于基于第 i 组特征预测对应账户为异常账户的概率。

[84] 延用图 3 中的例子，如果形成了 3 个子标注样本集，那么可以对应训练出 3 个子模型。

[85] 如此训练得到的子模型可以作为初步的子模型，协同作用，为未标注样本进行打标。

[86] 于是，接下来，在步骤 24，获取多个未标注样本，每个未标注样本包括对应的样

本特征，但是不具有样本标签。对于未标注样本的样本特征，按照处理标注样本同样的分组规则，将样本特征进行划分，划分为 n 个特征组。分组的过程不再赘述。

[87] 为了描述的方便，将上述多个未标注样本中的某个不特定样本称为第一样本，结合该第一样本描述利用上述 n 个子模型对其进行打标的过程。需要理解，此处第一样本中的“第一”，以及下文中相应的“第一”，仅仅是为了区分和描述方便，而不具有任何限定意义。

[88] 在步骤 25，将第一样本的 n 组特征中除第 i 组特征外的（n-1）组特征，分别对应输入所述 n 个子模型中除第 i 子模型外的（n-1）个子模型，分别得到所述（n-1）个子模型对该第一样本的（n-1）个打分。根据前述的各个子模型的训练说明，各个子模型用于基于对应的特征组预测对应账户为异常账户的概率，因此，各个子模型对第一样本的打分即表示，与该第一样本对应的账户为异常账户的概率。

[89] 然后在步骤 26，基于所述（n-1）个子模型输出的（n-1）个打分，得到第一样本的针对第 i 组特征的第一综合分。

[90] 在一个实施例中，对上述（n-1）个打分求和，将和值作为所述第一综合分。更具体地，在一个例子中，上述求和可以是加权求和。在这样的情况下，可以根据各个子模型的重要性、可靠性等因素，预先为各个子模型设置对应的权重。如此，对于上述（n-1）个子模型的打分，将各个子模型的权重作为对应打分的权重，对（n-1）个打分进行加权求和，得到上述第一综合分。

[91] 在另一实施例中，对所述（n-1）个打分求平均，将平均值作为所述第一综合分。在其他实施例中，还可以采用其他方式对（n-1）个打分进行综合，得到第一综合分。

[92] 接着，对该第一综合分进行判断。在步骤 27，在该第一综合分满足预定条件的情况下，为第一样本的第 i 组特征添加第一标签。

[93] 在不同实施例中，上述预定条件和对应的第一标签的内容可以有不同实现方式。

[94] 在一个实施例中，预先设定综合分的判断阈值，例如较高的第一阈值，和较低的第二阈值，根据阈值比较结果确定第一标签的添加。具体地，如果第一样本的针对第 i 组特征的第一综合分高于第一阈值，那么说明，除第 i 子模型外的其他各个子模型预测第一样本为异常账户的总体概率足够高，因此，为第一样本的第 i 组特征添加异常账户的标签。如果第一综合分低于第二阈值，则说明，除第 i 子模型外的其他各个子模型预测第一样本为异常账户的总体概率很低，因此，为第一样本的第 i 组特征添加正常账户

的标签。可选的，如果第一综合分在第一阈值和第二阈值之间，那么有可能其他各个子模型对第一样本是否为异常账户的预测概率差异较大，或者说，预测结果并不一致，在这样的情况下，可以暂时不为第一样本的第 i 组特征添加标签。

[95] 在另一实施例中，设定综合分的排名阈值，根据排名阈值确定第一标签的添加。可以理解，以上描述的获取第一样本针对第 i 组特征的第一综合分的过程，可以应用于步骤 24 获取的多个未标注样本中的每个样本，由此可以获得各个未标注样本的针对第 i 组特征的综合分，由此得到多个综合分。可以将这多个综合分进行从大到小的排序。

[96] 如果第一样本对应的第一综合分在上述排序中位于靠前的第一数目（例如 50 个）之内，为第一样本的第 i 组特征添加异常账户的标签。换而言之，如果第一样本对应的第一综合分属于所有未标注样本中综合分最高的第一数目个，例如综合分最高的 50 个，那么说明，除第 i 子模型外的其他各个子模型预测第一样本为异常账户的总体概率足够高，因此，为第一样本的第 i 组特征添加异常账户的标签。对应的，如果第一综合分在多个综合分的从大到小排序中位于后端的第二数目之内，即属于分数最低的第二数目个，则说明，除第 i 子模型外的其他各个子模型预测第一样本为异常账户的总体概率足够低，因此为第一样本的第 i 组特征添加正常账户的标签。上述第一数目和第二数目可以根据未标注样本的数量而设定，两者可以相等，也可以不相等。

[97] 如此，对于第一样本，将除第 i 组特征外的其他组特征对应输入除第 i 子模型外的其他子模型，基于各个子模型的预测结果，得出针对第 i 组特征的综合分，并基于该综合分为该第 i 组特征添加第一标签。第一样本的第 i 组特征，以及添加的第一标签，形成一个子标注样本，称为第一子标注样本。

[98] 接着，在步骤 28，将如此得到的第一子标注样本添加到前述的第 i 个子标注样本集，以更新所述第 i 个子标注样本集。

[99] 可以理解，通过对各个未标注样本的各个特征组 i 进行以上步骤 25 到步骤 28 的操作过程，可以不断筛选出（n-1）个子模型的预测结果相对一致的未标注样本，为其特征组 i 添加与预测结果对应的标签，从而得到新的子标注样本，如此不断扩充各个子标注样本集，增加训练样本的数量。

[100] 下面仍沿用之前 n=3 的例子，描述以上过程。

[101] 在一个例子中，获取了例如 1000 条未标注样本，每个未标注样本的样本特征被分为 3 个特征组。相应地，假定其中第一样本的样本特征被划分为 U1，U2 和 U3。在步

骤 25，在 i=1 的情况下，将这 3 组特征中除第 1 组特征 U1 外的 2 组特征，即 U2 和 U3，分别对应输入 3 个子模型中除第 1 子模型外的 2 个子模型，即 M2 和 M3，分别得到这 2 个子模型对该第一样本的 2 个打分，记为 c2 和 c3。

[102] 然后，在步骤 26，基于上述 c2 和 c3，得到第一样本的针对第 1 组特征的第一综合分 S1。例如，S1 可以是 c2 和 c3 的和值或均值，等等。

[103] 在步骤 27，判断该第一综合分 S1 是否满足预定条件，以此确定是否添加标签以及添加什么标签。例如，在一个例子中，如果 S1 大于阈值 T1，则为第一样本的第 1 组特征添加异常账户的标签；如果 S1 小于阈值 T2，则为其添加正常账户的标签。

[104] 或者，在另一例子中，根据该综合分的排名来确定标签的添加。例如，获取 1000 个未标注样本中各个未标注样本针对第 1 组特征的综合分，如此得到 1000 个综合分。可以对这 1000 个综合分进行排序。如果 S1 属于这 1000 个综合分中得分最高的例如前 50 个，那么就为第一样本的第 1 组特征添加异常账户的标签；如果 S1 属于这 1000 个综合分中得分最低的例如后 50 个，那么就为第一样本的第 1 组特征添加正常账户的标签。

[105] 于是，第一样本的第 1 组特征和对应的标签就构成一条子标注样本，添加到针对第 1 组特征的子标注样本集中。

[106] 类似的，可以针对每条未标注样本（1-1000 条）的每组特征（第 1 组特征，第 2 组特征，第 3 组特征）进行类似的处理，从而为部分未标注样本添加标签，扩充到训练样本集中。

[107] 在一个实施例中，在如此更新或扩充各个第 i 个子标注样本集后，用更新或扩充的第 i 个子标注样本集，再次训练第 i 子模型。之后，可以用再次训练的子模型来对接下来的未标注样本进行预测和打标。如此不断重复循环，利用多个子模型的预测结果进行自动打标来扩大对于另一子模型的训练样本集，再用扩大的训练样本集再次训练子模型，使得整个系统的鲁棒性不断提升。

[108] 图 4 示出根据一个实施例的协同训练过程的示意图。在图 4 的例子中，假设最初有 100 个账户的标注样本，每个标注样本的样本特征信息被分为 3 组，分别为账户的基本信息，动态信息和关系信息。由此，形成与 3 个特征组对应的 3 个子标注样本集，每个子标注样本集包含 100 个子标注样本。基于这 3 个子标注样本集，训练得到 3 个初始子模型，图 4 中表示为模型 1，模型 2 和模型 3。

[109] 另一方面，假定获取到 1000 条未标注样本数据。同样地，将每个未标注样本的特征信息划分为 3 组：基本信息，动态信息和关系信息。将各个未标注样本的各个特征组相应输入对应的子模型，例如将基本信息输入模型 1，动态信息输入模型 2，关系信息输入模型 3，分别获得各个子模型的预测结果，即打分，基于这些打分进行无标记数据的筛选和打标。

[110] 具体地，对于某条未标注样本，为了对其基本信息（组 1）进行打标，就考虑模型 2 对其动态信息（组 2）的打分 c2 以及模型 3 对其关系信息（组 3）的打分 c3，基于这两个打分得出一个综合分 S1。如果综合分 S1 满足一定条件，例如数值阈值条件，或者排名条件，则为该样本的基本信息对应添加异常账户/正常账户的标签，形成标记数据。

[111] 或者，从 1000 条未标注样本的整体来看，在利用模型 1，模型 2 和模型 3 分别对各个未标记样本的基本信息、动态信息、关系信息进行打分后，对各个样本，基于模型 2 和 3 的打分综合得到 S1，如此得到 1000 个样本各自的 S1。从中选择 S1 大小超过一定数值阈值的，或者选择 S1 最大的若干个（例如 50 个）样本，获取这些样本的基本信息，为其加上异常账户的标签；选择 S1 小于另一较小数值阈值的，或者选择 S1 最小的若干个（例如 50 个）样本，提取这些样本的基本信息，为其附上正常账户的标签。

[112] 同理，基于模型 1 和 3 的打分综合得到 S2。根据 S2 对应的数值阈值或排名，从 1000 个未标记样本中筛选出若干个，提取其动态信息，为其添加正常账户/异常账户的标签。基于模型 1 和 2 的打分综合得到 S3。根据 S3 对应的数值阈值或排名，选择若干个样本，提取其关系信息，并附上正常账户/异常账户的标签。

[113] 换而言之，选择任意两个模型认为最可靠的那部分未标记样本，添加上标签，作为另外一个模型的训练样本。

[114] 如此添加了标签的样本与原标注样本集中的样本可以融合在一起，形成更新或扩充的训练样本集。例如，如果针对一组特征，每次选择另外 2 个模型的综合分最高的 50 个样本，添加异常账户标签，选择综合分最低的 50 个样本添加正常账户标签，那么执行一次上述过程，可以将各个子标注样本集的样本数目扩充到 200 个。

[115] 之后，可以利用更新的训练样本集再次训练各个子模型，如此不断重复循环，训练样本集越来越丰富，子模型的预测性能也越来越可靠，整个系统的鲁棒性不断提升。如此，利用数量较少的人工标注样本，就可以得到性能可靠的预测系统。

[116] 在反复训练各个子模型，使其可靠性达到一定程度之后，就可以利用如此训练得

到的子模型所构成的总模型，对未知账户进行预测。

[117] 图 5 示出根据一个实施例的账户预测方法的流程图。如图 5 所示，该预测方法包括：步骤 51，获取待测账户的账户特征；步骤 52，按照预定分组规则将所述账户特征划分为 n 组特征；步骤 53，将所述 n 组特征分别输入 n 个子模型，得到所述 n 个子模型对所述待测账户异常概率的 n 个打分；步骤 54，根据所述 n 个打分，确定所述待测账户的总得分；步骤 55，根据所述总得分，确定所述待测账户的预测结果。

[118] 下面结合图 6 的例子描述以上过程。图 6 示出基于图 4 训练得到的模型进行账户预测的过程。

[119] 如图 5 和图 6 所示，首先在步骤 51，获取待测账户的账户特征。一般地，账户特征包括与账户相关的多方面的特征，维度可达上千或者几千维。

[120] 接着，在步骤 52，按照预定分组规则，将账户特征划分为 n 组特征。可以理解，此处的分组规则与模型训练过程中对训练样本的样本特征进行分组的规则一致。

[121] 例如，在图 6 中，与图 4 对应的，将待测账户的特征分为 3 组，即基本信息，动态信息，关系信息。

[122] 接着，在步骤 53，将 n 组特征分别输入 n 个子模型，得到 n 个子模型对所述待测账户异常概率的 n 个打分。可以理解，此处的 n 个子模型是利用图 2 方法获取的训练样本训练得到的。因此，这 n 个子模型与 n 组特征分别对应，第 i 子模型被训练为，基于第 i 组特征为对应账户打分，该打分表示对应账户为异常账户的概率。如此，n 个子模型对应输出 n 个打分。

[123] 在图 6 中，将待测账户的基本信息输入模型 1，得到预测结果 1；将动态信息输入模型 2，得到预测结果 2；将关系信息输入模型 3，得到预测结果 3。各个预测结果即对应于上述打分，表示对应模型预测的该账户为异常账户的概率。

[124] 接着，在步骤 54，根据上述 n 个打分，确定待测账户的总得分。

[125] 在一个实施例中，在步骤 54，对所述 n 个打分求和，将和值作为总得分。更具体地，在一个例子中，上述求和可以是加权求和。即根据各个子模型的重要性、可靠性等因素，预先为各个子模型设置对应的权重。如此，对于上述 n 个子模型的打分，将各个子模型的权重作为对应打分的权重，对 n 个打分进行加权求和，得到总得分。

[126] 在另一实施例中，还可以对所述 n 个打分求平均，将均值作为总得分。或者，在

其他实施例中，还可以采取其他方式基于这 n 个打分确定出总得分。

[127] 接着，在步骤 55，根据上述总得分，确定待测账户的预测结果。

[128] 在一个实施例中，输出的预测结果为，待测账户是正常账户还是异常账户的判断结果。在这样的情况下，可以将上述总得分与一概率阈值进行比较，在总得分大于该概率阈值的情况下，确定待测账户为异常账户，否则，确定待测账户为正常账户。

[129] 在另一实施例中，输出的预测结果为，待测账户是异常账户的概率。更具体地，在一个例子中，上述总得分是对 n 个打分求平均而得到；在这样的情况下，可以直接将上述总得分作为异常账户的概率值，而输出作为预测结果。在另一例子中，上述总得分通过其他方式计算得到，在这样的情况下，可以对上述总得分进行简单的处理运算，例如归一化处理，将处理结果作为异常账户的概率值，输出作为预测结果。

[130] 在图 6 中，该过程简单示出为，将预测结果 1，预测结果 2 和预测结果 3 进行综合，得到最终预测结果。

[131] 通过以上过程可以看到，在实施例的方案中，将待测账户的特征数据划分为多个组，分别输入对应的多个子模型，再对子模型的预测结果进行综合。如此，既避免了特征数据维度太高对模型计算性能的影响，又不会因为丢弃数据造成信息损失。

[132] 综合以上，说明书实施例的方案采用半监督和多个子模型协同训练的方式，基于较少的人工标注数据，训练出多个可靠的子模型。在对待测账户进行预测时，利用如此训练的多个子模型分别进行预测，然后对结果进行综合，从而得到可靠的预测结果。

[133] 根据另一方面的实施例，还提供一种获取训练样本的装置。图 7 示出根据一个实施例的获取训练样本的装置的示意性框图。如图 7 所示，该装置 700 包括：

[134] 标注样本获取单元 71，配置为获取标注样本集，所述标注样本集包括 M 个标注样本，每个标注样本包括与账户信息相关联的样本特征，以及该账户是否为异常账户的样本标签，其中所述样本特征按照预定分组规则被划分为 n 组特征，其中 n 为大于 2 的自然数；

[135] 子样本集形成单元 72，配置为形成 n 个子标注样本集，其中第 i 个子标注样本集包括 M 个子标注样本，每个子标注样本包括所述 n 组特征中的第 i 组特征作为子样本特征，以及所述样本标签作为其子样本标签；

[136] 子模型训练单元 73，配置为分别利用所述 n 个子标注样本集训练得到 n 个子模型，

其中第 i 子模型用于基于第 i 组特征预测对应账户为异常账户的概率;

[137] 未标注样本获取单元 74，配置为获取多个未标注样本，每个未标注样本包括按照所述预定分组规则进行划分的 n 组特征，所述多个未标注样本包括第一样本;

[138] 打分获取单元 75，配置为将所述第一样本的 n 组特征中除第 i 组特征外的（n-1）组特征，分别对应输入所述 n 个子模型中除第 i 子模型外的（n-1）个子模型，分别得到所述（n-1）个子模型对该第一样本的（n-1）个打分，所述打分表示与该第一样本对应的账户为异常账户的概率;

[139] 综合分获取单元 76，配置为基于所述（n-1）个打分，得到针对第 i 组特征的第一综合分;

[140] 标签添加单元 77，配置为在所述第一综合分满足预定条件的情况下，为所述第一样本的第 i 组特征添加第一标签，所述第 i 组特征和所述第一标签形成第一子标注样本;

[141] 样本添加单元 78，配置为将所述第一子标注样本添加到所述第 i 个子标注样本集，以更新所述第 i 个子标注样本集。

[142] 根据一种实施方式，所述 n 组特征包括以下特征组中的多个：账户对应的用户的基本属性特征；用户的历史行为特征；用户的关联关系特征；用户的交互特征。

[143] 在一个实施例中，上述综合分获取单元 76 配置为:

[144] 对所述（n-1）个打分求和，将和值作为所述第一综合分；或者

[145] 对所述（n-1）个打分求平均，将平均值作为所述第一综合分。

[146] 根据一种可能的设计，所述标签添加单元 77 配置为:

[147] 在所述第一综合分高于第一阈值的情况下，为所述第一样本的第 i 组特征添加异常账户的标签;

[148] 在所述第一综合分低于第二阈值的情况下，为所述第一样本的第 i 组特征添加正常账户的标签，所述第二阈值小于所述第一阈值。

[149] 在一种实施方式中，综合分获取单元 76 配置为，针对所述多个未标注样本，对应得到针对第 i 组特征的多个综合分;

[150] 相应地，标签添加单元 77 配置为:

[151] 如果所述第一综合分在所述多个综合分的从大到小排序中位于前端的第一数目之

内，为所述第一样本的第 i 组特征添加异常账户的标签；

[152] 如果所述第一综合分在所述多个综合分的从大到小排序中位于后端的第二数目之内，为所述第一样本的第 i 组特征添加正常账户的标签。

[153] 在一个实施例中，子模型训练单元 73 还配置为，用更新后的第 i 个子标注样本集，再次训练所述第 i 子模型。

[154] 根据又一方面的实施例，还提供一种账户预测装置。图 8 示出根据一个实施例的账户预测装置的示意性框图。如图 8 所示，该装置 800 包括：

[155] 特征获取单元 81，配置为获取待测账户的账户特征；

[156] 特征分组单元 82，配置为按照预定分类规则将所述账户特征划分为 n 组特征；

[157] 打分获取单元 83，配置为将所述 n 组特征分别输入 n 个子模型，得到所述 n 个子模型对所述待测账户异常概率的 n 个打分，所述 n 个子模型利用权利要求 11 的装置所获取的训练样本训练得到；

[158] 总分确定单元 84，配置为根据所述 n 个打分，确定所述待测账户的总得分；

[159] 结果确定单元 85，配置为根据所述总得分，确定所述待测账户的预测结果。

[160] 在一个实施例中，总分确定单元 84 配置为：对所述 n 个打分求和，将和值作为总得分；或者，对所述 n 个打分求平均，将均值作为总得分。

[161] 根据一种可能的设计，结果确定单元 85 配置为：在所述总得分大于预定阈值的情况下，确定所述待测账户为异常账户。

[162] 根据另一种可能的设计，结果确定单元 85 配置为：根据所述总得分，确定所述待测账户为异常账户的概率值，将该概率值作为预测结果。

[163] 通过图 7 和图 8 的装置，采用半监督和多个子模型协同训练的方式，基于较少的人工标注数据，训练出多个可靠的子模型。在对待测账户进行预测时，利用如此训练的多个子模型分别进行预测，然后对结果进行综合，从而得到可靠的预测结果。

[164] 根据另一方面的实施例，还提供一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行结合图 2 和图 5 所描述的方法。

[165] 根据再一方面的实施例，还提供一种计算设备，包括存储器和处理器，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现结合图 2 和图 5 所述

的方法。

[166] 本领域技术人员应该可以意识到，在上述一个或多个示例中，本发明所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码进行传输。

[167] 以上所述的具体实施方式，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施方式而已，并不用于限定本发明的保护范围，凡在本发明的技术方案的基础之上，所做的任何修改、等同替换、改进等，均应包括在本发明的保护范围之内。

权利要求书

1.一种获取训练样本的方法，包括:

获取标注样本集，所述标注样本集包括 M 个标注样本，每个标注样本包括与账户信息相关联的样本特征，以及该账户是否为异常账户的样本标签，其中所述样本特征按照预定分组规则被划分为 n 组特征，其中 n 为大于 2 的自然数;

形成 n 个子标注样本集，其中第 i 个子标注样本集包括 M 个子标注样本，每个子标注样本包括所述 n 组特征中的第 i 组特征作为子样本特征，以及所述样本标签作为其子样本标签;

分别利用所述 n 个子标注样本集训练得到 n 个子模型，其中第 i 子模型用于基于第 i 组特征预测对应账户为异常账户的概率;

获取多个未标注样本，每个未标注样本包括按照所述预定分组规则进行划分的 n 组特征，所述多个未标注样本包括第一样本;

将所述第一样本的 n 组特征中除第 i 组特征外的（n-1）组特征，分别对应输入所述 n 个子模型中除第 i 子模型外的（n-1）个子模型，分别得到所述（n-1）个子模型对该第一样本的（n-1）个打分，所述打分表示与该第一样本对应的账户为异常账户的概率;

基于所述（n-1）个打分，得到针对第 i 组特征的第一综合分;

在所述第一综合分满足预定条件的情况下，为所述第一样本的第 i 组特征添加第一标签，所述第 i 组特征和所述第一标签形成第一子标注样本;

将所述第一子标注样本添加到所述第 i 个子标注样本集，以更新所述第 i 个子标注样本集。

2.根据权利要求 1 所述的方法，其中所述 n 组特征包括以下特征组中的多个：账户对应的用户的基本属性特征；用户的历史行为特征；用户的关联关系特征；用户的交互特征。

3.根据权利要求 1 所述的方法，其中基于所述（n-1）个打分，得到针对第 i 组特征的第一综合分包括:

对所述（n-1）个打分求和，将和值作为所述第一综合分；或者

对所述（n-1）个打分求平均，将平均值作为所述第一综合分。

4.根据权利要求 1 所述的方法，其中在所述第一综合分满足预定条件的情况下，为所述第一样本的第 i 组特征添加第一标签包括:

在所述第一综合分高于第一阈值的情况下，为所述第一样本的第 i 组特征添加异常账户的标签;

在所述第一综合分低于第二阈值的情况下，为所述第一样本的第 i 组特征添加正常账户的标签，所述第二阈值小于所述第一阈值。

5.根据权利要求 1 所述的方法，还包括，针对所述多个未标注样本，对应得到针对第 i 组特征的多个综合分；

所述在所述第一综合分满足预定条件的情况下，为所述第一样本的第 i 组特征添加第一标签包括：

如果所述第一综合分在所述多个综合分的从大到小排序中位于前端的第一数目之内，为所述第一样本的第 i 组特征添加异常账户的标签；

如果所述第一综合分在所述多个综合分的从大到小排序中位于后端的第二数目之内，为所述第一样本的第 i 组特征添加正常账户的标签。

6.根据权利要求 1 所述的方法，还包括，用更新后的第 i 个子标注样本集，再次训练所述第 i 子模型。

7.一种账户预测方法，包括：

获取待测账户的账户特征；

按照预定分类规则将所述账户特征划分为 n 组特征；

将所述 n 组特征分别输入 n 个子模型，得到所述 n 个子模型对所述待测账户异常概率的 n 个打分，所述 n 个子模型利用权利要求 1 的方法所获取的训练样本训练得到；

根据所述 n 个打分，确定所述待测账户的总得分；

根据所述总得分，确定所述待测账户的预测结果。

8.根据权利要求 7 所述的方法，其中根据所述 n 个打分，确定所述账户的总得分包括：

对所述 n 个打分求和，将和值作为总得分；或者

对所述 n 个打分求平均，将均值作为总得分。

9.根据权利要求 7 所述的方法，其中根据所述总得分，确定所述待测账户的预测结果包括：

在所述总得分大于预定阈值的情况下，确定所述待测账户为异常账户。

10.根据权利要求 7 所述的方法，其中根据所述总得分，确定所述待测账户的预测结果包括：

根据所述总得分，确定所述待测账户为异常账户的概率值，将该概率值作为预测结果。

11.一种获取训练样本的装置，包括：

标注样本获取单元，配置为获取标注样本集，所述标注样本集包括 M 个标注样本，每个标注样本包括与账户信息相关联的样本特征，以及该账户是否为异常账户的样本标签，其中所述样本特征按照预定分组规则被划分为 n 组特征，其中 n 为大于 2 的自然数；

子样本集形成单元，配置为形成 n 个子标注样本集，其中第 i 个子标注样本集包括 M 个子标注样本，每个子标注样本包括所述 n 组特征中的第 i 组特征作为子样本特征，以及所述样本标签作为其子样本标签；

子模型训练单元，配置为分别利用所述 n 个子标注样本集训练得到 n 个子模型，其中第 i 子模型用于基于第 i 组特征预测对应账户为异常账户的概率；

未标注样本获取单元，配置为获取多个未标注样本，每个未标注样本包括按照所述预定分组规则进行划分的 n 组特征，所述多个未标注样本包括第一样本；

打分获取单元，配置为将所述第一样本的 n 组特征中除第 i 组特征外的（n-1）组特征，分别对应输入所述 n 个子模型中除第 i 子模型外的（n-1）个子模型，分别得到所述（n-1）个子模型对该第一样本的（n-1）个打分，所述打分表示与该第一样本对应的账户为异常账户的概率；

综合分获取单元，配置为基于所述（n-1）个打分，得到针对第 i 组特征的第一综合分；

标签添加单元，配置为在所述第一综合分满足预定条件的情况下，为所述第一样本的第 i 组特征添加第一标签，所述第 i 组特征和所述第一标签形成第一子标注样本；

样本添加单元，配置为将所述第一子标注样本添加到所述第 i 个子标注样本集，以更新所述第 i 个子标注样本集。

12.根据权利要求 11 所述的装置，其中所述 n 组特征包括以下特征组中的多个：账户对应的用户的基本属性特征；用户的历史行为特征；用户的关联关系特征；用户的交互特征。

13.根据权利要求 11 所述的装置，其中所述综合分获取单元配置为：

对所述（n-1）个打分求和，将和值作为所述第一综合分；或者

对所述（n-1）个打分求平均，将平均值作为所述第一综合分。

14.根据权利要求 11 所述的装置，其中所述标签添加单元配置为：

在所述第一综合分高于第一阈值的情况下，为所述第一样本的第 i 组特征添加异常账户的标签；

在所述第一综合分低于第二阈值的情况下，为所述第一样本的第 i 组特征添加正常账户的标签，所述第二阈值小于所述第一阈值。

15.根据权利要求 11 所述的装置，所述综合分获取单元配置为，针对所述多个未标注样本，对应得到针对第 i 组特征的多个综合分;

所述标签添加单元配置为:

如果所述第一综合分在所述多个综合分的从大到小排序中位于前端的第一数目之内，为所述第一样本的第 i 组特征添加异常账户的标签;

如果所述第一综合分在所述多个综合分的从大到小排序中位于后端的第二数目之内，为所述第一样本的第 i 组特征添加正常账户的标签。

16.根据权利要求 11 所述的装置，所述子模型训练单元还配置为，用更新后的第 i 个子标注样本集，再次训练所述第 i 子模型。

17.一种账户预测装置，包括:

特征获取单元，配置为获取待测账户的账户特征;

特征分组单元，配置为按照预定分类规则将所述账户特征划分为 n 组特征;

打分获取单元，配置为将所述 n 组特征分别输入 n 个子模型，得到所述 n 个子模型对所述待测账户异常概率的 n 个打分，所述 n 个子模型利用权利要求 11 的装置所获取的训练样本训练得到;

总分确定单元，配置为根据所述 n 个打分，确定所述待测账户的总得分;

结果确定单元，配置为根据所述总得分，确定所述待测账户的预测结果。

18.根据权利要求 17 所述的装置，其中所述总分确定单元配置为:

对所述 n 个打分求和，将和值作为总得分; 或者

对所述 n 个打分求平均，将均值作为总得分。

19.根据权利要求 17 所述的装置，其中所述结果确定单元配置为:

在所述总得分大于预定阈值的情况下，确定所述待测账户为异常账户。

20.根据权利要求 17 所述的装置，其中所述结果确定单元配置为:

根据所述总得分，确定所述待测账户为异常账户的概率值，将该概率值作为预测结果。

21.一种计算设备，包括存储器和处理器，其特征在于，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现权利要求 1-10 中任一项所述的方法。

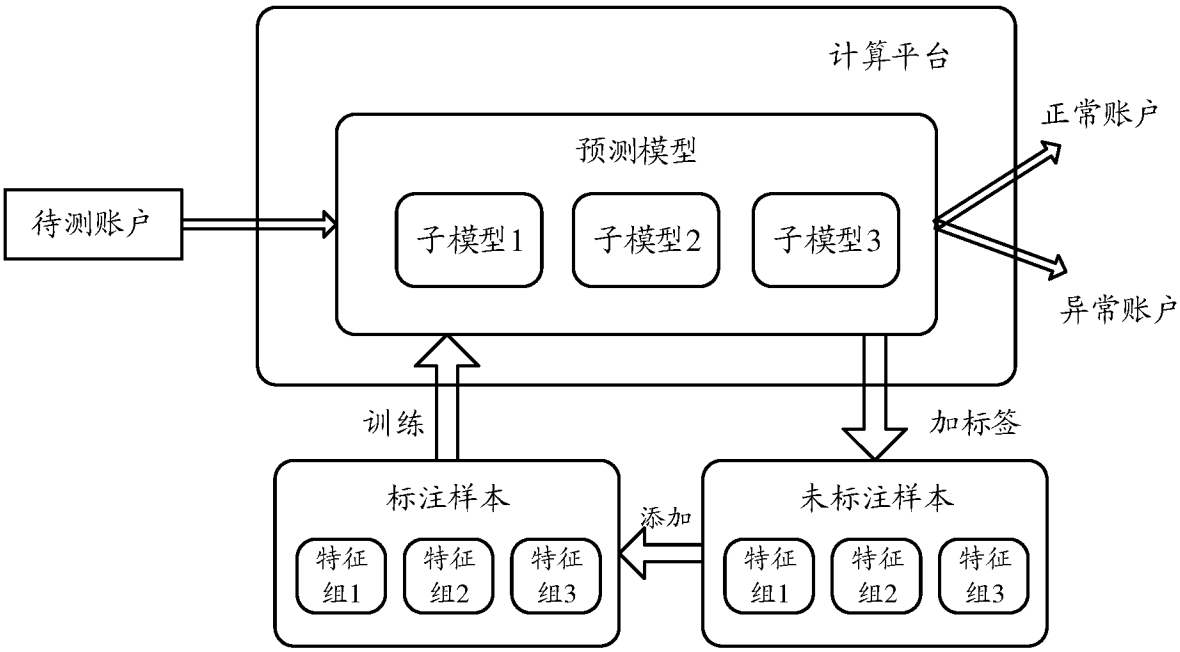
计算平台
预测模型
子模型1
子模型2
子模型3
待测账户
正常账户
异常账户
训练
加标签
标注样本
特征组1
特征组2
特征组3
添加
未标注样本
特征组1
特征组2
特征组3

图 1

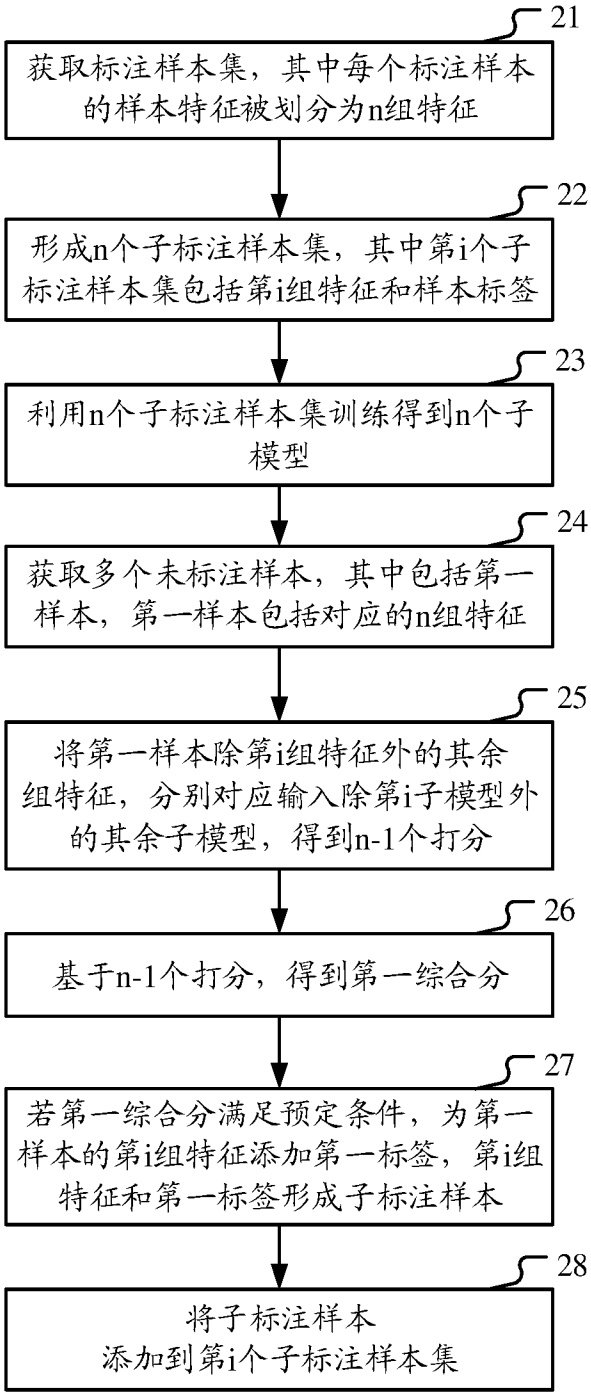
21
获取标注样本集，其中每个标注样本的样本特征被划分为n组特征
22
形成n个子标注样本集，其中第i个子标注样本集包括第i组特征和样本标签
23
利用n个子标注样本集训练得到n个子模型
24
获取多个未标注样本，其中包括第一样本，第一样本包括对应的n组特征
25
将第一样本除第i组特征外的其余组特征，分别对应输入除第i子模型外的其余子模型，得到n-1个打分
26
基于n-1个打分，得到第一综合分
27
若第一综合分满足预定条件，为第一样本的第i组特征添加第一标签，第i组特征和第一标签形成子标注样本
28
将子标注样本
添加到第i个子标注样本集

图 2

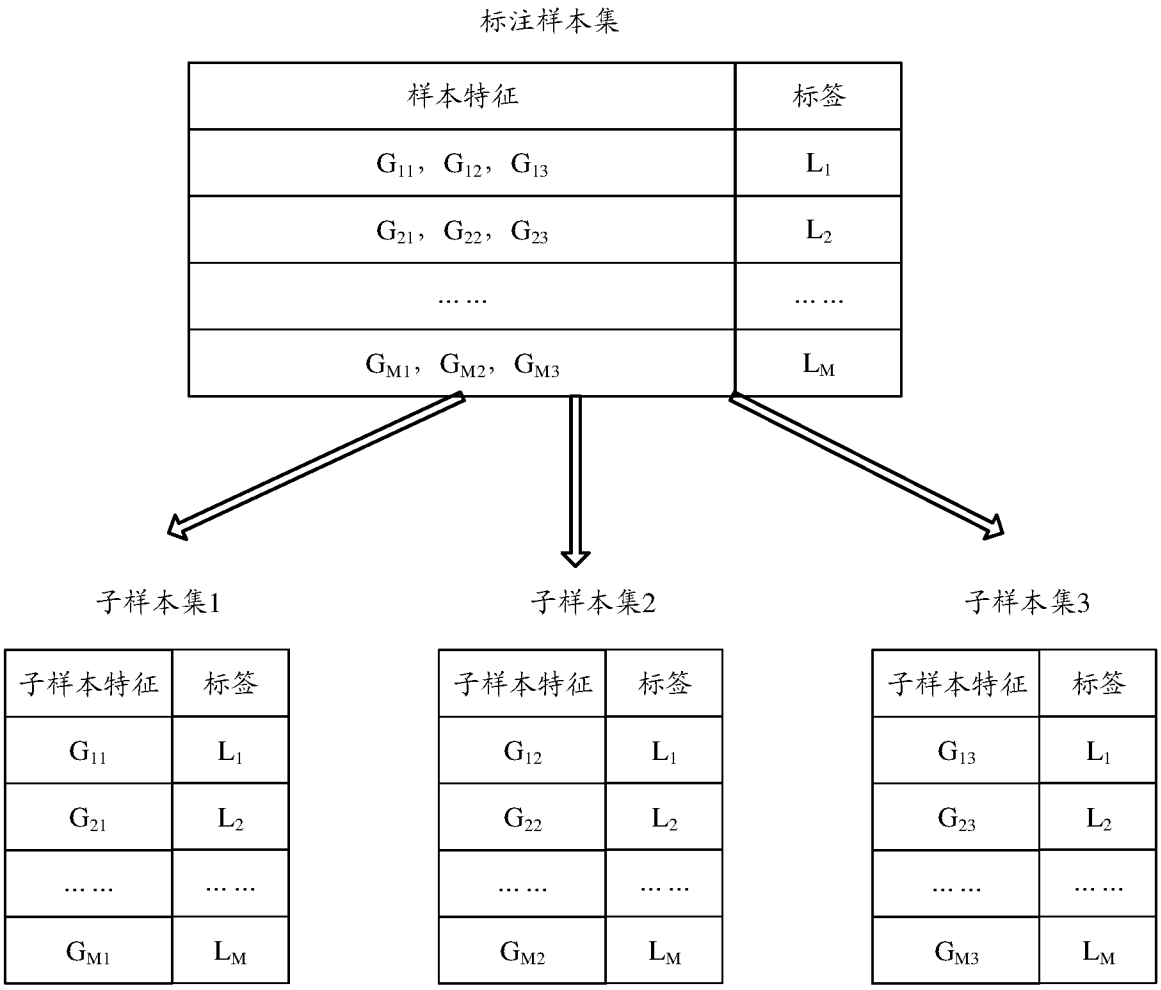
标注样本集
样本特征
标签
G_{11}，G_{12}，G_{13}
L_1
G_{21}，G_{22}，G_{23}
L_2
……
……
G_{M1}，G_{M2}，G_{M3}
L_M
子样本集1
子样本集2
子样本集3
子样本特征
标签
G_{11}
L_1
G_{21}
L_2
……
……
G_{M1}
L_M
子样本特征
标签
G_{12}
L_1
G_{22}
L_2
……
……
G_{M2}
L_M
子样本特征
标签
G_{13}
L_1
G_{23}
L_2
……
……
G_{M3}
L_M

图 3

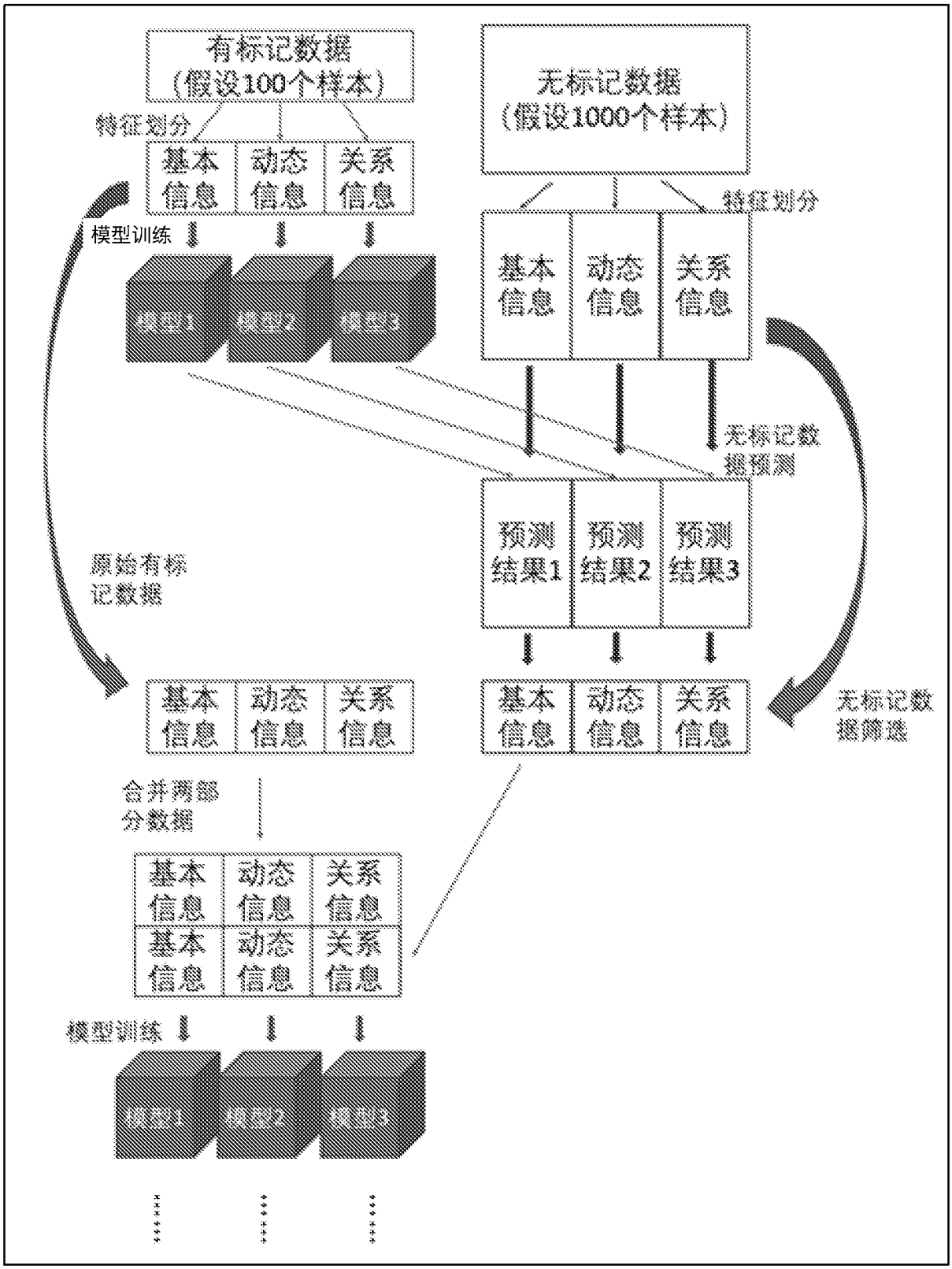
有标记数据
（假设100个样本）
特征划分
基本信息
动态信息
关系信息
模型训练
模型1
模型2
模型3
无标记数据
（假设1000个样本）
特征划分
基本信息
动态信息
关系信息
无标记数据预测
预测结果1
预测结果2
预测结果3
原始有标记数据
基本信息
动态信息
关系信息
无标记数据筛选
合并两部分数据
基本信息
动态信息
关系信息
基本信息
动态信息
关系信息
模型训练
模型1
模型2
模型3

图 4

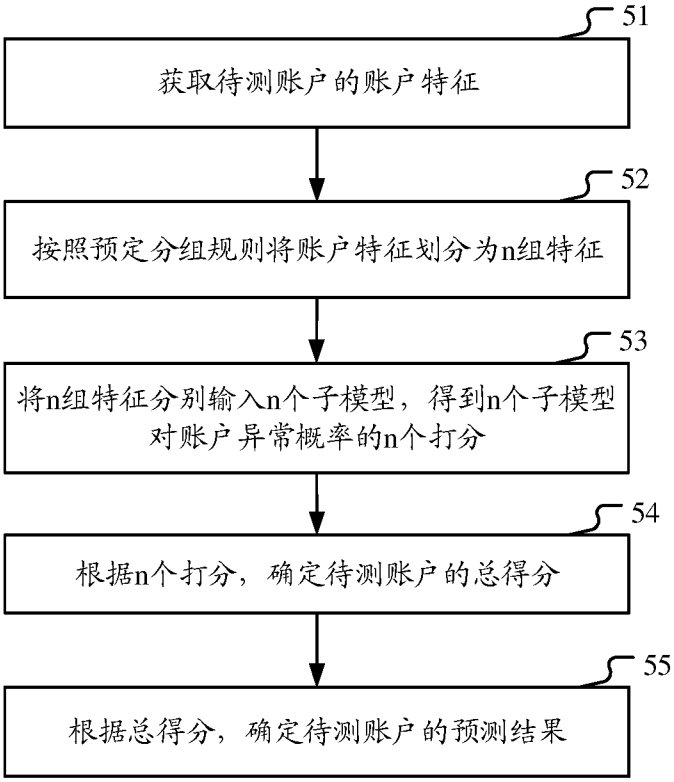
51
获取待测账户的账户特征
52
按照预定分组规则将账户特征划分为n组特征
53
将n组特征分别输入n个子模型，得到n个子模型对账户异常概率的n个打分
54
根据n个打分，确定待测账户的总得分
55
根据总得分，确定待测账户的预测结果

图 5

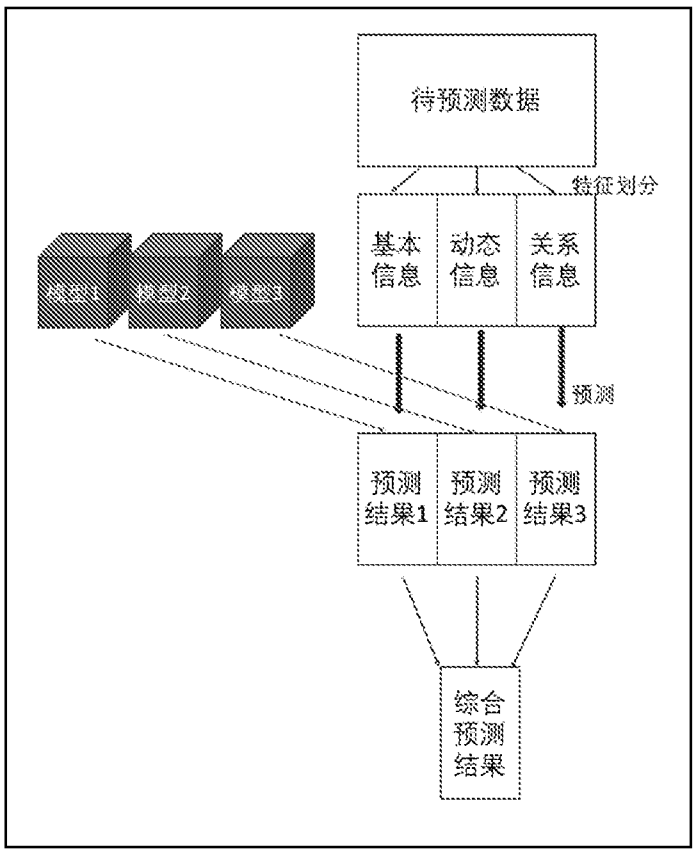
待预测数据
特征划分
基本信息
动态信息
关系信息
预测
预测结果1
预测结果2
预测结果3
综合预测结果

图 6

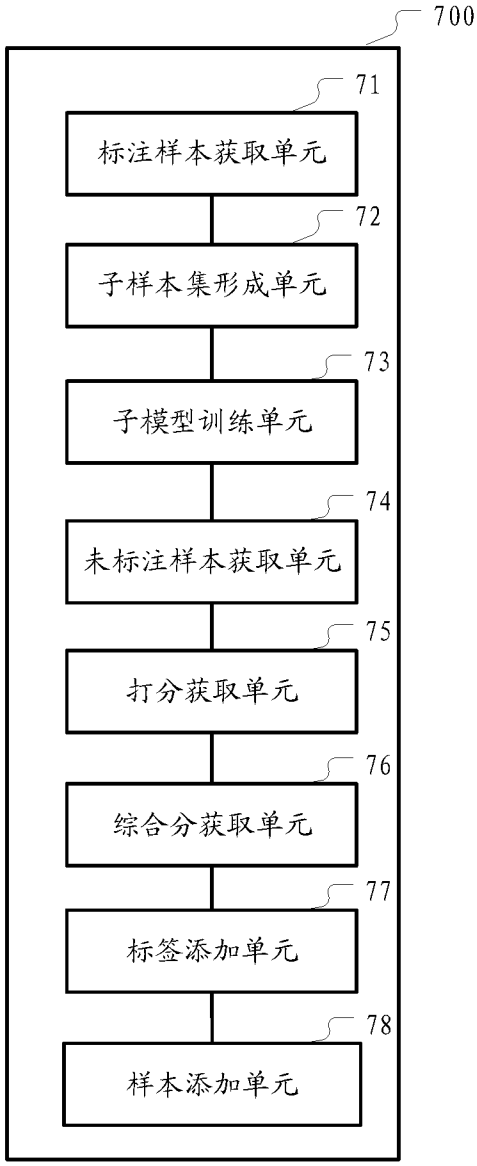
700
71
标注样本获取单元
72
子样本集形成单元
73
子模型训练单元
74
未标注样本获取单元
75
打分获取单元
76
综合分获取单元
77
标签添加单元
78
样本添加单元

图 7

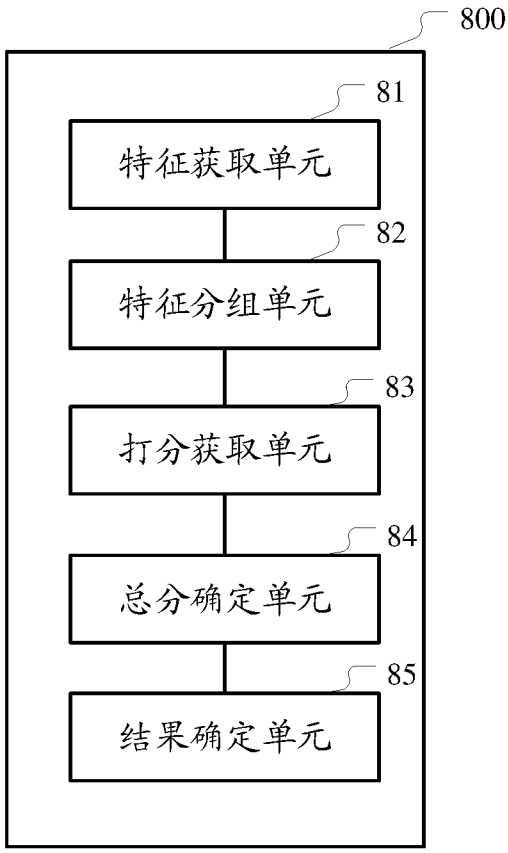

图 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/097090

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62(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)

G06K; G06F; G06Q

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, EPODOC, WPI: 样本, 账号, 账户, 交易, 标注, 标定, 标签, 特征, 分组, 组, 模型, 降维, 打分, 异常, 非法, training sample, sample, account, transaction, label, mark, feature, character, set, group, model, Dimensionality reduction, score, abnormal, illegal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109583468 A (ALIBABA GROUP HOLDING LIMITED) 05 April 2019 (2019-04-05) claims 1-21, and description, paragraphs [0068]-[0164]	1-21
X	CN 108595495 A (ALIBABA GROUP HOLDING LIMITED) 28 September 2018 (2018-09-28) description, paragraphs [0038]-[0084]	1-21
A	CN 106709513 A (ZHONGTAI SECURITIES STOCK CO., LTD.) 24 May 2017 (2017-05-24) entire document	1-21
A	CN 106294590 A (CHONGQING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 04 January 2017 (2017-01-04) entire document	1-21
A	US 2016314174 A1 (CHINA UNIONPAY CO., LTD.) 27 October 2016 (2016-10-27) entire document	1-21

☐ 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	Date of mailing of the international search report
18 September 2019	**26 September 2019**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/097090

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109583468	A	05 April 2019	None			
CN	108595495	A	28 September 2018	None			
CN	106709513	A	24 May 2017	None			
CN	106294590	A	04 January 2017	None			
US	2016314174	A1	27 October 2016	CN	104699717	A	10 June 2015
				WO	2015085916	A1	18 June 2015
				EP	3082051	A1	19 October 2016

国际检索报告

国际申请号
PCT/CN2019/097090

A. 主题的分类

G06K 9/62(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)

G06K; G06F; G06Q

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, EPODOC, WPI: 样本，账号，账户，交易，标注，标定，标签，特征，分组，组，模型，降维，打分，异常，非法，training sample, sample, account, transaction, label, mark, feature, character, set, group, model, Dimensionality reduction, score, abnormal, illegal

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 109583468 A (阿里巴巴集团控股有限公司) 2019年 4月 5日 (2019-04-05) 权利要求1-21，说明书第[0068]-[0164]段	1-21
X	CN 108595495 A (阿里巴巴集团控股有限公司) 2018年 9月 28日 (2018-09-28) 说明书第[0038]-[0084]段	1-21
A	CN 106709513 A (中泰证券股份有限公司) 2017年 5月 24日 (2017-05-24) 全文	1-21
A	CN 106294590 A (重庆邮电大学) 2017年 1月 4日 (2017-01-04) 全文	1-21
A	US 2016314174 A1 (CHINA UNIONPAY CO., LTD.) 2016年 10月 27日 (2016-10-27) 全文	1-21

☐ 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	国际检索报告邮寄日期
2019年 9月 18日	2019年 9月 26日

ISA/CN的名称和邮寄地址	受权官员
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	魏玲
传真号 (86-10)62019451	电话号码 86-(10)-53961737

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/097090

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	109583468	A	2019年 4月 5日	无			
CN	108595495	A	2018年 9月 28日	无			
CN	106709513	A	2017年 5月 24日	无			
CN	106294590	A	2017年 1月 4日	无			
US	2016314174	A1	2016年 10月 27日	CN	104699717	A	2015年 6月 10日
				WO	2015085916	A1	2015年 6月 18日
				EP	3082051	A1	2016年 10月 19日