

(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 5 月 22 日 (22.05.2020)



(10) 国际公布号
WO 2020/098308 A1

(51) 国际专利分类号:
G06K 9/62 (2006.01)

(21) 国际申请号: PCT/CN2019/097892

(22) 国际申请日: 2019 年 7 月 26 日 (26.07.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811340717.5 2018 年 11 月 12 日 (12.11.2018) 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 518000 (CN)。

(72) 发明人: 金戈(JIN, Ge); 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 518000 (CN)。 徐亮(XU, Liang); 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 518000 (CN)。 肖京(XIAO, Jing); 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 518000 (CN)。

(74) 代理人: 北京中强智尚知识产权代理有限公司 (BEIJING STRONG WISDOM INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市朝阳区北苑东路 19 号院 3 号楼 2501 室, Beijing 100012 (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,

(54) **Title:** METHOD, DEVICE AND EQUIPMENT FOR ESTABLISHING CROWD PORTRAIT CLASSIFICATION MEDEL AND STORAGE MEDIUM

(54) 发明名称: 人群画像分类模型的建立方法、装置、设备和存储介质

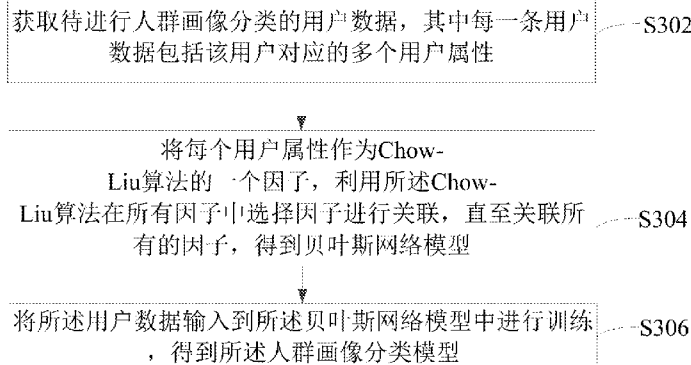


图 3

S302 Acquiring user data to be performed crowd portrait classification thereon, each piece of user data comprising a plurality of user attributes corresponding to the user

S304 Taking each user attribute as a factor of Chow-liu algorithm, and using the Chow-liu algorithm to select a factor among all factors to establish association until all factors are associated to obtain a Bayesian network model

S306 Inputting all user data into the Bayesian network model for training to obtain the crowd portrait classification model

(57) **Abstract:** The present invention relates to a method, device and computer equipment for establishing a crowd portrait classification model and a storage medium. The method comprises: acquiring user data to be performed crowd portrait classification thereon, each piece of user data comprising a plurality of user attributes corresponding to the user; taking each user attribute as a factor of Chow-liu algorithm, and using the Chow-liu algorithm to select a factor among all factors to establish association until all factors are associated to obtain a Bayesian network model; and inputting all user data into the Bayesian network model for training to obtain the crowd portrait classification model. The crowd portrait classification model according to the described method has a better explanatory feature, and can well reflect the correlation between respective user attributes of the user data.



WO 2020/098308 A1



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)。

根据细则4.17的声明:

- 关于申请人有权申请并被授予专利 (细则4.17(ii))
- 发明人资格 (细则4.17(iv))

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 本申请涉及一种人群画像分类模型的建立方法、装置、计算机设备和存储介质, 所述方法包括: 获取待进行人群画像分类的用户数据, 其中每一条用户数据包括该用户对应的多个用户属性; 将每个用户属性作为Chow-Liu算法的一个因子, 利用所述Chow-Liu算法在所有因子中选择因子进行关联, 直至关联所有的因子, 得到贝叶斯网络模型; 将所述用户数据输入到所述贝叶斯网络模型中进行训练, 得到所述人群画像分类模型。上述方法构建出的人群画像分类模型的解释性较佳, 能够很好地反映出用户数据的各个用户属性之间的相关性。

人群画像分类模型的建立方法、装置、设备和存储介质

本申请要求与 2018 年 11 月 12 日提交中国专利局、申请号为 201811340717.5、申请名称为“人群画像分类模型的建立方法、装置、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申请中。

5

技术领域

本申请涉及数据处理技术领域，特别是涉及的人群画像分类模型的建立方法、装置、计算机设备和存储介质。

背景技术

10 人群画像分类是指通过人群画像分类模型，对新输入的用户数据进行人群画像分类的过程。其中，人群画像分类模型是利用预设模型对海量的用户数据进行训练而构建的。

以员工画像分类为例，员工数据包括：员工的职位、工龄、教育、性别、部门等多个员工属性。利用预设模型对海量的员工数据进行训练，构建出员工画像分类模型，再通过员工画像分类模型得到若干的员工画像，从而完成对各个员工的分类。在员工画像分类的过程中，可以利用员工的离职情况构建员工离职
15 预测模型，进而通过员工离职预测模型，预测某个员工离职的概率。

目前，构建人群画像分类模型所基于的预设模型主要是分类模型和聚类模型，例如 SVM，神经网络，k-means 等。然而，在对现有技术的研究与实践过程中，本申请的发明人发现，现有技术存在以下问题：无论是基于分类模型还是聚类模型构建人群画像分类模型，构建得到的人群画像分类模型仅能够用于分类，其解释性较差，不能很好地反映出用户数据的各个用户属性之间的相关性以及用户数据的各个用户属性与类别归属的关联。
20

发明内容

基于此，有必要针对目前构建出的人群画像分类模型的解释性较差，不能很好地反映出用户数据的各个用户属性之间的相关性的问题，提供一种人群画像分类模型的建立方法、装置、计算机设备和存储介质。

一种人群画像分类模型的建立方法，所述人群画像分类模型的建立方法包括：获取待进行人群画像
25 分类的用户数据，其中每一条用户数据包括该用户对应的多个用户属性；将每个用户属性作为 Chow-Liu 算法的一个因子，利用所述 Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子，得到贝叶斯网络模型；将所述用户数据输入到所述贝叶斯网络模型中进行训练，得到所述人群画像分类模型。

在其中一个实施例中，所述方法还包括：对所述用户数据进行数据预处理。

在其中一个实施例中，所述数据预处理包括：数据清洗以及标准化处理；所述数据清洗包括：删除用户数据中的空缺数据、噪声数据、重复数据以及错误数据；所述标准化处理包括：将同一个用户对应的多个数据进行整合。

在其中一个实施例中，所述利用所述 Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子，包括：对于每一个因子，根据公式一在所有未选取的因子中选择与其 KL 距离最小的因子作为该因子的关联因子，直至所有因子均被选取；

$$\text{所述公式一为： } KL(P(X) || T(X)) = -\sum I(X_i, Pa(X_i)) + \sum H(X_i) - H(X_1, X_2, \dots, X_n)$$

其中， $KL(P(X) || T(X))$ 表示该因子与所有未选择的因子中任一因子的 KL 距离， $P(X)$ 表示进行关联之前所有因子的分布情况， $T(X)$ 表示进行关联之后所有因子的分布情况； X_i 表示第 i 个因子， H 表示熵， $Pa(X_i)$ 表示 X_i 的父节点； I 表示互信息，是通过公式二计算得到的，所述公式二为：

$$I(X_1, X_2) = \sum_{a \in X_1} \sum_{b \in X_2} P(a, b) \log \frac{p(a, b)}{p(a)p(b)}$$

其中， $p(a)$ 表示数值 a 出现的概率， $p(b)$ 表示数值 b 出现的概率， $p(a, b)$ 表示数值 b 出现的前提下数值 a 出现的概率， X_1 和 X_2 代表所述多个用户属性中任两个用户属性，数值 a 为属于用户属性 X_1 的任一数值，数值 b 为属于用户属性 X_2 的任一数值。

在其中一个实施例中，所述用户数据中包括标签数据以及未标签数据，所述将所述用户数据输入到所述贝叶斯网络模型中进行训练的步骤包括：采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练。

在其中一个实施例中，所述采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练，包括：利用所述贝叶斯网络模型对未标签数据进行标签预测；利用所述贝叶斯网络模型对标签数据进行训练；重复交替执行上述两个步骤，直至训练过程收敛。

在其中一个实施例中，所述方法还包括：在接收到新输入的用户数据时，利用所述人群画像分类模型对所述用户数据进行人群画像分类，得到对应的分类结果。

一种人群画像分类模型的建立装置，所述人群画像分类模型的建立装置包括：数据获取单元，用于获取待进行人群画像分类的用户数据，其中每一条用户数据包括该用户对应的多个用户属性；因子关联单元，用于将每个用户属性作为 Chow-Liu 算法的一个因子，利用所述 Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子，得到贝叶斯网络模型；数据训练单元，用于将所述用户数据输入到所述贝叶斯网络模型中进行训练，得到所述人群画像分类模型。

在其中一个实施例中，所述装置还包括：包括预处理单元 802，用于对所述用户数据进行数据预处理。

在其中一个实施例中，当数据预处理包括：数据清洗以及标准化处理时，所述预处理单元包括：数据清洗模块和标准化处理模块。所述数据清洗模块，用于删除用户数据中的空缺数据、噪声数据、重复数据以及错误数据；所述标准化处理模块，用于将同一个用户对应的多个数据进行整合。

在其中一个实施例中，因子关联单元 704 具体用于执行以下步骤：

对于每一个因子，根据公式一在所有未选取的因子中选择与其 KL 距离最小的因子作为该因子的关联因子，直至所有因子均被选取；

所述公式一为： $KL(P(X) || T(X)) = -\sum I(X_i, Pa(X_i)) + \sum H(X_i) - H(X_1, X_2, \dots, X_n)$

其中， $KL(P(X) || T(X))$ 表示该因子与所有未选择的因子中任一因子的 KL 距离， $P(X)$ 表示进行关联之前所有因子的分布情况， $T(X)$ 表示进行关联之后所有因子的分布情况； X_i 表示第 i 个因子， H 表示熵， $Pa(X_i)$ 表示 X_i 的父节点； I 表示互信息，是通过公式二计算得到的，所述公式二为：

$$I(X_1, X_2) = \sum_{a \in X_1} \sum_{b \in X_2} P(a, b) \log \frac{p(a, b)}{p(a)p(b)}$$

其中， $p(a)$ 表示数值 a 出现的概率， $p(b)$ 表示数值 b 出现的概率， $p(a, b)$ 表示数值 b 出现的前提下数值 a 出现的概率， X_1 和 X_2 代表所述多个用户属性中任两个用户属性，数值 a 为属于用户属性 X_1 的任一数值，数值 b 为属于用户属性 X_2 的任一数值。

在其中一个实施例中，当用户数据中包括标签数据以及未标签数据时，所述数据训练单元具体用于采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练。

在其中一个实施例中，当用户数据中包括标签数据以及未标签数据时，数据训练单元具体用于执行以下步骤：利用所述贝叶斯网络模型对未标签数据进行标签预测；利用所述贝叶斯网络模型对标签数据进行训练；重复交替执行上述两个步骤，直至训练过程收敛。

在其中一个实施例中，人群画像分类模型的建立装置还可以包括分类单元，用于在接收到新输入的用户数据时，利用所述人群画像分类模型对所述用户数据进行人群画像分类，得到对应的分类结果。

一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述计算机可读指令被所述处理器执行时，使得所述处理器执行上述人群画像分类模型的建立方法的步骤。

一种存储有计算机可读指令的非易失性可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行上述人群画像分类模型的建立方法的步骤。

上述人群画像分类模型的建立方法、装置、计算机设备和存储介质，获取待进行人群画像分类的用户数据，其中每一条用户数据包括该用户对应的多个用户属性；将每个用户属性作为 Chow-Liu 算法的一个因子，利用 Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子，得到贝叶斯网络模型；将用户数据输入到贝叶斯网络模型中进行训练，得到人群画像分类模型。上述人群画像分类模型的建立方法，将用户数据包括的多个用户属性中每个用户属性作为 Chow-Liu 算法的一个因子，利用 Chow-Liu 算法进行因子选择以及关联，由于 Chow-Liu 算法能够较好反映地各个因子之间的关联，同时能够反映因子与类别归属的关联，所以基于 Chow-Liu 算法构建的人群画像分类模型能够很好地反映出用户数据的各个用户属性之间的相关性，同时能够反映用户数据的各个用户属性与类别归属的关联。

附图说明

10 图 1 为一个实施例中提供的人群画像分类模型的建立方法的实施环境图；

图 2 为一个实施例中计算机设备的内部结构框图；

图 3 为一个实施例中人群画像分类模型的建立方法的流程图；

图 4 为一个实施例中人群画像分类模型的建立方法的流程图；

图 5 为一个实施例中人群画像分类的方法的流程图；

15 图 6 为一个实施例中人群画像分类模型的建立方法的流程图；

图 7 为一个实施例中人群画像分类模型的建立装置的结构框图；

图 8 为一个实施例中人群画像分类模型的建立装置的结构框图；

图 9 为一个实施例中预处理单元的结构框图；

图 10 为一个实施例中人群画像分类模型的建立装置的结构框图。

20 具体实施方式

为了使本申请的目的、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，此处所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。

图 1 为一个实施例中提供的人群画像分类模型的建立方法的实施环境图，如图 1 所示，在该实施环境中，包括计算机设备 110 以及数据库 120。数据库 120 中存储有待进行人群画像分类的用户数据以及新输入的用户数据。在预先对待进行人群画像分类的用户数据中的部分用户数据打标签的情况下，数据库 120 中存储的待进行人群画像分类的用户数据包括标签数据和未标签数据。计算机设备 110 为用于对用户数据进行处理以建立人群画像分类模型的设备，计算机设备 110 从数据库 120 中获取待进行人群画像分类的用户数据以及新输入的用户数据。在数据库 120 中存储的待进行人群画像分类的用户数据包括标签数据和未

标签数据的情况下，计算机设备 110 从数据库 120 中获取标签数据和未标签数据。当需要建立人群画像分类模型时，模型建立人员可以利用计算机设备 110 获取待进行人群画像分类的用户数据，然后根据该用户数据包括的多个用户属性，得到贝叶斯网络模型，然后将待进行人群画像分类的用户数据输入该贝叶斯网络模型进行训练，得到人群画像分类模型。

5 需要说明的是，计算机设备 110 和数据库 120 均分别可为智能手机、平板电脑、笔记本电脑、台式计算机、服务器等，但并不局限于此。计算机设备 110 与数据库 120 可以通过蓝牙、USB(Universal Serial Bus, 通用串行总线)或者其他通讯连接方式进行连接，本申请在此不做限制。数据库 120 可以独立于计算机设备 110 (如图 1 所示)，或者，数据库 120 可以集成与计算机设备 110 内部(图 1 未示出)。

图 2 为一个实施例中计算机设备的内部结构示意图。如图 2 所示，该计算机设备包括通过系统总线
10 连接的处理器、非易失性可读存储介质、存储器和网络接口。其中，该计算机设备的非易失性可读存储介质存储有操作系统、数据库和计算机可读指令，数据库中可存储有待进行人群画像分类的用户数据以及新输入的用户数据。在预先对待进行人群画像分类的用户数据中的部分用户数据打标签的情况下，数据库中存储的待进行人群画像分类的用户数据包括标签数据和未标签数据。该计算机可读指令被处理器执行时，可使得处理器实现一种人群画像分类模型的建立方法。该计算机设备的处理器用于提供计算和控制能力，
15 支撑整个计算机设备的运行。该计算机设备的存储器中可存储有计算机可读指令，该计算机可读指令被处理器执行时，可使得处理器执行一种人群画像分类模型的建立方法。该计算机设备的网络接口用于与外部通信连接。本领域技术人员可以理解，图 2 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。

20 如图 3 所示，在一个实施例中，提出了一种人群画像分类模型的建立方法，该人群画像分类模型的建立方法可以应用于上述的计算机设备 110 中，包括以下步骤：

步骤 S302，获取待进行人群画像分类的用户数据，其中每一条用户数据包括该用户对应的多个用户属性；

在本实施例中，用户数据包括用户对应的多个用户属性。以用户数据是员工数据为例，员工数据包
25 括员工对应的多个员工属性：职位、工龄、教育、性别、部门等。针对多个用户属性中的每个用户属性，属于该用户属性的数值有若干个，若干个数值与若干条用户数据一一对应，也即与若干个用户一一对应。示例地，以用户数据是员工数据为例，员工数据包括员工对应的多个员工属性：职位、工龄、教育、性别、部门等，假设有员工 A 和员工 B，员工 A 和员工 B 的员工数据分别如表 1 所示。

表 1 员工数据示意图

员工属性 员工	职位	工龄/年	教育	性别	部门
员工 A	总经理 (编号 0011)	5	研究生 (编号 01)	男 (编号 0)	管理部 (编号 001)
员工 B	工程师 (编号 0012)	2	本科 (编号 02)	女 (编号 1)	工程部 (编号 002)

从表 1 可以看出,属于职位这一用户属性的数值有 2 个: 0011 和 0012,分别与员工 A 的职位以及员工 B 的职位一一对应,也即分别与员工 A 和员工 B 一一对应。同理,属于工龄这一用户属性的数值有 2 个: 5 和 2,分别与员工 A 的工龄以及员工 B 的工龄一一对应,也即分别与员工 A 和员工 B 一一对应。待

5 进行人群画像分类的用户数据是即将输入预设模型的样本数据,该预设模型为根据多个用户属性得到的贝叶斯网络模型。

步骤 S304,将每个用户属性作为 Chow-Liu 算法的一个因子,利用所述 Chow-Liu 算法在所有因子中选择因子进行关联,直至关联所有的因子,得到贝叶斯网络模型;

在本实施例中,考虑到 Chow-Liu 算法能够较好地反映各个因子之间的关联,同时能够反映因子与类别归属的关联。进而在进行预测的同时,能够从模型中归纳各个因子对于预测结果的影响。示例地,以对

10 员工离职的概率进行预测为例,通过利用 Chow-Liu 算法,可以预测某个员工离职的概率,同时发现哪个员工属性是导致高离职率的影响因素,从而为后续的员工招聘工作提供参考。在具体实施过程中,执行多次关联。第一次关联的过程为:首先,将多个用户属性中的任何一个用户属性作为 Chow-Liu 算法的第一个因子,然后,将多个用户属性中剩余的各个用户属性作为 Chow-Liu 算法的其他各个因子,从这些其他

15 各个因子中选择一个因子与第一个因子关联。第二次关联的过程为:将上一次关联过程中剩余的各个用户属性中的任何一个用户属性作为 Chow-Liu 算法的第二个因子,将本次关联过程中剩余的各个用户属性作为 Chow-Liu 算法的其他各个因子,从这些其他各个因子中选择一个因子与第二个因子关联。第三次关联过程与第二次关联过程类似。如此进行多次关联,直至关联所有的因子。需要注意的是,每次关联过程中需要避免产生回路。在一种实施方式中,步骤 304 包括:对于每一个因子,根据公式一在所有未选取的因

20 子中选择与其 KL 距离最小的因子作为该因子的关联因子,直至所有因子均被选取;所述公式一为:

$$KL(P(X) || T(X)) = -\sum I(X_i, Pa(X_i)) + \sum H(X_i) - H(X_1, X_2, \dots, X_n)$$

其中, $KL(P(X) || T(X))$ 表示该因子与所有未选择的因子中任一因子的 KL 距离, $P(X)$ 表示进行关联之前所有因子的分布情况, $T(X)$ 表示进行关联之后所有因子的分布情况; X_i 表示第 i 个因子, H 表示熵, $Pa(X_i)$ 表示 X_i 的父节点; I 表示互信息, 是通过公式二计算得到的, 所述公式二为:

$$I(X_1, X_2) = \sum_{a \in X_1} \sum_{b \in X_2} P(a, b) \log \frac{p(a, b)}{p(a)p(b)}$$

5 其中, $p(a)$ 表示数值 a 出现的概率, $p(b)$ 表示数值 b 出现的概率, $p(a, b)$ 表示数值 b 出现的前提下数值 a 出现的概率, X_1 和 X_2 代表所述多个用户属性中任两个用户属性, 数值 a 为属于用户属性 X_1 的任一数值, 数值 b 为属于用户属性 X_2 的任一数值。在具体实施过程中, 从多个因子中选择一个因子与另一个因子关联, 包括以下步骤: 步骤 1): 根据公式一和公式二, 计算多个因子中每个因子与该另一个因子的 KL 距离, 公式一和公式二分别为:

$$10 \quad KL(P(X) || T(X)) = -\sum I(X_i, Pa(X_i)) + \sum H(X_i) - H(X_1, X_2, \dots, X_n) \quad \text{公式一}$$

$$I(X_1, X_2) = \sum_{a \in X_1} \sum_{b \in X_2} P(a, b) \log \frac{p(a, b)}{p(a)p(b)} \quad \text{公式二}$$

步骤 2): 从多个因子中确定出与该另一个因子的 KL 距离最小的因子, 第一次确定过程的标准为 KL 距离最小, 第二次确定过程的标准为 KL 距离次小, 第三次确定过程的标准为 KL 距离第三小, 依次类推; 步骤 3): 判断如果确定出的因子与该另一个因子关联, 是否会产生回路, 若判断结果为否, 则转入步骤 4); 15 若判断结果为是, 则返回步骤 2); 步骤 4): 将确定出的因子与该另一个因子关联。

步骤 S306, 将所述用户数据输入到所述贝叶斯网络模型中进行训练, 得到所述人群画像分类模型。

在根据多个用户属性得到的贝叶斯网络模型, 再将待进行人群画像分类的用户数据输入贝叶斯网络模型进行训练, 得到人群画像分类模型。示例地, 以用户数据是员工数据且构建员工离职预测模型为例, 针对待进行离职概率预测的员工数据, 首先, 根据员工数据包括员工对应的多个员工属性: 职位、工龄、20 教育、性别、部门等, 得到贝叶斯网络模型; 然后, 将员工数据输入到得到的贝叶斯网络模型进行训练, 得到员工离职预测模型。上述人群画像分类模型的建立方法, 将用户数据包括的多个用户属性中每个用户属性作为 Chow-Liu 算法的一个因子, 利用 Chow-Liu 算法进行因子选择以及关联, 由于 Chow-Liu 算法能够较好反映地各个因子之间的关联, 同时能够反映因子与类别归属的关联, 所以基于 Chow-Liu 算法构建的人群画像分类模型能够很好地反映出用户数据的各个用户属性之间的相关性, 同时能够反映用户数据的25 各个用户属性与类别归属的关联。

图 4 示出了在一个实施例中, 当用户数据中包括标签数据以及未标签数据时, 人群画像分类模型的

建立方法的实现流程图，包括以下步骤：步骤 S402，获取待进行人群画像分类的用户数据，其中每一条用户数据包括该用户对应的多个用户属性；步骤 S404，将每个用户属性作为 Chow-Liu 算法的一个因子，利用所述 Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子，得到贝叶斯网络模型；步骤 S406，采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练，得到所述人群画像分类模型。其中，步骤 S402 以及步骤 S404 的实现过程分别与步骤 S302 以及步骤 S304 的实现过程类似，在此就不再赘述。为了提升得到的人群画像分类模型的精度，可以预先对部分待进行人群画像分类的用户数据打标签，因而待进行人群画像分类的用户数据包括标签数据以及未标签数据，然后将标签数据和未标签数据输入贝叶斯网络模型进行训练，得到精度更高的人群画像分类模型。示例地，以用户数据是员工数据且构建员工离职预测模型为例，针对待进行离职概率预测的员工数据，一方面，根据员工数据包括员工对应的多个员工属性：职位、工龄、教育、性别、部门等，得到贝叶斯网络模型；另一方面，对部分员工数据打标签，根据这部分员工的实际离职情况，对这部分员工数据打标签：已离职员工和为离职员工。然后，将打标签的员工数据（已知离职情况）和未打标签的员工数据（未知离职情况）输入得到的贝叶斯网络模型进行训练，得到员工离职预测模型。

在一种实施方式中，步骤 S406 包括以下步骤：利用所述贝叶斯网络模型对未标签数据进行标签预测；利用所述贝叶斯网络模型对标签数据进行训练；重复交替执行上述两个步骤，直至训练过程收敛。在具体实施过程中，半监督学习方法包括 E-step 和 M-step。用户数据包括标签数据以及未标签数据。将用户数据输入到执行步骤 S404 后得到的贝叶斯网络模型中进行训练的过程如下：首先，进行 E-step，即利用执行步骤 S404 后得到的贝叶斯网络模型对未标签数据进行标签预测。随后，进行 M-step，即利用标签数据，重新训练贝叶斯网络模型，并交替重复 E-step 与 M-step，直至训练过程收敛，最终得到人群画像分类模型。

在本实施例中，当获取的待进行人群画像分类的用户数据包括未标签数据时，根据多个用户属性得到的贝叶斯网络模型同样能够将此类未标签数据用于训练，也就是说，除将标签数据作为训练数据外，可以加入未标签数据作为训练数据，以避免训练数据量过低的问题，从而提升最终得到的人群画像分类模型的精度。图 5 示出了一个实施例中人群画像分类的方法的实现流程图，包括以下步骤：步骤 S502，获取待进行人群画像分类的用户数据，其中每一条用户数据包括该用户对应的多个用户属性；步骤 S504，将每个用户属性作为 Chow-Liu 算法的一个因子，利用所述 Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子，得到贝叶斯网络模型；步骤 S506，将所述用户数据输入到所述贝叶斯网络模型中进行训练，得到所述人群画像分类模型；步骤 S508，在接收到新输入的用户数据时，利用所述人群画像分类模型对所述用户数据进行人群画像分类，得到对应的分类结果。其中，步骤 S502-步骤 S506 的实现过程分别与

步骤 S302-步骤 S306 的实现过程类似，在此就不再赘述。在得到人群画像分类模型之后，即可利用该人群画像分类模型实现人群画像分类。具体地，接收新输入的用户数据，然后将该用户数据输入执行步骤 S502-步骤 S506 后所得到的人群画像分类模型，该人群画像分类模型的输出即为分类结果。示例地，以用户数据是员工数据且人群画像分类模型是员工离职预测模型为例，将一个某个员工的员工数据输入该员工离职预测模型，通过该员工离职预测模型即可预测该员工离职的概率，也即是说，该员工离职预测模型的输出即为该员工离职的概率。

图 6 示出了一个实施例中人群画像分类模型的建立方法的实现流程图，包括以下步骤：步骤 S602，获取待进行人群画像分类的用户数据，其中每一条用户数据包括该用户对应的多个用户属性；步骤 S604，对所述用户数据进行数据预处理；步骤 S606，将每个用户属性作为 Chow-Liu 算法的一个因子，利用所述 Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子，得到贝叶斯网络模型；步骤 S608，将所述用户数据输入到所述贝叶斯网络模型中进行训练，得到所述人群画像分类模型。其中，步骤 S602 的实现过程与步骤 S302 的实现过程类似，在此就不再赘述。在一种实施方式中，步骤 S604 中的数据预处理包括：数据清洗以及标准化处理；所述数据清洗包括：删除用户数据中的空缺数据、噪声数据、重复数据以及错误数据；所述标准化处理包括：将同一个用户对应的多个数据进行整合。在本实施例中，考虑到执行步骤 S602 后原始获取的用户数据存在“脏数据”，包括数据空缺和噪声、不一致、重复、错误等问题，为了保证后期数据处理的准确性，以及，在利用人群画像分类模型得到分类结果后，为了减少该分类结果对最终决策造成的影响，有必要对原始获取的用户数据进行预处理。即删除原始获取的用户数据中的空缺数据、噪声数据、重复数据以及错误数据。此外，人群画像的建立需要有整合多源数据的能力，例如一个用户可能使用多个设备，在网络上拥有多个账号。因此需要把同一用户的多个账号组合，即将同一个用户对应的多个数据进行整合，进而建立统一的标准，以完整标识用户的人群画像。在执行完步骤 S604 之后，步骤 S606 中将经过步骤 S604 中的预处理后的用户数据包括的每个用户属性作为 Chow-Liu 算法的一个因子，其余与步骤 S304 类似。同理，步骤 S606 中将经过步骤 S604 中的预处理后的用户数据，输入到执行步骤 S606 后所得到的贝叶斯网络模型中进行训练，其余与步骤 S306 类似。

如图 7 所示，在一个实施例中，提供了一种人群画像分类模型的建立装置，该人群画像分类模型的建立装置可以集成于上述的计算机设备 110 中，可以包括数据获取单元 702、因子关联单元 704、以及数据训练单元 706。数据获取单元 702，用于获取待进行人群画像分类的用户数据，其中每一条用户数据包括该用户对应的多个用户属性；因子关联单元 704，用于将每个用户属性作为 Chow-Liu 算法的一个因子，利用所述 Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子，得到贝叶斯网络模型；数

据训练单元 706, 用于将所述用户数据输入到所述贝叶斯网络模型中进行训练, 得到所述人群画像分类模型。如图 8 所示, 人群画像分类模型的建立装置还可以包括预处理单元 802。预处理单元 802, 用于对所述用户数据进行数据预处理。如图 9 所示, 在一个实施例中, 当数据预处理包括: 数据清洗以及标准化处理时, 预处理单元 802 包括: 数据清洗模块 802A 和标准化处理模块 802B。数据清洗模块 802A, 用于删

5 除用户数据中的空缺数据、噪声数据、重复数据以及错误数据; 标准化处理模块 802B, 用于将同一个用户对应的多个数据进行整合。在一个实施例中, 因子关联单元 704 具体用于执行以下步骤: 对于每一个因子, 根据公式一在所有未选取的因子中选择与其 KL 距离最小的因子作为该因子的关联因子, 直至所有因子均被选取; 所述公式一为: $KL(P(X) || T(X)) = -\sum I(X_i, Pa(X_i)) + \sum H(X_i) - H(X_1, X_2, \dots, X_n)$

其中, $KL(P(X) || T(X))$ 表示该因子与所有未选择的因子中任一因子的 KL 距离, $P(X)$ 表示进行
10 关联之前所有因子的分布情况, $T(X)$ 表示进行关联之后所有因子的分布情况; X_i 表示第 i 个因子, H 表示熵, $Pa(X_i)$ 表示 X_i 的父节点; I 表示互信息, 是通过公式二计算得到的, 所述公式二为:

$$I(X_1, X_2) = \sum_{a \in X_1} \sum_{b \in X_2} P(a, b) \log \frac{p(a, b)}{p(a)p(b)}$$

其中, $p(a)$ 表示数值 a 出现的概率, $p(b)$ 表示数值 b 出现的概率, $p(a, b)$ 表示数值 b 出现的前提
下数值 b 出现的概率, X_1 和 X_2 代表所述多个用户属性中任两个用户属性, 数值 a 为属于用户属性 X_1 的任
15 一数值, 数值 b 为属于用户属性 X_2 的任一数值。在一个实施例中, 当用户数据中包括标签数据以及未标签数据时, 数据训练单元 706 具体用于采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练。

在一个实施例中, 当用户数据中包括标签数据以及未标签数据时, 数据训练单元 706 具体用于执行以下步骤: 利用所述贝叶斯网络模型对未标签数据进行标签预测; 利用所述贝叶斯网络模型对标签数据进行训练; 重复交替执行上述两个步骤, 直至训练过程收敛。如图 10 所示, 人群画像分类模型的建立装置还可以包

20 括分类单元 1002。分类单元 1002, 用于在接收到新输入的用户数据时, 利用所述人群画像分类模型对所述用户数据进行人群画像分类, 得到对应的分类结果。在一个实施例中, 提出了一种计算机设备, 所述计算机设备包括非易失性可读存储介质、处理器及存储在所述非易失性可读存储介质上并可在所述处理器上运行的计算机可读指令, 所述处理器执行所述计算机可读指令时实现以下步骤: 获取待进行人群画像分类的用户数据, 其中每一条用户数据包括该用户对应的多个用户属性; 将每个用户属性作为 Chow-Liu 算法
25 的一个因子, 利用所述 Chow-Liu 算法在所有因子中选择因子进行关联, 直至关联所有的因子, 得到贝叶斯网络模型; 将所述用户数据输入到所述贝叶斯网络模型中进行训练, 得到所述人群画像分类模型。在一个实施例中, 处理器执行计算机可读指令时还执行以下步骤: 对所述用户数据进行数据预处理。在一个实

施例中, 所述数据预处理包括: 数据清洗以及标准化处理; 所述处理器所执行的对所述用户数据进行数据预处理的步骤包括: 删除用户数据中的空缺数据、噪声数据、重复数据以及错误数据; 将同一个用户对应的多个数据进行整合。在一个实施例中, 所述处理器所执行的利用所述 Chow-Liu 算法在所有因子中选择因子进行关联, 直至关联所有的因子的步骤包括: 对于每一个因子, 根据公式一在所有未选取的因子中选择与其 KL 距离最小的因子作为该因子的关联因子, 直至所有因子均被选取; 所述公式一为:

$$KL(P(X) || T(X)) = -\sum I(X_i, Pa(X_i)) + \sum H(X_i) - H(X_1, X_2, \dots, X_n)$$

其中, $KL(P(X) || T(X))$ 表示该因子与所有未选择的因子中任一因子的 KL 距离, $P(X)$ 表示进行关联之前所有因子的分布情况, $T(X)$ 表示进行关联之后所有因子的分布情况; X_i 表示第 i 个因子, H 表示熵, $Pa(X_i)$ 表示 X_i 的父节点; I 表示互信息, 是通过公式二计算得到的, 所述公式二为:

$$I(X_1, X_2) = \sum_{a \in X_1} \sum_{b \in X_2} P(a, b) \log \frac{p(a, b)}{p(a)p(b)}$$

其中, $p(a)$ 表示数值 a 出现的概率, $p(b)$ 表示数值 b 出现的概率, $p(a, b)$ 表示数值 b 出现的前提下数值 a 出现的概率, X_1 和 X_2 代表所述多个用户属性中任两个用户属性, 数值 a 为属于用户属性 X_1 的任一数值, 数值 b 为属于用户属性 X_2 的任一数值。在一个实施例中, 所述用户数据中包括标签数据以及未标签数据, 所述处理器所执行的将所述用户数据输入到所述贝叶斯网络模型中进行训练的步骤包括: 采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练。在一个实施例中, 所述处理器所执行的采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练的步骤包括: 利用所述贝叶斯网络模型对未标签数据进行标签预测; 利用所述贝叶斯网络模型对标签数据进行训练; 重复交替执行上述两个步骤, 直至训练过程收敛。在一个实施例中, 处理器执行计算机可读指令时还执行以下步骤: 在接收到新输入的用户数据时, 利用所述人群画像分类模型对所述用户数据进行人群画像分类, 得到对应的分类结果。

在一个实施例中, 提出了一种存储有计算机可读指令的非易失性可读存储介质, 该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执行以下步骤: 获取待进行人群画像分类的用户数据, 其中每一条用户数据包括该用户对应的多个用户属性; 将每个用户属性作为 Chow-Liu 算法的一个因子, 利用所述 Chow-Liu 算法在所有因子中选择因子进行关联, 直至关联所有的因子, 得到贝叶斯网络模型; 将所述用户数据输入到所述贝叶斯网络模型中进行训练, 得到所述人群画像分类模型。在一个实施例中, 处理器执行计算机可读指令时还执行以下步骤: 对所述用户数据进行数据预处理。在一个实施例中, 所述数据预处理包括: 数据清洗以及标准化处理; 所述处理器所执行的对所述用户数据进行数据预处理的步骤包括: 删除用户数据中的空缺数据、噪声数据、重复数据以及错误数据; 将同一个用户对应的多个数据进

行整合。在一个实施例中，所述处理器所执行的利用所述 Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子的步骤包括：对于每一个因子，根据公式一在所有未选取的因子中选择与其 KL 距离最小的因子作为该因子的关联因子，直至所有因子均被选取；所述公式一为：

$$KL(P(X) || T(X)) = -\sum I(X_i, Pa(X_i)) + \sum H(X_i) - H(X_1, X_2, \dots, X_n)$$

5 其中， $KL(P(X) || T(X))$ 表示该因子与所有未选择的因子中任一因子的 KL 距离， $P(X)$ 表示进行关联之前所有因子的分布情况， $T(X)$ 表示进行关联之后所有因子的分布情况； X_i 表示第 i 个因子， H 表示熵， $Pa(X_i)$ 表示 X_i 的父节点； I 表示互信息，是通过公式二计算得到的，所述公式二为：

$$I(X_1, X_2) = \sum_{a \in x_1} \sum_{b \in x_2} P(a, b) \log \frac{p(a, b)}{p(a)p(b)}$$

其中， $p(a)$ 表示数值 a 出现的概率， $p(b)$ 表示数值 b 出现的概率， $p(a, b)$ 表示数值 b 出现的前提下数值 b 出现的概率， X_1 和 X_2 代表所述多个用户属性中任两个用户属性，数值 a 为属于用户属性 X_1 的任一数值，数值 b 为属于用户属性 X_2 的任一数值。在一个实施例中，所述用户数据中包括标签数据以及未标签数据，所述处理器所执行的将所述用户数据输入到所述贝叶斯网络模型中进行训练的步骤包括：采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练。在一个实施例中，所述处理器所执行的采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练的步骤包括：利用所述贝叶斯网络模型对未标签数据进行标签预测；利用所述贝叶斯网络模型对标签数据进行训练；重复交替执行上述两个步骤，直至训练过程收敛。在一个实施例中，处理器执行计算机可读指令时还执行以下步骤：在接收到新输入的用户数据时，利用所述人群画像分类模型对所述用户数据进行人群画像分类，得到对应的分类结果。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，是可以通过计算机程序来指令相关的硬件来完成，该计算机程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，前述的存储介质可为磁碟、光盘、只读存储记忆体（Read-Only Memory, ROM）等非易失性存储介质，或随机存储记忆体（Random Access Memory, RAM）等。以上所述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。以上所述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本申请专利范围的限制。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、一种人群画像分类模型的建立方法，其特征在于，包括：

获取待进行人群画像分类的用户数据，其中每一条用户数据包括该用户对应的多个用户属性；

将每个用户属性作为 Chow-Liu 算法的一个因子，利用所述 Chow-Liu 算法在所有因子中选择因子进

5 行关联，直至关联所有的因子，得到贝叶斯网络模型；

将所述用户数据输入到所述贝叶斯网络模型中进行训练，得到所述人群画像分类模型。

2、根据权利要求 1 所述的方法，其特征在于，在获取待进行人群画像分类的用户数据的步骤之后，
所述方法还包括：对所述用户数据进行数据预处理。

3、根据权利要求 2 所述的方法，其特征在于，所述数据预处理包括：数据清洗以及标准化处理；

10 所述数据清洗包括：删除用户数据中的空缺数据、噪声数据、重复数据以及错误数据；

所述标准化处理包括：将同一个用户对应的多个数据进行整合。

4、根据权利要求 1 所述的方法，其特征在于，所述利用所述 Chow-Liu 算法在所有因子中选择因子
进行关联，直至关联所有的因子，包括：

对于每一个因子，根据公式一在所有未选取的因子中选择与其 KL 距离最小的因子作为该因子的关

15 联因子，直至所有因子均被选取；

所述公式一为： $KL(P(X) || T(X)) = -\sum I(X_i, Pa(X_i)) + \sum H(X_i) - H(X_1, X_2, \dots, X_n)$

其中， $KL(P(X) || T(X))$ 表示该因子与所有未选择的因子中任一因子的 KL 距离， $P(X)$ 表示进行
关联之前所有因子的分布情况， $T(X)$ 表示进行关联之后所有因子的分布情况； X_i 表示第 i 个因子，H 表
示熵， $Pa(X_i)$ 表示 X_i 的父节点；I 表示互信息，是通过公式二计算得到的，所述公式二为：

20
$$I(X_1, X_2) = \sum_{a \in X_1} \sum_{b \in X_2} P(a, b) \log \frac{p(a, b)}{p(a)p(b)}$$

其中， $p(a)$ 表示数值 a 出现的概率， $p(b)$ 表示数值 b 出现的概率， $p(a, b)$ 表示数值 b 出现的前提
下数值 a 出现的概率， X_1 和 X_2 代表所述多个用户属性中任两个用户属性，数值 a 为属于用户属性 X_1 的任
一数值，数值 b 为属于用户属性 X_2 的任一数值。

5、根据权利要求 1 所述的方法，其特征在于，所述用户数据中包括标签数据以及未标签数据，所述
25 将所述用户数据输入到所述贝叶斯网络模型中进行训练的步骤包括：采用半监督学习方法对输入到贝叶斯
网络模型中的用户数据进行训练。

6、根据权利要求5所述的方法，其特征在于，所述采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练，包括：

利用所述贝叶斯网络模型对未标签数据进行标签预测；

利用所述贝叶斯网络模型对标签数据进行训练；

5 重复交替执行上述两个步骤，直至训练过程收敛。

7、根据权利要求1所述的方法，其特征在于，在得到所述人群画像分类模型之后，所述方法还包括：

在接收到新输入的用户数据时，利用所述人群画像分类模型对所述用户数据进行人群画像分类，得到对应的分类结果。

8、一种人群画像分类模型的建立装置，所述人群画像分类模型的建立装置包括：

10 数据获取单元，用于获取待进行人群画像分类的用户数据，其中每一条用户数据包括该用户对应的多个用户属性；

因子关联单元，用于将每个用户属性作为 Chow-Liu 算法的一个因子，利用所述 Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子，得到贝叶斯网络模型；

15 数据训练单元，用于将所述用户数据输入到所述贝叶斯网络模型中进行训练，得到所述人群画像分类模型。

9、根据权利要求8所述的装置，其特征在于，所述装置还包括：预处理单元，用于对所述用户数据进行数据预处理。

10、根据权利要求9所述的装置，其特征在于，当数据预处理包括：数据清洗以及标准化处理时，所述预处理单元包括：数据清洗模块和标准化处理模块；所述数据清洗模块，用于删除用户数据中的空缺
20 数据、噪声数据、重复数据以及错误数据；标准化处理模块，用于将同一个用户对应的多个数据进行整合。

11、根据权利要求8所述的装置，其特征在于，所述因子关联单元具体用于执行以下步骤：

对于每一个因子，根据公式一在所有未选取的因子中选择与其 KL 距离最小的因子作为该因子的关联因子，直至所有因子均被选取；

所述公式一为： $KL(P(X) || T(X)) = -\sum I(X_i, Pa(X_i)) + \sum H(X_i) - H(X_1, X_2, \dots, X_n)$

25 其中， $KL(P(X) || T(X))$ 表示该因子与所有未选择的因子中任一因子的 KL 距离， $P(X)$ 表示进行关联之前所有因子的分布情况， $T(X)$ 表示进行关联之后所有因子的分布情况； X_i 表示第 i 个因子，H 表示熵， $Pa(X_i)$ 表示 X_i 的父节点；I 表示互信息，是通过公式二计算得到的，所述公式二为：

$$I(X_1, X_2) = \sum_{a \in X_1} \sum_{b \in X_2} P(a, b) \log \frac{p(a, b)}{p(a)p(b)}$$

其中, $p(a)$ 表示数值 a 出现的概率, $p(b)$ 表示数值 b 出现的概率, $p(a, b)$ 表示数值 b 出现的前提下数值 a 出现的概率, X_1 和 X_2 代表所述多个用户属性中任两个用户属性, 数值 a 为属于用户属性 X_1 的任一数值, 数值 b 为属于用户属性 X_2 的任一数值。

5 12、根据权利要求 8 所述的装置, 其特征在于, 当所述用户数据中包括标签数据以及未标签数据时, 数据训练单元具体用于采用半监督学习方法对输入到贝叶斯网络模型中的用户数据进行训练。

13、根据权利要求 12 所述的装置, 其特征在于, 当所述用户数据中包括标签数据以及未标签数据时, 所述数据训练单元具体用于执行以下步骤:

利用所述贝叶斯网络模型对未标签数据进行标签预测;

10 利用所述贝叶斯网络模型对标签数据进行训练;

重复交替执行上述两个步骤, 直至训练过程收敛。

14、根据权利要求 8 所述的装置, 其特征在于, 所述装置还包括: 分类单元, 用于在接收到新输入的用户数据时, 利用所述人群画像分类模型对所述用户数据进行人群画像分类, 得到对应的分类结果。

15 15、一种计算机设备, 包括非易失性可读存储介质和处理器, 所述非易失性可读存储介质中存储有计算机可读指令, 所述计算机可读指令被所述处理器执行时实现人群画像分类模型的建立方法, 包括: 获取待进行人群画像分类的用户数据, 其中每一条用户数据包括该用户对应的多个用户属性; 将每个用户属性作为 Chow-Liu 算法的一个因子, 利用所述 Chow-Liu 算法在所有因子中选择因子进行关联, 直至关联所有的因子, 得到贝叶斯网络模型; 将所述用户数据输入到所述贝叶斯网络模型中进行训练, 得到所述人群画像分类模型。

20 16、根据权利要求 15 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实现在获取待进行人群画像分类的用户数据的步骤之后, 所述方法还包括: 对所述用户数据进行数据预处理。

17、根据权利要求 16 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实现所述数据预处理包括: 数据清洗以及标准化处理; 所述数据清洗包括: 删除用户数据中的空缺数据、噪声数据、重复数据以及错误数据; 所述标准化处理包括: 将同一个用户对应的多个数据进行整合。

25 18、一种存储有计算机可读指令的非易失性可读存储介质, 所述计算机可读指令被一个或多个处理器执行时实现人群画像分类模型的建立方法, 包括: 获取待进行人群画像分类的用户数据, 其中每一条用户数据包括该用户对应的多个用户属性; 将每个用户属性作为 Chow-Liu 算法的一个因子, 利用所述

Chow-Liu 算法在所有因子中选择因子进行关联，直至关联所有的因子，得到贝叶斯网络模型；将所述用户数据输入到所述贝叶斯网络模型中进行训练，得到所述人群画像分类模型。

19、根据权利要求 18 所述的非易失性可读存储介质，其特征在于，所述处理器执行所述计算机可读指令时实现在获取待进行人群画像分类的用户数据的步骤之后，所述方法还包括：对所述用户数据进行数据预处理。

20、根据权利要求 19 所述的非易失性可读存储介质，其特征在于，所述处理器执行所述计算机可读指令时实现所述数据预处理包括：数据清洗以及标准化处理；所述数据清洗包括：删除用户数据中的空缺数据、噪声数据、重复数据以及错误数据；所述标准化处理包括：将同一个用户对应的多个数据进行整合。

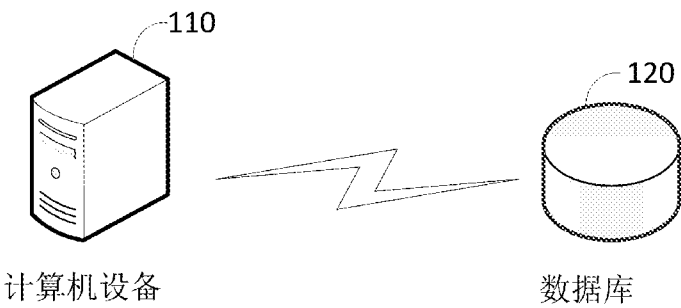


图 1

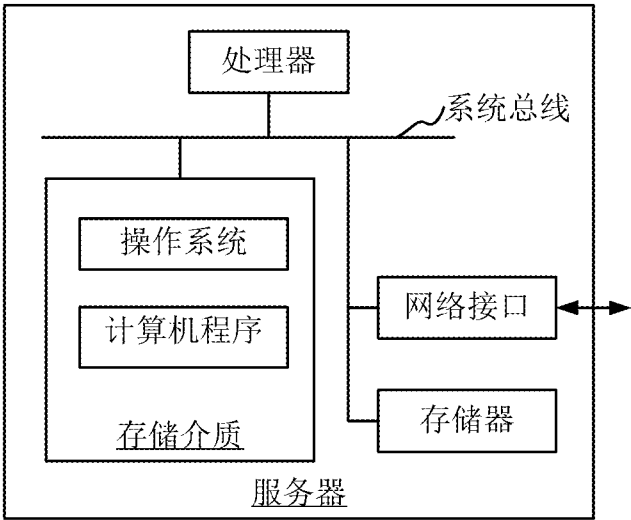


图 2

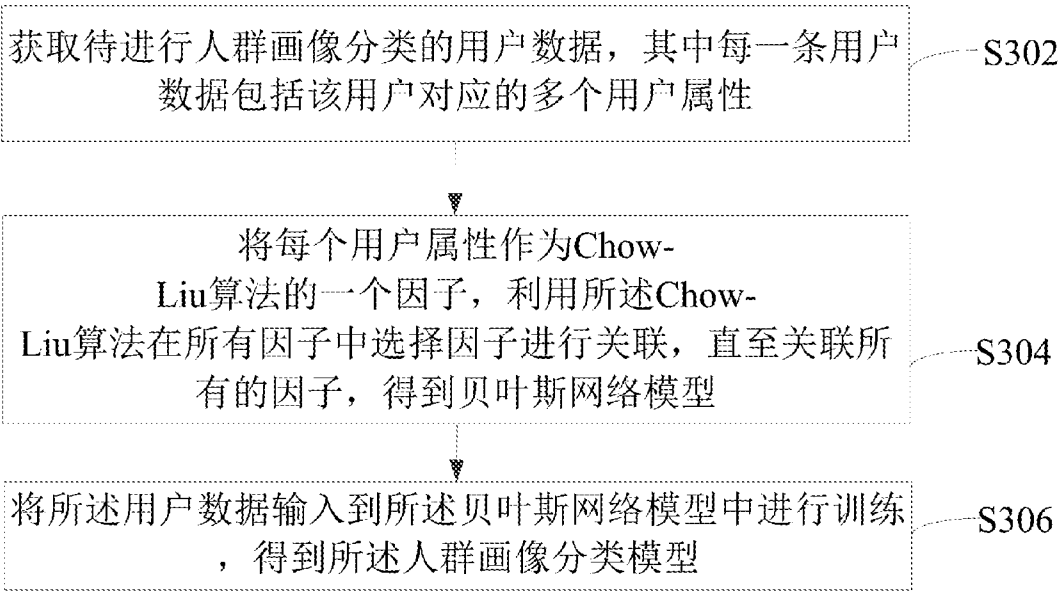


图 3

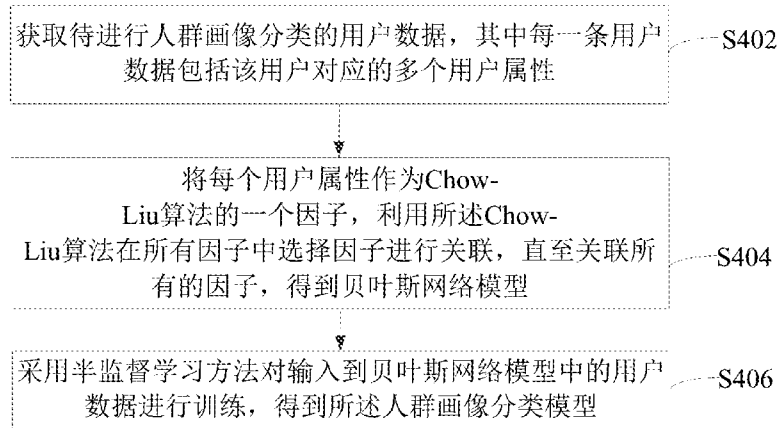


图 4

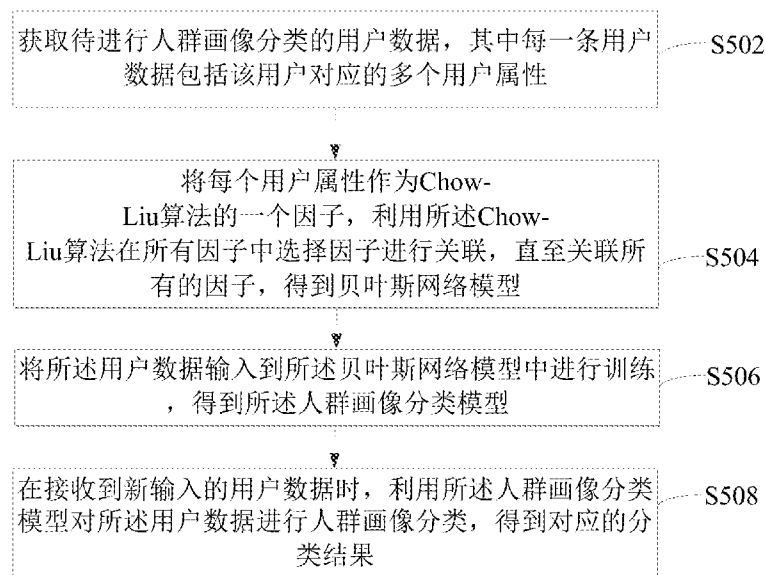


图 5

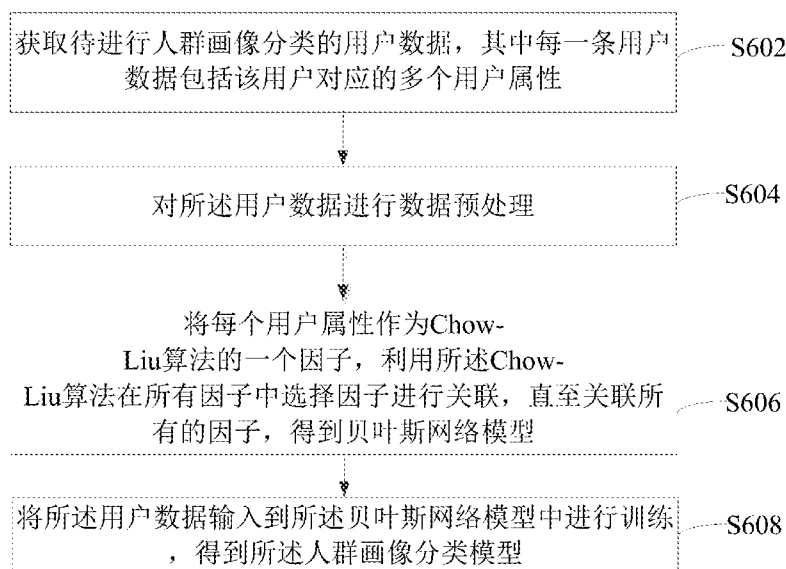


图 6

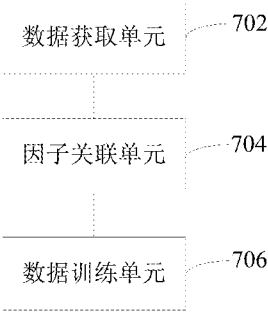


图 7

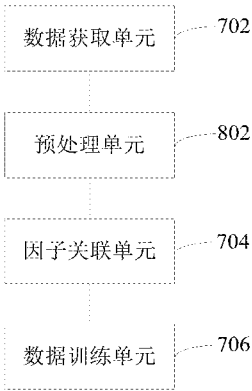


图 8

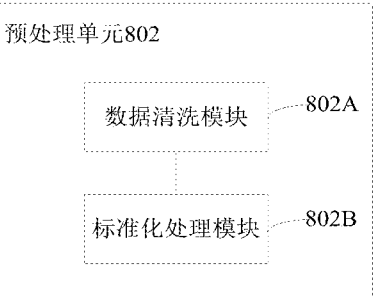


图 9

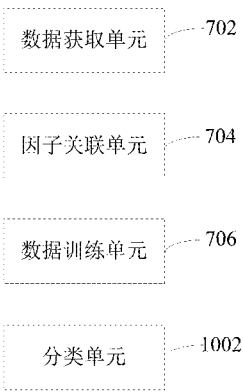


图 10

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/097892

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, G06Q, 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)

CNPAT, CNKI, IEEE, EPODOC, WPI: 画像, 标签, 建模, 模型, 贝叶斯, 分类, 属性, 关联, 依赖, 训练, 学习, user profile, user model+, user label+, persona, Bayesian, Bayes, class+, attribute, associat+, relat+, depend+, train+, learn+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109740620 A (PING'AN TECHNOLOGY (SHENZHEN) CO., LTD.) 10 May 2019 (2019-05-10) entire document	1-20
Y	CN 106651424 A (STATE GRID SHANDONG ELECTRIC POWER RESEARCH INSTITUTE et al.) 10 May 2017 (2017-05-10) description, paragraphs [0010]-[0044] and [0053]-[0087], and figures 1-2	1-3, 5-10, 12-20
Y	孙文静 (SUN, Wenjing). "基于依赖分析和假设检验的贝叶斯分类器 (Bayesian Classifier Based on Dependency Analysis and Hypothesis Testing)" 中国优秀硕士学位论文全文数据库 信息科技辑 (Electronic Technology & Information Science, China Master's Theses Full-Text Database), No. no. 11, 15 November 2014 (2014-11-15), chapter 1, section 1.1.2	1-3, 5-10, 12-20
A	CN 105893406 A (LETV CLOUD COMPUTING CO., LTD.) 24 August 2016 (2016-08-24) entire document	



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

"D" document cited by the applicant in the international application

"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

28 September 2019

Date of mailing of the international search report

29 October 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

International application No.

PCT/CN2019/097892**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 107391603 A (BEIJING QIHOO TECHNOLOGY CO., LTD.) 24 November 2017 (2017-11-24) entire document	1-20
A	CN 108021700 A (BAOFENG GROUP CO., LTD.) 11 May 2018 (2018-05-11) entire document	1-20
A	US 2008195431 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 14 August 2008 (2008-08-14) entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/097892

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109740620	A	10 May 2019	None			
CN	106651424	A	10 May 2017	None			
CN	105893406	A	24 August 2016	WO	2017080170	A1	18 May 2017
CN	107391603	A	24 November 2017	None			
CN	108021700	A	11 May 2018	None			
US	2008195431	A1	14 August 2008	None			

国际检索报告

国际申请号

PCT/CN2019/097892

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, G06Q, 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)) CNPAT, CNKI, IEEE, EPDOC, WPI: 画像, 标签, 建模, 模型, 贝叶斯, 分类, 属性, 关联, 依赖, 训练, 学习, user profile, user model+, user label+, persona, Bayesian, Bayes, class+, attribute, associat+, relat+, depend+, train+, learn+																							
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 109740620 A (平安科技深圳有限公司) 2019年 5月 10日 (2019 - 05 - 10) 全文</td> <td>1-20</td> </tr> <tr> <td>Y</td> <td>CN 106651424 A (国网山东省电力公司电力科学研究院 等) 2017年 5月 10日 (2017 - 05 - 10) 说明书第[0010]-[0044]、[0053]-[0087]段, 附图1-2</td> <td>1-3, 5-10, 12-20</td> </tr> <tr> <td>Y</td> <td>孙文静. “基于依赖分析和假设检验的贝叶斯分类器” 《中国优秀硕士学位论文全文数据库 信息科技辑》, 第11期, 2014年 11月 15日 (2014 - 11 - 15), 第一章第1.1.2节</td> <td>1-3, 5-10, 12-20</td> </tr> <tr> <td>A</td> <td>CN 105893406 A (乐视云计算有限公司) 2016年 8月 24日 (2016 - 08 - 24) 全文</td> <td></td> </tr> <tr> <td>A</td> <td>CN 107391603 A (北京奇虎科技有限公司) 2017年 11月 24日 (2017 - 11 - 24) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 108021700 A (暴风集团股份有限公司) 2018年 5月 11日 (2018 - 05 - 11) 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 109740620 A (平安科技深圳有限公司) 2019年 5月 10日 (2019 - 05 - 10) 全文	1-20	Y	CN 106651424 A (国网山东省电力公司电力科学研究院 等) 2017年 5月 10日 (2017 - 05 - 10) 说明书第[0010]-[0044]、[0053]-[0087]段, 附图1-2	1-3, 5-10, 12-20	Y	孙文静. “基于依赖分析和假设检验的贝叶斯分类器” 《中国优秀硕士学位论文全文数据库 信息科技辑》, 第11期, 2014年 11月 15日 (2014 - 11 - 15), 第一章第1.1.2节	1-3, 5-10, 12-20	A	CN 105893406 A (乐视云计算有限公司) 2016年 8月 24日 (2016 - 08 - 24) 全文		A	CN 107391603 A (北京奇虎科技有限公司) 2017年 11月 24日 (2017 - 11 - 24) 全文	1-20	A	CN 108021700 A (暴风集团股份有限公司) 2018年 5月 11日 (2018 - 05 - 11) 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 109740620 A (平安科技深圳有限公司) 2019年 5月 10日 (2019 - 05 - 10) 全文	1-20																					
Y	CN 106651424 A (国网山东省电力公司电力科学研究院 等) 2017年 5月 10日 (2017 - 05 - 10) 说明书第[0010]-[0044]、[0053]-[0087]段, 附图1-2	1-3, 5-10, 12-20																					
Y	孙文静. “基于依赖分析和假设检验的贝叶斯分类器” 《中国优秀硕士学位论文全文数据库 信息科技辑》, 第11期, 2014年 11月 15日 (2014 - 11 - 15), 第一章第1.1.2节	1-3, 5-10, 12-20																					
A	CN 105893406 A (乐视云计算有限公司) 2016年 8月 24日 (2016 - 08 - 24) 全文																						
A	CN 107391603 A (北京奇虎科技有限公司) 2017年 11月 24日 (2017 - 11 - 24) 全文	1-20																					
A	CN 108021700 A (暴风集团股份有限公司) 2018年 5月 11日 (2018 - 05 - 11) 全文	1-20																					
☒ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																							
<table border="0"> <tr> <td style="vertical-align: top;"> * 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 </td> <td style="vertical-align: top;"> “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 </td> </tr> </table>			* 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																			
* 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 28日		2019年 10月 29日																					
ISA/CN的名称和邮寄地址		授权官员																					
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 传真号 (86-10)62019451		贺艳娟 电话号码 86-(10)-53961817																					

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	US 2008195431 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2008年 8月 14日 (2008 - 08 - 14) 全文	1-20

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/097892

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	109740620	A	2019年 5月 10日	无	
CN	106651424	A	2017年 5月 10日	无	
CN	105893406	A	2016年 8月 24日	W0 2017080170 A1	2017年 5月 18日
CN	107391603	A	2017年 11月 24日	无	
CN	108021700	A	2018年 5月 11日	无	
US	2008195431	A1	2008年 8月 14日	无	