

(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
2021 年 4 月 22 日 (22.04.2021)



(10) 国际公布号
WO 2021/073259 A1

(51) 国际专利分类号:
G06K 9/62 (2006.01) **G06N 5/02 (2006.01)**

(21) 国际申请号: PCT/CN2020/111413

(22) 国际申请日: 2020 年 8 月 26 日 (26.08.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910991392.5 2019 年 10 月 18 日 (18.10.2019) 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: 廖希洋(LIAO, Xiyang); 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)。
孙行智(SUN, Xingzhi); 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)。
赵婷婷(ZHAO, Tingting); 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)。
田静涛(TIAN, Jingtao); 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)。
赵丽君(ZHAO, Lijun); 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)。

(54) Title: KNOWLEDGE-DRIVEN AND DATA-DRIVEN GROUPING METHOD AND RELATED DEVICE

(54) 发明名称: 基于知识驱动和数据驱动的分群方法及相关设备

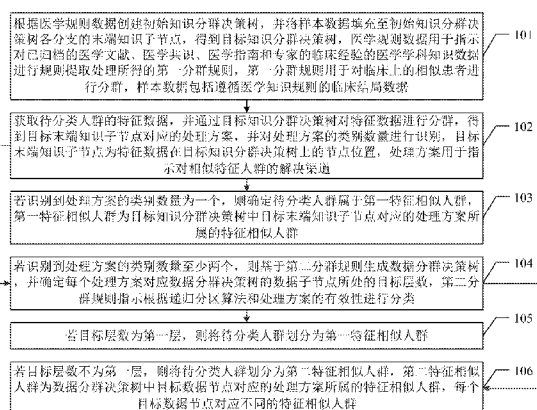


图 1

101 On the basis of medical rule data, create an initial knowledge grouping decision tree and fill sample data into end knowledge nodes of each branch of the initial knowledge grouping decision tree to obtain a target knowledge grouping decision tree, the medical rule data being used for indicating a first grouping rule obtained by performing rule extraction processing on medical discipline knowledge data of archived medical literature, medical consensus, medical guidelines, and expert clinical experience, the first grouping rule being used for grouping clinically similar patients, and the sample data comprising clinical outcome data that follows rules of medical knowledge

102 Acquire feature data of a group of people to be classified, group the feature data by means of the target knowledge grouping decision tree to obtain treatment plans corresponding to the target end knowledge sub-nodes, and identify the number of types of treatment plans, the target end knowledge sub-node being the position of the node of the feature data on the target knowledge grouping decision tree, and the treatment plans being used for indicating a solution channel for a group of people with similar features

103 If the number of types of treatment plan identified is one, then determine that the group of people to be classified belongs to a first group with similar features, the first group with similar features being a group with similar features to which a treatment plan corresponding to a target end knowledge sub-node in the target knowledge grouping decision tree belongs

104 If the number of types of treatment plan identified is at least two, then generate a data grouping decision tree on the basis of a second grouping rule, and determine the target level number on which a data sub-node of the data grouping decision tree corresponding to each treatment plan is located, the second grouping rule indicating that grouping is performed on the basis of a recursive partitioning algorithm and the effectiveness of the treatment plan

105 If the target level number is the first level, then divide the group of people to be classified into the first group with similar features

106 If the target level number is not the first level, then divide the group of people to be classified into a second group with similar features, the second group of people with similar features being a group with similar features to which a treatment plan corresponding to a target data node in the data grouping decision tree belongs, each target data node corresponding to a different group of people with similar features

(57) Abstract: The present application relates to the field of artificial intelligence, and discloses a knowledge-driven and data-driven grouping method and related device for increasing the accuracy of grouping a group of people to be classified into groups of people with similar features. The method of the present application comprises: creating a target knowledge grouping decision tree; identifying the number of types of treatment plan corresponding to target end knowledge sub-nodes on the target knowledge grouping decision tree of the feature data of the group of people to be classified; when the number of types identified is one, determining that the group of people to be classified belongs to a first group of people with similar features corresponding to a target end knowledge sub-node in the target knowledge grouping decision tree; when the number of types identified is at least two, generating a data grouping decision tree on the basis of a preset second grouping rule, and determining a target level number of the data grouping decision tree; and, on the basis of the target level number, dividing the group of people to be classified into the first group of people with similar features or a second group of people with similar features corresponding to a target data node in the data grouping decision tree.

(74) 代理人:北京市京大律师事务所(BEIJING JINGDA LAW FIRM); 中国北京市海淀区海淀中街16号7层2单元703, Beijing 100080 (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, IT, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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: 本申请涉及人工智能领域, 公开了一种基于知识驱动和数据驱动的分群方法及相关设备, 提高了将待分类人群分群到相似特征人群的精准度。本申请方法包括: 创建目标知识分群决策树; 识别待分类人群的特征数据在所述目标知识分群决策树上目标末端知识子节点对应的处理方案的类别数量; 当识别到所述类别数量为一个时, 确定所述待分类人群属于目标知识分群决策树中目标末端知识子节点对应的第一特征相似人群; 当识别到所述类别数量为至少两个时, 基于预置第二分群规则生成数据分群决策树, 并确定所述数据分群决策树的目标层数; 根据所述目标层数将所述待分类人群划分为所述第一特征相似人群或所述数据分群决策树中目标数据节点对应的第二特征相似人群。

基于知识驱动和数据驱动的分群方法及相关设备

本申请要求于2019年10月18日提交中国专利局、申请号为201910991392.5、发明名称为“基于知识驱动和数据驱动的分群方法及相关设备”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。

技术领域

本申请涉及人工智能领域，尤其涉及基于知识驱动和数据驱动的分群方法及相关设备。

背景技术

个性化精准治疗是目前医学治疗领域一个重要的研究课题，也是医学研究领域一个值得期待的发展方向。而个性化精准治疗主要从群体出发，并根据患者的个体差异提供个性化治疗，在这个过程中，最困难的是如何确定不同群体，并针对群体中患者的个性化治疗给出充足的医学知识支撑。现有的分群系统中，常使用完全基于数据驱动或者完全基于知识驱动的分群方法，但两种方法都存在较多弊端。

完全基于数据驱动的分群方法中，其核心思想是基于机器学习或深度学习算法学习数据集中医生开药规则，并将学习到的规则应用于同类病人之中。而完全基于知识驱动的分群方法，其核心是参照已有指南、共识等相关权威知识构建规则，根据规则获取相应的辅助决策方案。然而，发明人发现，不同于其他行业，使用完全基于数据驱动的分群方法并不能保证算法学习到的医学规则符合医学知识，并且无法做到针对相同群体中不同病人进行个性化推荐对应的辅助决策方案；使用完全基于知识驱动的推荐方法则存在着无法对推荐的多种辅助决策方案进行合理的选择，并存在指南、共识等知识中完全无法获取的辅助信息，如：药品价格。因而，导致将待分类人群分群到相似特征人群的精准度低。

发明内容

本申请提供了一种基于知识驱动和数据驱动的分群方法及相关设备，用于创建目标知识分群决策树、确定待分类人群属于第一特征相似人群、创建数据分群决策树和根据目标层数将待分类人群划分为第一特征相似人群或第二特征相似人群等步骤，提高将待分类人群分群到相似特征人群的精准度，进而为医生决策提供合理规范性且全面性精细化的循证医学证据。

本申请实施例的第一方面提供一种基于知识驱动和数据驱动的分群方法，包括：根据医学规则数据创建初始知识分群决策树，并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，所述医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，所述第一分群规则用于对临床上的相似特征人群进行分群，所述样本数据包括遵循医学知识规则的临床上的相似特征人群对应的临床结局数据；获取待分类人群的特征数据，并通过所述目标知识分群决策树对所述特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对所述处理方案的类别数量进行识别，所述目标末端知识子节点为所述特征数据在所述目标知识分群决策树上的节点位置，所述处理方案用于指示对相似特征人群的解决渠道；若识别到所述处理方案的类别数量为一个，则确定所述待分类人群属于第一特征相似人群，所述第一特征相似人群为所述目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群；若识别到所述处理方案的类别数量为至少两个，则基于第二分群规则生成数据分群决策树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数，所述第二分群规则指示根据递归分区算法和处理方案的有效性进行分类；若所述目标层数为第一层，则将所述待分类人群划分为所述第一特征相似人群；若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群，所述第二特征相似人群为所述数据分群决策树中目标数据节点对应的处理方案所属的特征相似人群，每个目标数据节点对应不同的特征相似人群。

本申请实施例的第二方面提供一种用于基于知识驱动和数据驱动的分群装置，所述基于知识驱动和数据驱动的分群装置包括：第一创建单元，用于根据医学规则数据创建初始知识分群

决策树，并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，所述医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，所述第一分群规则用于对临床上的相似特征人群进行分群，所述样本数据包括遵循医学知识规则的临床上的相似特征人群对应的临床结局数据；识别单元，用于获取待分类人群的特征数据，并通过所述目标知识分群决策树对所述特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对所述处理方案的类别数量进行识别，所述目标末端知识子节点为所述特征数据在所述目标知识分群决策树上的节点位置，所述处理方案用于指示对相似特征人群的解决渠道；第二创建单元，用于若识别到所述处理方案的类别数量为至少两个，则基于第二分群规则生成数据分群决策树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数，所述第二分群规则指示根据递归分区算法和处理方案的有效性进行分类；第一划分单元，用于若所述目标层数为第一层，则将所述待分类人群划分为所述第一特征相似人群；第二划分单元，用于若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群，所述第二特征相似人群为所述数据分群决策树中目标数据节点对应的处理方案所属的特征相似人群，每个目标数据节点对应不同的特征相似人群。

本申请实施例的第三方面提供了一种基于知识驱动和数据驱动的分群装置，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时执行如下所述的基于知识驱动和数据驱动的分群方法的步骤：根据医学规则数据创建初始知识分群决策树，并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，所述医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，所述第一分群规则用于对临床上的相似特征人群进行分群，所述样本数据包括遵循医学知识规则的临床上的相似特征人群对应的临床结局数据；获取待分类人群的特征数据，并通过所述目标知识分群决策树对所述特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对所述处理方案的类别数量进行识别，所述目标末端知识子节点为所述特征数据在所述目标知识分群决策树上的节点位置，所述处理方案用于指示对相似特征人群的解决渠道；若识别到所述处理方案的类别数量为一个，则确定所述待分类人群属于第一特征相似人群，所述第一特征相似人群为所述目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群；若识别到所述处理方案的类别数量为至少两个，则基于第二分群规则生成数据分群决策树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数，所述第二分群规则指示根据递归分区算法和处理方案的有效性进行分类；若所述目标层数为第一层，则将所述待分类人群划分为所述第一特征相似人群；若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群，所述第二特征相似人群为所述数据分群决策树中目标数据节点对应的处理方案所属的特征相似人群，每个目标数据节点对应不同的特征相似人群。

本申请实施例的第四方面提供了一种计算机可读存储介质，包括指令，当所述指令在计算机上运行时，使得计算机执行如下所述的基于知识驱动和数据驱动的分群方法的步骤：根据医学规则数据创建初始知识分群决策树，并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，所述医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，所述第一分群规则用于对临床上的相似特征人群进行分群，所述样本数据包括遵循医学知识规则的临床上的相似特征人群对应的临床结局数据；获取待分类人群的特征数据，并通过所述目标知识分群决策树对所述特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对所述处理方案的类别数量进行识别，所述目标末端知识子节点为所述特征数据在所述目标知识分群决策树上的节点位置，所述处理方案用于指示对相似特征人群的解决渠道；若识别到所述处理方案的类别数量为一个，则确定所述待分类人群属于第一特征相似人群，所述第一特征相似人群为所述目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群；若识别到所述处理方案的类别数量为至少两个，则基于第二分群规则生成数据分群决策

树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数，所述第二分群规则指示根据递归分区算法和处理方案的有效性进行分类；若所述目标层数为第一层，则将所述待分类人群划分为所述第一特征相似人群；若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群，所述第二特征相似人群为所述数据分群决策树中目标数据节点对应的处理方案所属的特征相似人群，每个目标数据节点对应不同的特征相似人群。

相较于现有技术，本申请实施例提供的技术方案中，通过创建目标知识分群决策树；识别待分类人群的特征数据在所述目标知识分群决策树上目标末端知识子节点对应的处理方案的类别数量；当识别到所述类别数量为一个时，确定所述待分类人群属于目标知识分群决策树中目标末端知识子节点对应的第一特征相似人群；当识别到所述类别数量为至少两个时，基于预置第二分群规则生成数据分群决策树，并确定所述数据分群决策树的目标层数；根据所述目标层数将所述待分类人群划分为所述第一特征相似人群或所述数据分群决策树中目标数据节点对应的第二特征相似人群。本申请实施例，通过创建目标知识分群决策树、确定待分类人群属于第一特征相似人群、创建数据分群决策树和根据目标层数将待分类人群划分为第一特征相似人群或第二特征相似人群等步骤，提高将待分类人群分群到相似特征人群的精准度，进而为医生决策提供合理规范性且全面性精细化的循证医学证据。

附图说明

图1为本申请实施例中基于知识驱动和数据驱动的分群方法的一个实施例示意图；
图2为本申请实施例中输出第一特征相似人群的一种示意图；
图3为本申请实施例中输出第二特征相似人群的一种示意图；
图4为本申请实施例中采用知识分群决策树和数据分群决策树分群的一种示意图；
图5为本申请实施例中基于知识驱动和数据驱动的分群方法的另一个实施例示意图；
图6为本申请实施例中基于知识驱动和数据驱动的分群装置的一个实施例示意图；
图7为本申请实施例中基于知识驱动和数据驱动的分群装置的另一个实施例示意图；
图8为本申请实施例中基于知识驱动和数据驱动的分群设备的另一个实施例示意图。

具体实施方式

本申请实施例提供了一种基于知识驱动和数据驱动的分群方法及相关设备，用于创建目标知识分群决策树、确定待分类人群属于第一特征相似人群、创建数据分群决策树和根据目标层数将待分类人群划分为第一特征相似人群或第二特征相似人群等步骤，提高将待分类人群分群到相似特征人群的精准度，进而为医生决策提供合理规范性且全面性精细化的循证医学证据。

为了使本技术领域的人员更好地理解本申请方案，下面将结合本申请实施例中的附图，对本申请实施例进行描述。

应当理解，此处所描述的具体实施例仅用以解释本申请，并不用于限定本申请。本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或模块的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或模块，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或模块，本申请中所出现的模块的划分，仅仅是一种逻辑上的划分，实际应用中实现时可以有另外的划分方式，例如多个模块可以结合成或集成在另一个系统中，或一些特征可以忽略，或不执行。

请参阅图1，本申请实施例提供的一种基于知识驱动和数据驱动的分群方法的流程图，以下对本申请提供一种基于知识驱动和数据驱动的分群方法进行举例说明，本申请不对执行主体的类型作限制，具体包括：

101、根据医学规则数据创建初始知识分群决策树，并将样本数据填充至初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，医学规则数据用于指示对已归档的医

学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，第一分群规则用于对临床上的相似患者进行分群，样本数据包括遵循医学知识规则的临床结局数据。

具体的，对医学指南、医学共识、医学文献和专家的临床经验的医学学科知识数据进行数据清理处理和数据规约处理，以保证医学学科知识数据的质量。对经过清理处理和数据规约处理的医学学科知识数据进行规则提取，将提取的规则构建较粗粒度的分群规则（即医学规则数据），用较粗粒度的分群规则构建初始知识分群决策树，以得到目标知识分群决策树。例如，如图2所示，将临床结局数据中遵循医学知识规则的临床结局数据（即样本数据）对应分群的临床上的相似特征人群的病症信息填充在对应的知识分群决策树各分支的末端知识子节点上，即末端知识子节点1、末端知识子节点2、末端知识子节点3、末端知识子节点4、末端知识子节点5、末端知识子节点6和末端知识子节点7，知识分群决策树之外的临床结局数据为不遵循医学知识规则的临床结局数据，则知识分群决策树各分支的末端知识子节点（末端知识子节点1、末端知识子节点2、末端知识子节点3、末端知识子节点4、末端知识子节点5、末端知识子节点6和末端知识子节点7）包含至少一种以上的病症信息对应的处理方案。通过从医学指南、医学共识、医学文献和专家的临床经验的医学学科知识数据获取的较粗粒度的分群规则和将第一样本数据填充到知识分群决策树各分支的末端知识子节点构建基于医学学科知识数据和临床结局数据的知识分群决策树。上述图2内容和举例内容仅作举例说明参考，其内容的准确性和实际操作与否不作考虑。

102、获取待分类人群的特征数据，并通过目标知识分群决策树对特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对处理方案的类别数量进行识别，目标末端知识子节点为特征数据在目标知识分群决策树上的节点位置，处理方案用于指示对相似特征人群的解决渠道。

终端获取待分类人群的特征数据，并通过目标知识分群决策树基于医学学科知识根据特征数据将待分类人群进行分群，获得待分类人群被分群到目标知识分群决策树所处的目标末端知识子节点，并获取目标末端知识子节点对应的相似特征人群的处理方案，对处理方案的类别数量进行识别，以判断是否对目标末端知识子节点的特征相似人群进行进一步分群。其中，处理方案用于指示对相似特征人群的解决渠道，例如，临床上的相似特征人群对应病症的治疗方案或者处方。

103、若识别到处理方案的类别数量为一个，则确定待分类人群属于第一特征相似人群，第一特征相似人群为目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群。

终端若识别到处理方案的类别数量为一个，则将待分类人群划分到所位于目标知识分群决策树所处的目标末端知识子节点对应的特征相似人群（即第一特征相似人群），得到仅通过目标知识分群决策树对待分类人群进行分群的第一特征相似人群。

例如，具体的，可通过处理方案之间的用药处方的相似度是否处于预设判断值来进行分析来分析各分支的末端子节点的处理方案的类别数量是否为一个。如图2的末端子节点1的处理方案有三个，分别为处理方案甲、处理方案乙和处理方案丙，预设判断值为相似度为97%~100%，处理方案甲与处理方案乙的用药处方的相似度为96%，处理方案甲与处理方案丙的用药处方的相似度为98%，处理方案乙与处理方案丙的用药处方的相似度为97.5%，则末端子节点1的处理方案的类别数量为2个。当分析到各分支的末端知识子节点的处理方案的类别数量是一个时，该末端知识子节点不进行后续的数据驱动分群模型的创建的操作，表示该末端知识子节点的人群是同一类的特征相似人群（即第一特征相似人群）。上述举例内容仅作举例说明参考，其内容的准确性和实际操作与否不作考虑。

104、若识别到处理方案的类别数量至少两个，则基于第二分群规则生成数据分群决策树，并确定每个处理方案对应数据分群决策树的数据子节点所处的目标层数，第二分群规则指示根据递归分区算法和处理方案的有效性进行分类。

具体的，当分析到各分支的末端知识子节点的处理方案的类别数量不是一个时，进入后续

的数据分群决策树的创建的操作，数据分群决策树通过机器学习或者深度学习算法学习数据集中的分群方式以对临床上的相似特征人群进行分群。以目标末端知识子节点作为父节点，以父节点开始，根据递归分区算法和处理方案的有效性进行分类的第二分群规则对数据根节点对应的相似特征人群的临床结局数据进行分支，以创建数据分群决策树，并确定每个处理方案对应数据分群决策树的数据子节点所处的目标层数。通过基于在知识驱动分群中的医学学科知识数据进行数据驱动分群，以对相似人群进行进一步分群，提高分群的精度。

例如，如图3所示，根据第一层的数据根节点的相似特征人群的处理方案的有效性对第一层的父节点的相似特征人群进行分支，在第2层将甲数据节点分群为处理方案的有效性为百分之九十以下的第二特征相似人群，将乙数据节点分群为处理方案的有效性为百分之九十以上的第二特征相似人群，在第3层将丙数据节点分群为处理方案的有效性为百分之八十至百分之八十五的第二特征相似人群，将丁数据节点分群为处理方案的有效性为百分之八十六至百分之九十的第二特征相似人群。

其中，有效性包括但不限于药物疗效有效性，目标层数用于判断是将待分类人群划分为目标知识分群决策树对待分类人群进行分群所得的第一特征相似人群还是通过数据分群决策树对目标知识分群决策树对待分类人群进行分群所得的第一特征相似人群再进行进一步分群所得的第二特征相似人群。从数据分群决策树的父节点开始，每层的各分支的数据子节点由上至下，对父节点的特征相似人群的处理方案和子节点的特征相似人群的处理方案的药物疗效有效性是否存在差异进行分析，以处理方案的药物疗效有效性存在显著差异对应的子节点对应的层数作为数据分群决策树的选择层数，并将在父节点对应的知识分群决策树的该末端子节点上标记该目标层数。以目标层数作为是否调用数据分群决策树且通过数据分群决策树对相似人群进行基于数据驱动分群的层数的依据。其中，可通过治愈率或显效率或好转率或无效率或其他药物有效性的指标中的至少一项进行药物疗效有效性分析。

105、若目标层数为第一层，则将待分类人群划分为第一特征相似人群。

具体的，当目标层数为第一层，即数据分群决策树的父节点时，将待分类人群划分为目标知识分群决策树对待分类人群进行分群所得的第一特征相似人群，不再通过数据分群决策树分群对待分类人群进行分群。

106、若目标层数不为第一层，则将待分类人群划分为第二特征相似人群，第二特征相似人群为数据分群决策树中目标数据节点对应的处理方案所属的特征相似人群，每个目标数据节点对应不同的特征相似人群。

具体的，当目标层数为第一层，即数据分群决策树的父节点时，获取目标层数在数据分群决策树对应的层数的末端子节点，并获取该末端子节点对应的特征相似人群，将该特征相似人群作为结合目标知识分群决策树和数据分群决策树对待分类人群进行分群所得的第二特征相似人群。例如：如图3所示，若目标层数为2，则将待分类人群划分到乙数据节点对应的第二特征相似人群，若目标层数为3，则将待分类人群划分到丙节点和丁节点对应的第二特征相似人群。

例如，如图4所示，结合知识分群决策树和数据决策树对待分类人群进行分群，将待分类人群划分到对应的特征相似人群，根据目标层数判断是将待分类人群划分为目标知识分群决策树对待分类人群进行分群所得的第一特征相似人群还是通过数据分群决策树对目标知识分群决策树对待分类人群进行分群所得的第一特征相似人群再进行进一步分群所得的第二特征相似人群，医学辅助决策系统根据本申请的方法所获得的第一特征相似人群，得到第一特征相似人群对应的提供医学知识以及补充相应的真实世界证据的第一循证医学信息，根据本申请的方法所获得的第二特征相似人群，得到第二特征相似人群对应的提供医学知识、补充相应的真实世界证据和结合数据驱动进行分群的第二循证医学信息。不仅定义相似患者，还提供针对相同群体中不同病人的符合医学学科知识的个性化循证医学证据信息，以辅助医生决策。

本申请实施例，通过创建目标知识分群决策树、确定待分类人群属于第一特征相似人群、创建数据分群决策树和根据目标层数将待分类人群划分为第一特征相似人群或第二特征相似人群等步骤，提高将待分类人群分群到相似特征人群的精准度，进而为医生决策提供合理规范性且全面性精细化的循证医学证据。

请参阅图 5，本申请实施例中，501 至 505 的方法可参见 101 至 105，此处不再赘述。本申请实施例中基于知识驱动和数据驱动的分群方法的另一个实施例包括：

501、根据医学规则数据创建初始知识分群决策树，并将样本数据填充至初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，第一分群规则用于对临床上的相似患者进行分群，样本数据包括遵循医学知识规则的临床结局数据。

可选的，上述的根据医学规则数据创建初始知识分群决策树，并将样本数据填充至初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树之前，还包括：从多个系统、多个平台和多个医疗系统获取医学学科知识数据和临床数据，并对医学学科知识数据和临床数据进行数据处理，获得医学规则数据，数据处理包括数据清理处理、数据集成处理、数据规约处理、数据变换处理、特征提取处理、数据脱敏处理和分类处理。具体包括：从多个系统和多个平台获取医学学科知识数据，以及从多个医疗系统获取临床数据；分别对医学学科知识数据和临床数据进行数据预处理，数据预处理包括数据清理处理、数据集成处理、数据规约处理和数据变换处理；对经过数据预处理的医学学科知识数据进行特征提取得到特征信息，以及对经过数据预处理的临床数据进行数据脱敏处理和分类得到第一样本数据和第三样本数据；分析特征信息，按照预设分类规则对特征信息进行分类，获得医学规则数据，预设分类规则包括相同和/或相似病症。通过对医学学科知识数据进行数据清理处理、数据集成处理、数据规约处理和数据变换处理的数据预处理，以保证医学学科知识数据的质量，以便于后续对医学学科知识数据的数据分析和数据挖掘。通过对医疗数据中的某些敏感信息按照脱敏规则进行数据的变形，以对医疗数据中的敏感隐私数据进行可靠保护，在保证敏感隐私数据可靠保护的基础上对医疗数据的质量不产生影响。其中，可通过对一个页面可能执行一条结构化查询语言(structured query language, SQL)，以保证各个医疗数据进行脱敏互不干扰。

502、获取待分类人群的特征数据，并通过目标知识分群决策树对特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对处理方案的类别数量进行识别，目标末端知识子节点为特征数据在目标知识分群决策树上的节点位置，处理方案用于指示对相似特征人群的解决渠道。

可选的，上述的对处理方案的类别数量进行识别，具体还包括：计算处理方案之间的相似度获得多个相似度值，并比较多个相似度值是否均在预设范围内；若比较结果为是，则统计处理方案的类别为一个；若比较结果为否，则统计处理方案的类别为多个。通过判断处理方案之间的多个相似度值是否均在预设范围内，以判断处理方案的类别。例如，相似度包括但不限于用药处方的相似度、用药处方结合药品用途的相似度和病症信息结合用药处方的相似度。通过判断处理方案的之间的类型，根据处理方案的类型判断是否构建数据驱动分群模型，以便于根据输入的就诊数据的不同进行不同的操作，进而减少不必要的操作和提高效率，以及保证精准分群。

503、若识别到处理方案的类别数量为一个，则确定待分类人群属于第一特征相似人群，第一特征相似人群为目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群。

504、若识别到处理方案的类别数量至少两个，则基于第二分群规则生成数据分群决策树，并确定每个处理方案对应数据分群决策树的数据子节点所处的目标层数，第二分群规则指示根据递归分区算法和处理方案的有效性进行分类。

可选的，上述的基于第二分群规则生成数据分群决策树，并确定每个处理方案对应数据分群决策树的数据子节点所处的目标层数，具体包括：步骤一：将目标末端知识子节点作为数据分群决策树的数据根节点，计算数据根节点对应的处理方案的有效性分值；步骤二：根据每个处理方案的有效性分值对数据根节点进行分支，根据递归分区算法将每个处理方案依次划分到每个分支的数据子节点中，并继续对每个数据子节点对应的处理方案进行有效性分析；步骤三：若当前数据子节点所对应的处理方案的有效性分值大于第一阈值，则对当前数据子节点进行分

支；步骤四：重复执行步骤三，直至每个数据子节点所对应的处理方案的有效性分值小于或等于第一阈值，得到数据分群决策树，数据分群决策树包括多个数据子节点，多个数据子节点中每个末端数据叶子节点对应有一个处理方案；步骤五：确定每个处理方案对应的末端数据叶子节点在数据分群决策树中所处的目标层数。根据目标末端知识子节点对应的处理方案的有效性分值和递归分区算法构建数据分群决策树。例如，第一阈值为 60 分，将数据根节点的处理方案分支为 95 分以下和 95 分以上（即对特征相似人群进行分群），95 分以下的有效性分值均大于 60 分，则将 95 分以下的有效性分值对应的处理方案进行分支，95 分以上的有效性分值均大于 60 分，则将 95 分以上的有效性分值对应的处理方案进行分支，依次类推，直到最后分支对应的数据子节点所对应的处理方案的有效性分值小于或等于 60 分，此时的末端数据叶子节点在数据分群决策树中所处的层数即为目标层数。其中，数据分群决策树在分支时，可通过将处理方案的有效性分值作为参数，将通过测量或测定所得到的对处理方案进行划分的所有有效性分值的划分范围值作为所有观测值，据预置的目标函数和所有观测值对参数进行估计处理得到参数估计值，根据参数估计值评估参数在不同分区的稳定性；若评估结果为不稳定，则分析分区变量，并选择稳定性最低的分区变量进行稳定性评估，直到评估结果为稳定，分区变量用于指示分区的回归量；若评估结果为稳定，则停止评估参数在不同分区的稳定性；基于参数在不同分区的稳定性，根据二分法或变量自适应法获取优化的区域分割规则，二分法用于指示所选子节点的数目为 2 的一分为二方法，变量自适应法用于指示所选子节点的数目等于分区数量的方法；根据区域分割规则获取与区域分割规则对应的子节点，在与区域分割规则对应的子节点重复执行根据参数估计值评估参数在不同分区的稳定性和根据二分法或变量自适应法获取优化的区域分割规则，直到在评估到参数在不同分区稳定，以将待分类人群和目标知识分群决策树中目标末端子节点对应的特征相似人群进行分群，得到数据分群决策树。采用基于模型的递归分区与临床结局数据强相关的分群规则创建数据分群决策树。通过利用模型的目标函数估计参数和分割点，对每个节点参数不稳定性进行相应的模型评分，评估应该使用哪个变量进行分区，以利于创建数据分群决策树的非线性关系和自动检测解释变量之间的互动。在创建的过程中：每个节点都与模型相关联；对参数进行稳定性波动分析以评估节点是否需要分裂；若任何一个分区变量存在显著的不稳定性，则将节点分割为局部最优端，并重复该过程。参数的估计可通过得分函数或估计函数进行估计处理。参数在不同分区的稳定性的评估，可通过经验波动过程来捕捉。

可选的，上述的将目标末端知识子节点作为数据分群决策树的数据根节点，计算数据根节点对应的处理方案的有效性分值，具体包括：将目标末端知识子节点作为数据分群决策树的数据根节点；获取数据根节点对应的处理方案的目标特征信息，将目标特征信息拆分成因素因子，计算因素因子在不同水平上的关联度、对预置影响因素的影响度以及因素因子不控制的边缘关联度；通过预置的可压缩准则和可比较准则分析因素因子的关联度、影响度和边缘关联度，得到混杂因子；计算混杂因子与预置影响因素发生概率的回归系数的指数变换值，并以指数变换值作为判定得分，得到数据根节点对应的处理方案的有效性

$$p(y = k|x; w) = \frac{\exp(w_k^T x)}{\sum_{i=1}^K \exp(w_i^T x)}, \quad p(y = k|x; w) \text{ 为指数变换值,}$$

分值，指数变换值的计算公式如下：

w 为混杂因子与预置影响因素发生的概率，x 为混杂因子，k 为预置影响因素的类别数量，y 为输出的预测值，y 是 k 的条件概率，K 为 k 的最大值，T 为时间周期。例如，目标特征信息可为数据分群决策树的父节点的处理方案中的病症信息，预置影响因素为疾病，影响度为关于疾病的危险度，有效性为药物有效性；混杂因子是指一个与暴露和疾病都有联系的因子，它既是所研究疾病的独立危险因子，在非暴露组中它也必定是一个危险因子。对特征相似人群对应的处理方案进行药物有效性分析以获取存在统计分析意义的子节点，从而快速而准确地通过数据分群决策树对知识分群决策树分群之后再进一步分群。在药物有效性分析过程中，先选取出可能的混杂因子，然后将该混杂因子加入到多变量逻辑回归分析中，最后将药物变量的回归系数的

指数变换值作为最终的有效性的判定得分，以获取多个有效值。其中，可通过可压缩准则、可比较准则、虚拟事实模型、图模型作为混杂因子的识别准则，以获取混杂因子。

505、若目标层数为第一层，则将待分类人群划分为第一特征相似人群。

506、若目标层数不为第一层，判断目标层数对应处理方案的有效性分值是否大于第二阈值，第二阈值大于第一阈值。

当识别到目标层数不为第一层时，需要通过数据分群决策树对待分类人群进行进一步分群。获取目标层数对应处理方案的有效性分值，并判断有效性分值是否大于第二阈值，以保证对待分类人群分群的精准度。例如，如图3所示，目标层数为第2层，则获取乙节点的第二特征相似人群对应的处理方案的有效性分值，则分析该有效性分值是否大于第二阈值。

507、若目标层数对应处理方案的有效性分值大于第二阈值，则将有效性分值大于第二阈值的处理方案所位于的数据叶子节点作为目标数据节点，并将待分类人群划分为目标数据节点对应的处理方案所属的特征相似人群，得到第二特征相似人群。

当目标层数对应处理方案的有效性分值大于第二阈值时，不需要对待分类人群和特征相似人群进行进一步分群，该处已为结合知识驱动和数据驱动对待分类人群进行与临床结局数据强相关的分群所得的最终的特征相似人群，以目标层数对应处理方案的有效性分值大于第二阈值所位于数据决策树的目标数据节点对应特征相似人群作为将待分类人群划分的第二特征相似人群。例如，第二阈值为70分，目标层数对应处理方案的有效性分值为80分，则对应处理方案的有效性分值为80分的数据叶子节点作为目标数据节点，并将待分类人群划分为目标数据节点对应的处理方案所属的特征相似人群（即第二特征相似人群）。

508、若目标层数对应处理方案的有效性分值小于或者等于第二阈值，则分析目标层数的上一层数对应处理方案的有效性分值。

当目标层数对应处理方案的有效性分值小于或者等于第二阈值时，该目标层数对应的特征相似人群还不是对待分类人群分群所划分所属的最终人群，需要分析目标层数的上一层数对应处理方案的有效性分值。有效性分值小于或者等于第二阈值的处理方案对应的特征相似人群与待分类人群的匹配度低，对有效性分值小于或者等于第二阈值对应的节点或之后的节点不再进行分析，以提高操作效率和分群的精准度。

509、若目标层数的上一层数对应处理方案的有效性分值大于第二阈值，则将目标层数的上一层数对应处理方案所位于的数据叶子节点作为目标数据节点，并将待分类人群划分为目标数据节点对应的处理方案所属的特征相似人群，得到第二特征相似人群。

若目标层数的上一层数对应处理方案的有效性分值大于第二阈值，则将待分类人群划分到目标层数的上一层数对应处理方案的有效性分值大于第二阈值对应的数据分群决策树的目标节点的第二特征相似人群，该节点对应的第二特征相似人群为结合知识驱动和数据驱动对待分类人群进行与临床结局数据强相关的分群划分的最终的特征相似人群。若目标层数的上一层数对应处理方案的有效性分值小于或等于第二阈值，则对目标层数的上一层数的上一层数对应处理方案的有效性分值是否大于第二阈值进行判断，以此类推判断，直到分析到对应处理方案的有效性分值大于第二阈值，以处理方案的有效性分值大于第二阈值对应的数据分群决策树的目标节点的特征相似人群作为将待分类人群最终划分的第二特征相似人群。

可选的，上述的若目标层数不为第一层，则将待分类人群划分为第二特征相似人群之后，具体还包括：调用预置的解释模型对第一特征相似人群或第二特征相似人群进行解释处理和可视化图表统计分析处理，得到可解释性特征相似人群，解释处理包括有效性排序处理、匹配度计算处理和匹配度排序处理。具体的，可通过调用预置的解释模型对第一特征相似人群或第二特征相似人群的有效性按照值从大到小的顺序进行排序，得到第一排序信息；计算待分类人群与第一特征相似人群或第二特征相似人群的匹配度，并对匹配度按照值从大到小的顺序进行排序，得到第二排序信息；根据第一排序信息和/或第二排序信息对第一特征相似人群或第二特征相似人群进行排序，并将根据第一排序信息和/或第二排序信息标记在第一特征相似人群或第二特征相似人群上；对标记之后的第一特征相似人群或第二特征相似人群进行可视化图表统计分析处理，生成可视化图表，得到可解释性特征相似人群。通过调用预置的解释模型对第一特征

相似人群或第二特征相似人群，以便于用户便捷地获取将待分类人群划分的特征相似人群各项分群数据。

医学辅助决策系统根据本申请的方法所获得的第一特征相似人群，得到第一特征相似人群对应的提供医学知识以及补充相应的真实世界证据的第一循证医学信息，根据本申请的方法所获得的第二特征相似人群，得到第二特征相似人群对应的提供医学知识、补充相应的真实世界证据和结合数据驱动进行分群的第二循证医学信息之后具体的，包括：调用已创建的解释模型分别获取第一样本数据的第一特征信息、第一循证医学证据信息的第二特征信息和/或第二循证医学证据信息的第一特征信息，第一样本数据指示不遵循医学知识规则的临床结局数据；对第一特征信息和第二特征信息进行分析，得到第一分析数据，以及对第一特征信息和第三特征信息进行分析，得到第二分析数据；对第一特征信息、第二特征信息和第三特征信息进行统计和分析，生成可视化图表；调用已创建的知识图谱，获取知识图谱中与第二特征信息匹配的第一实物信息，以及获取知识图谱中与第三特征信息匹配的第二实物信息；将第一特征信息、第二特征信息、第三特征信息、第一分析数据、第二分析数据、第一实物信息、第二实物信息和可视化图表生成图文结合的分析报告，获得可解释性数据。将解释说明数据生成图文结合的形式显示，以便于用户更好地获悉可解释性数据的内容。通过对第一循证医学证据信息或第二循证医学证据信息进行解释说明得到可解释性数据，以有效地为辅助医生决策提供更加个性化精确化有效性的真实世界证据。

本申请实施例，通过创建目标知识分群决策树、确定待分类人群属于第一特征相似人群、创建数据分群决策树和根据目标层数将待分类人群划分为第一特征相似人群或第二特征相似人群等步骤，提高将待分类人群分群到相似特征人群的精准度，进而为医生决策提供合理规范性且全面性精细化的循证医学证据。

上面对本申请实施例中基于知识驱动和数据驱动的分群方法进行了描述，下面对本申请实施例中基于知识驱动和数据驱动的分群装置进行描述，请参阅图 6，本申请实施例中基于知识驱动和数据驱动的分群装置的一个实施例包括：

第一创建单元 601，用于根据医学规则数据创建初始知识分群决策树，并将样本数据填充至初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，第一分群规则用于对临床上的相似特征人群进行分群，样本数据包括遵循医学知识规则的临床上的相似特征人群对应的临床结局数据；

识别单元 602，用于获取待分类人群的特征数据，并通过目标知识分群决策树对特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对处理方案的类别数量进行识别，目标末端知识子节点为特征数据在目标知识分群决策树上的节点位置，处理方案用于指示对相似特征人群的解决渠道；

确定单元 603，用于若识别到处理方案的类别数量为一个，则确定待分类人群属于第一特征相似人群，第一特征相似人群为目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群；

第二创建单元 604，用于若识别到处理方案的类别数量为至少两个，则基于第二分群规则生成数据分群决策树，并确定每个处理方案对应数据分群决策树的数据子节点所处的目标层数，第二分群规则指示根据递归分区算法和处理方案的有效性进行分类；

第一划分单元 605，用于若目标层数为第一层，则将待分类人群划分为第一特征相似人群；
第二划分单元 606，用于若目标层数不为第一层，则将待分类人群划分为第二特征相似人群，第二特征相似人群为数据分群决策树中目标数据节点对应的处理方案所属的特征相似人群，每个目标数据节点对应不同的特征相似人群。

本申请实施例，通过创建目标知识分群决策树、确定待分类人群属于第一特征相似人群、创建数据分群决策树和根据目标层数将待分类人群划分为第一特征相似人群或第二特征相似人群等步骤，提高将待分类人群分群到相似特征人群的精准度，进而为医生决策提供合理规范性且全面性精细化的循证医学证据。

请参阅图 7, 本申请实施例中基于知识驱动和数据驱动的分群装置的另一个实施例包括:

第一创建单元 601, 用于根据医学规则数据创建初始知识分群决策树, 并将样本数据填充至初始知识分群决策树各分支的末端知识子节点, 得到目标知识分群决策树, 医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则, 第一分群规则用于对临床上的相似患者进行分群, 样本数据包括遵循医学知识规则的临床上的相似特征人群对应的临床结局数据;

识别单元 602, 用于获取待分类人群的特征数据, 并通过目标知识分群决策树对特征数据进行分群, 得到目标末端知识子节点对应的处理方案, 并对处理方案的类别数量进行识别, 目标末端知识子节点为特征数据在目标知识分群决策树上的节点位置, 处理方案用于指示对相似特征人群的解决渠道;

确定单元 603, 用于若识别到处理方案的类别数量为一个, 则确定待分类人群属于第一特征相似人群, 第一特征相似人群为目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群;

第二创建单元 604, 用于若识别到处理方案的类别数量至少两个, 则基于第二分群规则生成数据分群决策树, 并确定每个处理方案对应数据分群决策树的数据子节点所处的目标层数, 第二分群规则指示根据递归分区算法和处理方案的有效性进行分类;

第一划分单元 605, 用于若目标层数为第一层, 则将待分类人群划分为第一特征相似人群;

判断单元 606, 用于目标层数不为第一层, 判断目标层数对应处理方案的有效性分值是否大于第二阈值, 第二阈值大于第一阈值;

第三划分单元 607, 用于若目标层数对应处理方案的有效性分值大于第二阈值, 则将有效性分值大于第二阈值的处理方案所位于的数据叶子节点作为目标数据节点, 并将待分类人群划分为目标数据节点对应的处理方案所属的特征相似人群, 得到第二特征相似人群;

分析单元 608, 用于若目标层数对应处理方案的有效性分值小于或者等于第二阈值, 则分析目标层数的上一层数对应处理方案的有效性分值;

第三划分单元 609, 用于若目标层数的上一层数对应处理方案的有效性分值大于第二阈值, 则将目标层数的上一层数对应处理方案所位于的数据叶子节点作为目标数据节点, 并将待分类人群划分为目标数据节点对应的处理方案所属的特征相似人群, 得到第二特征相似人群。

可选的, 第二创建单元 604 具体用于: 步骤一: 将所述目标末端知识子节点作为所述数据分群决策树的数据根节点, 计算所述数据根节点对应的处理方案的有效性分值; 步骤二: 根据每个处理方案的有效性分值对所述数据根节点进行分支, 根据递归分区算法将每个处理方案依次划分到每个分支的数据子节点中, 并继续对每个数据子节点对应的处理方案进行有效性分析; 步骤三: 若当前数据子节点所对应的处理方案的有效性分值大于第一阈值, 则对所述当前数据子节点进行分支; 步骤四: 重复执行步骤三, 直至每个数据子节点所对应的处理方案的有效性分值小于或等于所述第一阈值, 得到数据分群决策树, 所述数据分群决策树包括多个数据子节点, 所述多个数据子节点中每个末端数据叶子节点对应有一个处理方案; 步骤五: 确定每个处理方案对应的末端数据叶子节点在所述数据分群决策树中所处的目标层数。

可选的, 第二创建单元 604 具体用于: 将所述目标末端知识子节点作为所述数据分群决策树的数据根节点; 获取所述数据根节点对应的处理方案的目标特征信息, 将目标特征信息拆分成因素因子; 计算因素因子在不同水平上的关联度、对预置影响因素的影响度以及因素因子不控制的边缘关联度; 通过预置的可压缩准则和可比较准则分析因素因子的关联度、影响度和边缘关联度, 得到混杂因子; 计算混杂因子与预置影响因素发生概率的回归系数的指数变换值, 并以指数变换值作为判定得分, 得到数据根节点对应的处理方案的有效性分值, 指数变换值的

$$p(y = k|x; w) = \frac{\exp(w_k^T x)}{\sum_{i=1}^K \exp(w_i^T x)}$$

计算公式如下: $p(y = k|x; w)$ 为指数变换值, w 为混杂因子与预置影响因素发生的概率, x 为混杂因子, k 为预置影响因素的类别数量, y 为输出的预测值, y 是 k

的条件概率, K 为 k 的最大值, T 为时间周期。

可选的, 基于知识驱动和数据驱动的分群装置, 还包括: 处理单元, 用于从多个系统、多个平台和多个医疗系统获取医学学科知识数据和临床数据, 并对医学学科知识数据和临床数据进行数据处理, 获得医学规则数据, 数据处理包括数据清理处理、数据集成处理、数据规约处理、数据变换处理、特征提取处理、数据脱敏处理和分类处理。处理单元具体用于: 从多个系统和多个平台获取医学学科知识数据, 以及从多个医疗系统获取临床数据; 分别对医学学科知识数据和临床数据进行数据预处理, 数据预处理包括数据清理处理、数据集成处理、数据规约处理和数据变换处理; 对经过数据预处理的医学学科知识数据进行特征提取得到特征信息, 以及对经过数据预处理的临床数据进行数据脱敏处理和分类得到第一样本数据和第三样本数据; 分析特征信息, 按照预设分类规则对特征信息进行分类, 获得医学规则数据, 预设分类规则包括相同和/或相似病症。

可选的, 识别单元 602 具体用于: 计算处理方案之间的相似度获得多个相似度值, 并比较多个相似度值是否均在预设范围内; 若比较结果为是, 则统计处理方案的类别为一个; 若比较结果为否, 则统计处理方案的类别为多个。

可选的, 基于知识驱动和数据驱动的分群装置, 还包括: 调用单元, 用于调用预置的解释模型对第一特征相似人群或第二特征相似人群进行解释处理和可视化图表统计分析处理, 得到可解释性特征相似人群, 解释处理包括有效性排序处理、匹配度计算处理和匹配度排序处理。调用单元具体用于: 调用预置的解释模型对第一特征相似人群或第二特征相似人群的有效性按照值从大到小的顺序进行排序, 得到第一排序信息; 计算待分类人群与第一特征相似人群或第二特征相似人群的匹配度, 并对匹配度按照值从大到小的顺序进行排序, 得到第二排序信息; 根据第一排序信息和/或第二排序信息对第一特征相似人群或第二特征相似人群进行排序, 并将根据第一排序信息和/或第二排序信息标记在第一特征相似人群或第二特征相似人群上; 对标记之后的第一特征相似人群或第二特征相似人群进行可视化图表统计分析处理, 生成可视化图表, 得到可解释性特征相似人群。

本申请实施例, 通过创建目标知识分群决策树、确定待分类人群属于第一特征相似人群、创建数据分群决策树和根据目标层数将待分类人群划分为第一特征相似人群或第二特征相似人群等步骤, 提高将待分类人群分群到相似特征人群的精准度, 进而为医生决策提供合理规范性且全面性精细化的循证医学证据。

上面图 6 至图 7 从模块化功能实体的角度对本申请实施例中的基于知识驱动和数据驱动的分群设备进行详细描述, 下面从硬件处理的角度对本申请实施例中基于知识驱动和数据驱动的分群设备进行详细描述。

图 8 是本申请实施例提供的一种基于知识驱动和数据驱动的分群设备的结构示意图, 该基于知识驱动和数据驱动的分群设备 800 可因配置或性能不同而产生比较大的差异, 可以包括一个或一个以上处理器 (central processing units, CPU) 801 (例如, 一个或一个以上处理器) 和存储器 809, 一个或一个以上存储应用程序 807 或数据 806 的存储介质 808 (例如一个或一个以上海量存储装置)。其中, 存储器 809 和存储介质 808 可以是短暂存储或持久存储。存储在存储介质 808 的程序可以包括一个或一个以上模块 (图示没标出), 每个模块可以包括对签到管理设备中的一系列指令操作。更进一步地, 处理器 801 可以设置为与存储介质 808 通信, 在基于知识驱动和数据驱动的分群设备 800 上执行存储介质 808 中的一系列指令操作。

基于知识驱动和数据驱动的分群设备 800 还可以包括一个或一个以上电源 802, 一个或一个以上有线或无线网络接口 803, 一个或一个以上输入输出接口 804, 和/或, 一个或一个以上操作系统 805, 例如 Windows Serve, Mac OS X, Unix, Linux, FreeBSD 等等。本领域技术人员可以理解, 图 8 中示出的基于知识驱动和数据驱动的分群设备结构并不构成对基于知识驱动和数据驱动的分群设备的限定, 可以包括比图示更多或更少的部件, 或者组合某些部件, 或者不同的部件布置。处理器 801 可以执行上述实施例中第一创建单元 601、识别单元 602、第二创建单元 603、第一分析单元 604、第一标记单元 605、第二分析单元 606、第二标记单元 607、第一输出单元 608、第二输出单元 609、分类单元 610 和第一调用单元 611 的功能。

下面结合图 8 对基于知识驱动和数据驱动的分群设备的各个构成部件进行具体的介绍：

处理器 801 是基于知识驱动和数据驱动的分群设备的控制中心，可以按照基于知识驱动和数据驱动的分群方法进行处理。处理器 801 利用各种接口和线路连接整个基于知识驱动和数据驱动的分群设备的各个部分，通过运行或执行存储在存储器 809 内的软件程序和/或模块，以及调用存储在存储器 809 内的数据，执行基于知识驱动和数据驱动的分群设备的各种功能和处理数据，从而实现提高案件检索系统的可用性的功能。存储介质 808 和存储器 809 都是存储数据的载体，本申请实施例中，存储介质 808 可以是指储存容量较小，但速度快的内存储器，而存储器 809 可以是储存容量大，但储存速度慢的外存储器。

存储器 809 可用于存储软件程序以及模块，处理器 801 通过运行存储在存储器 809 的软件程序以及模块，从而执行基于知识驱动和数据驱动的分群设备 800 的各种功能应用以及数据处理。存储器 809 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（获取待分类人群的特征数据等）等；存储数据区可存储根据签到管理设备的使用所创建的数据（通过目标知识分群决策树对特征数据进行分群，得到目标末端子节点对应的处理方案，并对处理方案的类别数量进行识别等）等。此外，存储器 809 可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。在本申请实施例中提供的基于知识驱动和数据驱动的分群方法程序和接收到的数据流存储在存储器中，当需要使用时，处理器 801 从存储器 809 中调用。

在计算机上加载和执行计算机程序指令时，全部或部分地产生按照本申请实施例的流程或功能。计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一计算机可读存储介质传输，例如，计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线（例如同轴电缆、光纤、双绞线）或无线（例如红外、无线、微波等）方式向另一个网站站点、计算机、服务器或数据中心进行传输。计算机可读存储介质可以是计算机能够存储的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。可用介质可以是磁性介质，（例如，软盘、硬盘、磁带）、光介质（例如，光盘）、或者半导体介质（例如固态硬盘（solid state disk，SSD））等。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

权 利 要 求

1、一种基于知识驱动和数据驱动的分群方法，其中，所述方法包括：

根据医学规则数据创建初始知识分群决策树，并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，所述医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，所述第一分群规则用于对临床上的相似特征人群进行分群，所述样本数据包括遵循医学知识规则的临床上的相似特征人群对应的临床结局数据；

获取待分类人群的特征数据，并通过所述目标知识分群决策树对所述特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对所述处理方案的类别数量进行识别，所述目标末端知识子节点为所述特征数据在所述目标知识分群决策树上的节点位置，所述处理方案用于指示对相似特征人群的解决渠道；

若识别到所述处理方案的类别数量为一个，则确定所述待分类人群属于第一特征相似人群，所述第一特征相似人群为所述目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群；

若识别到所述处理方案的类别数量为至少两个，则基于第二分群规则生成数据分群决策树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数，所述第二分群规则指示根据递归分区算法和处理方案的有效性进行分类；

若所述目标层数为第一层，则将所述待分类人群划分为所述第一特征相似人群；

若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群，所述第二特征相似人群为所述数据分群决策树中目标数据节点对应的处理方案所属的特征相似人群，每个目标数据节点对应不同的特征相似人群。

2、根据权利要求1所述的基于知识驱动和数据驱动的分群方法，其中，所述基于第二分群规则生成数据分群决策树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数，包括：

步骤一：将所述目标末端知识子节点作为所述数据分群决策树的数据根节点，计算所述数据根节点对应的样本数据中处理方案的有效性分值；

步骤二：根据每个处理方案的有效性分值对所述数据根节点进行分支，根据递归分区算法将每个处理方案依次划分到每个分支的数据子节点中，并继续对每个数据子节点对应的处理方案进行有效性分析；

步骤三：若当前数据子节点所对应的处理方案的有效性分值大于第一阈值，则对所述当前数据子节点进行分支；

步骤四：重复执行步骤三，直至每个数据子节点所对应的处理方案的有效性分值小于或等于所述第一阈值，得到数据分群决策树，所述数据分群决策树包括多个数据子节点，所述多个数据子节点中每个末端数据叶子节点对应有一个处理方案；

步骤五：确定每个处理方案对应的末端数据叶子节点在所述数据分群决策树中所处的目标层数。

3、根据权利要求2所述的基于知识驱动和数据驱动的分群方法，其中，所述将所述目标末端知识子节点作为所述数据分群决策树的数据根节点，计算所述数据根节点对应的处理方案的有效性分值，包括：

将所述目标末端知识子节点作为所述数据分群决策树的数据根节点；

获取所述数据根节点对应的样本数据中处理方案的目标特征信息，将所述目标特征信息拆分成因素因子；

计算所述因素因子在不同水平上的关联度、对预置影响因素的影响度以及所述因素因子不控制的边缘关联度；

通过预置的可压缩准则和可比较准则分析所述关联度、所述影响度和所述边缘关联度，得到混杂因子；

计算所述混杂因子与所述预置影响因素发生概率的回归系数的指数变换值，并以所述指数

变换值作为判定得分，得到所述数据根节点对应的处理方案的有效性分值，所述指数变换值的计算公式如下：

$$p(y = k|x; w) = \frac{\exp(w_k^T x)}{\sum_{i=1}^K \exp(w_i^T x)},$$

所述 $p(y = k|x; w)$ 为所述指数变换值，所述 w 为所述混杂因子与所述预置影响因素发生的概率，所述 x 为所述混杂因子，所述 k 为所述预置影响因素的类别数量，所述 y 为输出的预测值，所述 y 是 k 的条件概率，所述 K 为 k 的最大值，所述 T 为时间周期。

4、根据权利要求 2 所述的基于知识驱动和数据驱动的分群方法，其中，所述若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群，包括：

若所述目标层数不为第一层，判断所述目标层数对应处理方案的有效性分值是否大于第二阈值，所述第二阈值大于所述第一阈值；

若所述目标层数对应处理方案的有效性分值大于第二阈值，则将有效性分值大于第二阈值的处理方案所位于的数据叶子节点作为目标数据节点，并将所述待分类人群划分为所述目标数据节点对应的处理方案所属的特征相似人群，得到第二特征相似人群；

若所述目标层数对应处理方案的有效性分值小于或者等于第二阈值，则分析所述目标层数的上一层数对应处理方案的有效性分值；

若所述目标层数的上一层数对应处理方案的有效性分值大于第二阈值，则将所述目标层数的上一层数对应处理方案所位于的数据叶子节点作为目标数据节点，并将所述待分类人群划分为所述目标数据节点对应的处理方案所属的特征相似人群，得到第二特征相似人群。

5、根据权利要求 1 所述的基于知识驱动和数据驱动的分群方法，其中，所述根据医学规则数据创建初始知识分群决策树，并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树之前，所述方法还包括：

从多个系统、多个平台和多个医疗系统获取医学学科知识数据和临床数据，并对所述医学学科知识数据和所述临床数据进行数据处理，获得医学规则数据，所述数据处理包括数据清理处理、数据集成处理、数据规约处理、数据变换处理、特征提取处理、据脱敏处理和分类处理。

6、根据权利要求 1 所述的基于知识驱动和数据驱动的分群方法，其中，所述对所述处理方案的类别数量进行识别，包括：

计算所述处理方案之间的相似度获得多个相似度值，并比较多个所述相似度值是否均在预设范围内；

若比较结果为是，则统计所述处理方案的类别为一个；

若比较结果为否，则统计所述处理方案的类别为多个。

7、根据权利要求 1-6 中任一项所述的基于知识驱动和数据驱动的分群方法，其中，所述若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群之后，所述方法还包括：

调用预置的解释模型对所述第一特征相似人群或所述第二特征相似人群进行解释处理和可视化图表统计分析处理，得到可解释性特征相似人群，所述解释处理包括有效性排序处理、匹配度计算处理和匹配度排序处理。

8、一种基于知识驱动和数据驱动的分群装置，其中，所述装置包括：

第一创建单元，用于根据医学规则数据创建初始知识分群决策树，并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，所述医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，所述第一分群规则用于对临床上的相似特征人群进行分群，所述样本数据包括遵循医学知识规则的临床上的相似特征人群对应的临床结局数据；

识别单元，用于获取待分类人群的特征数据，并通过所述目标知识分群决策树对所述特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对所述处理方案的类别数量进行识别，所述目标末端知识子节点为所述特征数据在所述目标知识分群决策树上的节点位置，所

述处理方案用于指示对相似特征人群的解决渠道；

确定单元，用于若识别到所述处理方案的类别数量为一个，则确定所述待分类人群属于第一特征相似人群，所述第一特征相似人群为所述目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群；

5 第二创建单元，用于若识别到所述处理方案的类别数量为至少两个，则基于第二分群规则生成数据分群决策树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数，所述第二分群规则指示根据递归分区算法和处理方案的有效性进行分类各分支的末端子节点的；

10 第一划分单元，用于若所述目标层数为第一层，则将所述待分类人群划分为所述第一特征相似人群；

第二划分单元，用于若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群，所述第二特征相似人群为所述数据分群决策树中目标数据节点对应的处理方案所属的特征相似人群，每个目标数据节点对应不同的特征相似人群。

15 9、一种基于知识驱动和数据驱动的分群设备，其中，所述基于知识驱动和数据驱动的分群设备包括：

存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如下所述的基于知识驱动和数据驱动的分群方法的步骤：

20 根据医学规则数据创建初始知识分群决策树，并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，所述医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，所述第一分群规则用于对临床上的相似特征人群进行分群，所述样本数据包括遵循医学知识规则的临床上的相似特征人群对应的临床结局数据；

25 获取待分类人群的特征数据，并通过所述目标知识分群决策树对所述特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对所述处理方案的类别数量进行识别，所述目标末端知识子节点为所述特征数据在所述目标知识分群决策树上的节点位置，所述处理方案用于指示对相似特征人群的解决渠道；

若识别到所述处理方案的类别数量为一个，则确定所述待分类人群属于第一特征相似人群，所述第一特征相似人群为所述目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群；

30 若识别到所述处理方案的类别数量为至少两个，则基于第二分群规则生成数据分群决策树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数，所述第二分群规则指示根据递归分区算法和处理方案的有效性进行分类；

若所述目标层数为第一层，则将所述待分类人群划分为所述第一特征相似人群；

35 若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群，所述第二特征相似人群为所述数据分群决策树中目标数据节点对应的处理方案所属的特征相似人群，每个目标数据节点对应不同的特征相似人群。

10、根据权利要求9所述的基于知识驱动和数据驱动的分群设备，其中，所述基于第二分群规则生成数据分群决策树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数，包括以下步骤：

40 步骤一：将所述目标末端知识子节点作为所述数据分群决策树的数据根节点，计算所述数据根节点对应的样本数据中处理方案的有效性分值；

步骤二：根据每个处理方案的有效性分值对所述数据根节点进行分支，根据递归分区算法将每个处理方案依次划分到每个分支的数据子节点中，并继续对每个数据子节点对应的处理方案进行有效性分析；

45 步骤三：若当前数据子节点所对应的处理方案的有效性分值大于第一阈值，则对所述当前数据子节点进行分支；

步骤四：重复执行步骤三，直至每个数据子节点所对应的处理方案的有效性分值小于或等

于所述第一阈值,得到数据分群决策树,所述数据分群决策树包括多个数据子节点,所述多个数据子节点中每个末端数据叶子节点对应有一个处理方案;

步骤五:确定每个处理方案对应的末端数据叶子节点在所述数据分群决策树中所处的目标层数。

- 5 11、根据权利要求10所述的基于知识驱动和数据驱动的分群设备,其中,所述将所述目标末端知识子节点作为所述数据分群决策树的数据根节点,计算所述数据根节点对应的处理方案的有效性分值,包括以下步骤:

将所述目标末端知识子节点作为所述数据分群决策树的数据根节点;

- 10 获取所述数据根节点对应的样本数据中处理方案的目标特征信息,将所述目标特征信息拆分成因素因子;

计算所述因素因子在不同水平上的关联度、对预置影响因素的影响度以及所述因素因子不控制的边缘关联度;

通过预置的可压缩准则和可比较准则分析所述关联度、所述影响度和所述边缘关联度,得到混杂因子;

- 15 计算所述混杂因子与所述预置影响因素发生概率的回归系数的指数变换值,并以所述指数变换值作为判定得分,得到所述数据根节点对应的处理方案的有效性分值,所述指数变换值的计算公式如下:

$$p(y=k|x;w)=\frac{\exp(w_k^T x)}{\sum_{i=1}^K \exp(w_i^T x)},$$

- 20 所述 $p(y=k|x;w)$ 为所述指数变换值,所述 w 为所述混杂因子与所述预置影响因素发生的概率,所述 x 为所述混杂因子,所述 k 为所述预置影响因素的类别数量,所述 y 为输出的预测值,所述 y 是 k 的条件概率,所述 K 为 k 的最大值,所述 T 为时间周期。

12、根据权利要求10所述的基于知识驱动和数据驱动的分群设备,其中,所述若所述目标层数不为第一层,则将所述待分类人群划分为第二特征相似人群,包括以下步骤:

- 25 若所述目标层数不为第一层,判断所述目标层数对应处理方案的有效性分值是否大于第二阈值,所述第二阈值大于所述第一阈值;

若所述目标层数对应处理方案的有效性分值大于第二阈值,则将有效性分值大于第二阈值的处理方案所位于的数据叶子节点作为目标数据节点,并将所述待分类人群划分为所述目标数据节点对应的处理方案所属的特征相似人群,得到第二特征相似人群;

- 30 若所述目标层数对应处理方案的有效性分值小于或者等于第二阈值,则分析所述目标层数的上一层数对应处理方案的有效性分值;

若所述目标层数的上一层数对应处理方案的有效性分值大于第二阈值,则将所述目标层数的上一层数对应处理方案所位于的数据叶子节点作为目标数据节点,并将所述待分类人群划分为所述目标数据节点对应的处理方案所属的特征相似人群,得到第二特征相似人群。

- 35 13、根据权利要求9所述的基于知识驱动和数据驱动的分群设备,其中,所述根据医学规则数据创建初始知识分群决策树,并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点,得到目标知识分群决策树之前,还包括以下步骤:

从多个系统、多个平台和多个医疗系统获取医学学科知识数据和临床数据,并对所述医学学科知识数据和所述临床数据进行数据处理,获得医学规则数据,所述数据处理包括数据清理处理、数据集成处理、数据规约处理、数据变换处理、特征提取处理、据脱敏处理和分类处理。

- 40 14、根据权利要求9所述的基于知识驱动和数据驱动的分群设备,其中,所述对所述处理方案的类别数量进行识别,包括以下步骤:

计算所述处理方案之间的相似度获得多个相似度值,并比较多个所述相似度值是否均在预设范围内;

若比较结果为是,则统计所述处理方案的类别为一个;

- 45 若比较结果否,则统计所述处理方案的类别为多个。

15、根据权利要求 9-14 中任一项所述的基于知识驱动和数据驱动的分群设备，其中，所述若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群之后，还包括以下步骤：

调用预置的解释模型对所述第一特征相似人群或所述第二特征相似人群进行解释处理和可视化图表统计分析处理，得到可解释性特征相似人群，所述解释处理包括有效性排序处理、匹配度计算处理和匹配度排序处理。

16、一种计算机可读存储介质，其中，包括指令，当所述指令在计算机上运行时，使得计算机执行如下所述的基于知识驱动和数据驱动的分群方法的步骤：

根据医学规则数据创建初始知识分群决策树，并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树，所述医学规则数据用于指示对已归档的医学文献、医学共识、医学指南和专家的临床经验的医学学科知识数据进行规则提取处理所得的第一分群规则，所述第一分群规则用于对临床上的相似特征人群进行分群，所述样本数据包括遵循医学知识规则的临床上的相似特征人群对应的临床结局数据；

获取待分类人群的特征数据，并通过所述目标知识分群决策树对所述特征数据进行分群，得到目标末端知识子节点对应的处理方案，并对所述处理方案的类别数量进行识别，所述目标末端知识子节点为所述特征数据在所述目标知识分群决策树上的节点位置，所述处理方案用于指示对相似特征人群的解决渠道；

若识别到所述处理方案的类别数量为一个，则确定所述待分类人群属于第一特征相似人群，所述第一特征相似人群为所述目标知识分群决策树中目标末端知识子节点对应的处理方案所属的特征相似人群；

若识别到所述处理方案的类别数量为至少两个，则基于第二分群规则生成数据分群决策树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数，所述第二分群规则指示根据递归分区算法和处理方案的有效性进行分类；

若所述目标层数为第一层，则将所述待分类人群划分为所述第一特征相似人群；

若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群，所述第二特征相似人群为所述数据分群决策树中目标数据节点对应的处理方案所属的特征相似人群，每个目标数据节点对应不同的特征相似人群。

17、根据权利要求 16 所述的计算机可读存储介质，其中，所述基于知识驱动和数据驱动的分群的指令被计算机执行所述基于第二分群规则生成数据分群决策树，并确定每个处理方案对应所述数据分群决策树的数据子节点所处的目标层数的步骤时，包括以下步骤：

步骤一：将所述目标末端知识子节点作为所述数据分群决策树的数据根节点，计算所述数据根节点对应的样本数据中处理方案的有效性分值；

步骤二：根据每个处理方案的有效性分值对所述数据根节点进行分支，根据递归分区算法将每个处理方案依次划分到每个分支的数据子节点中，并继续对每个数据子节点对应的处理方案进行有效性分析；

步骤三：若当前数据子节点所对应的处理方案的有效性分值大于第一阈值，则对所述当前数据子节点进行分支；

步骤四：重复执行步骤三，直至每个数据子节点所对应的处理方案的有效性分值小于或等于所述第一阈值，得到数据分群决策树，所述数据分群决策树包括多个数据子节点，所述多个数据子节点中每个末端数据叶子节点对应有一个处理方案；

步骤五：确定每个处理方案对应的末端数据叶子节点在所述数据分群决策树中所处的目标层数。

18、根据权利要求 17 所述的计算机可读存储介质，其中，所述基于知识驱动和数据驱动的分群的指令被计算机执行所述将所述目标末端知识子节点作为所述数据分群决策树的数据根节点，计算所述数据根节点对应的处理方案的有效性分值的步骤时，包括以下步骤：

将所述目标末端知识子节点作为所述数据分群决策树的数据根节点；

获取所述数据根节点对应的样本数据中处理方案的目标特征信息，将所述目标特征信息拆

分成因素因子；

计算所述因素因子在不同水平上的关联度、对预置影响因素的影响度以及所述因素因子不控制的边缘关联度；

5 通过预置的可压缩准则和可比较准则分析所述关联度、所述影响度和所述边缘关联度，得到混杂因子；

计算所述混杂因子与所述预置影响因素发生概率的回归系数的指数变换值，并以所述指数变换值作为判定得分，得到所述数据根节点对应的处理方案的有效性分值，所述指数变换值的计算公式如下：

$$p(y = k|x; w) = \frac{\exp(w_k^T x)}{\sum_{i=1}^K \exp(w_i^T x)},$$

10 所述 $p(y = k|x; w)$ 为所述指数变换值，所述 w 为所述混杂因子与所述预置影响因素发生的概率，所述 x 为所述混杂因子，所述 k 为所述预置影响因素的类别数量，所述 y 为输出的预测值，所述 y 是 k 的条件概率，所述 K 为 k 的最大值，所述 T 为时间周期。

19、根据权利要求 17 所述的计算机可读存储介质，其中，所述基于知识驱动和数据驱动的分群的指令被计算机执行所述若所述目标层数不为第一层，则将所述待分类人群划分为第二特征相似人群的步骤时，包括以下步骤：

若所述目标层数不为第一层，判断所述目标层数对应处理方案的有效性分值是否大于第二阈值，所述第二阈值大于所述第一阈值；

20 若所述目标层数对应处理方案的有效性分值大于第二阈值，则将有效性分值大于第二阈值的处理方案所位于的数据叶子节点作为目标数据节点，并将所述待分类人群划分为所述目标数据节点对应的处理方案所属的特征相似人群，得到第二特征相似人群；

若所述目标层数对应处理方案的有效性分值小于或者等于第二阈值，则分析所述目标层数的上一层数对应处理方案的有效性分值；

25 若所述目标层数的上一层数对应处理方案的有效性分值大于第二阈值，则将所述目标层数的上一层数对应处理方案所位于的数据叶子节点作为目标数据节点，并将所述待分类人群划分为所述目标数据节点对应的处理方案所属的特征相似人群，得到第二特征相似人群。

20、根据权利要求 16 所述的计算机可读存储介质，其中，所述基于知识驱动和数据驱动的分群的指令被计算机执行所述根据医学规则数据创建初始知识分群决策树，并将样本数据填充至所述初始知识分群决策树各分支的末端知识子节点，得到目标知识分群决策树之前，还包括以下步骤：

30 从多个系统、多个平台和多个医疗系统获取医学学科知识数据和临床数据，并对所述医学学科知识数据和所述临床数据进行数据处理，获得医学规则数据，所述数据处理包括数据清理处理、数据集成处理、数据规约处理、数据变换处理、特征提取处理、据脱敏处理和分类处理。

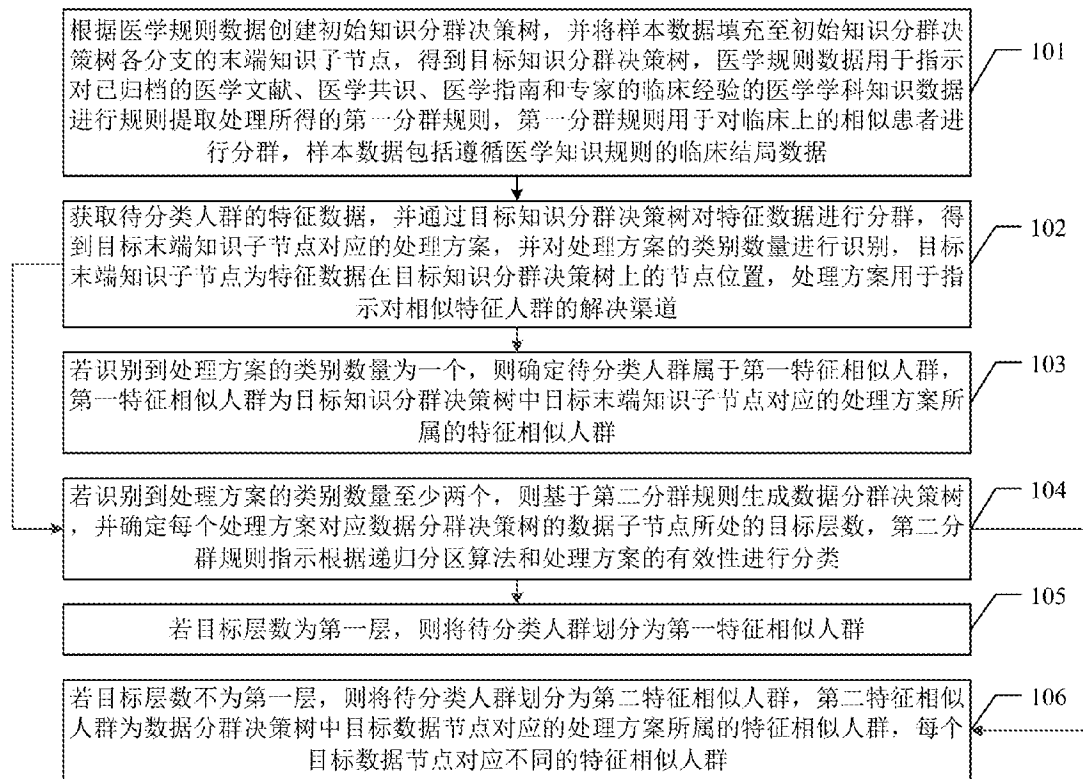


图 1

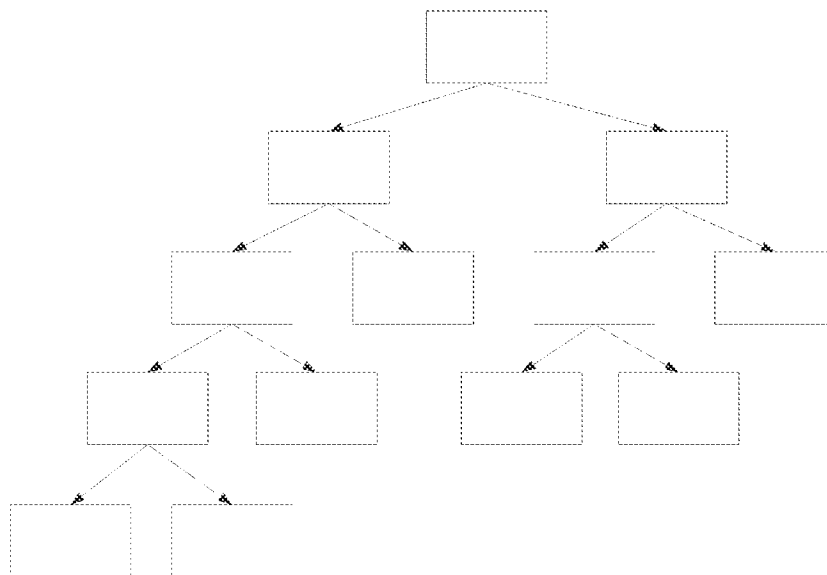


图 2

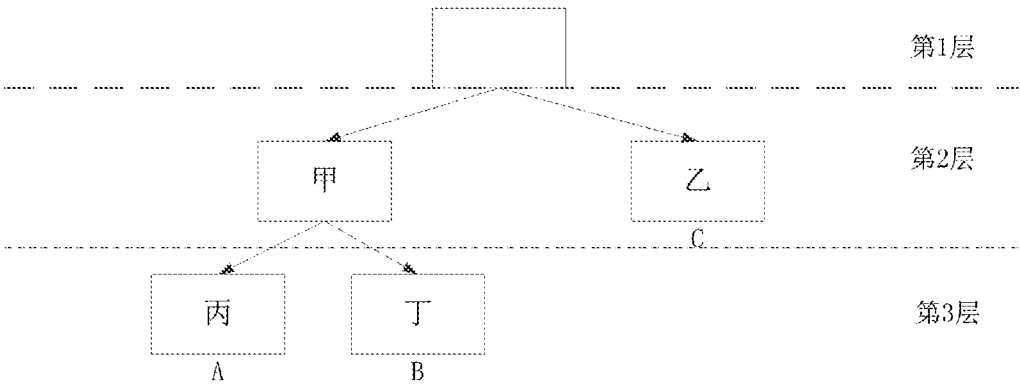


图 3

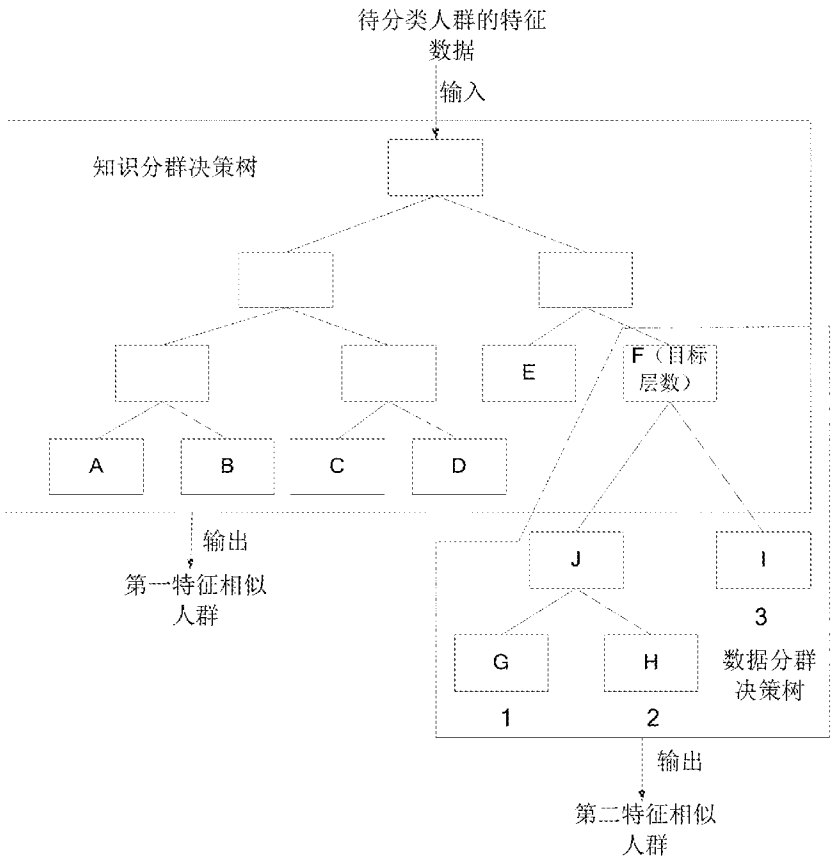


图 4

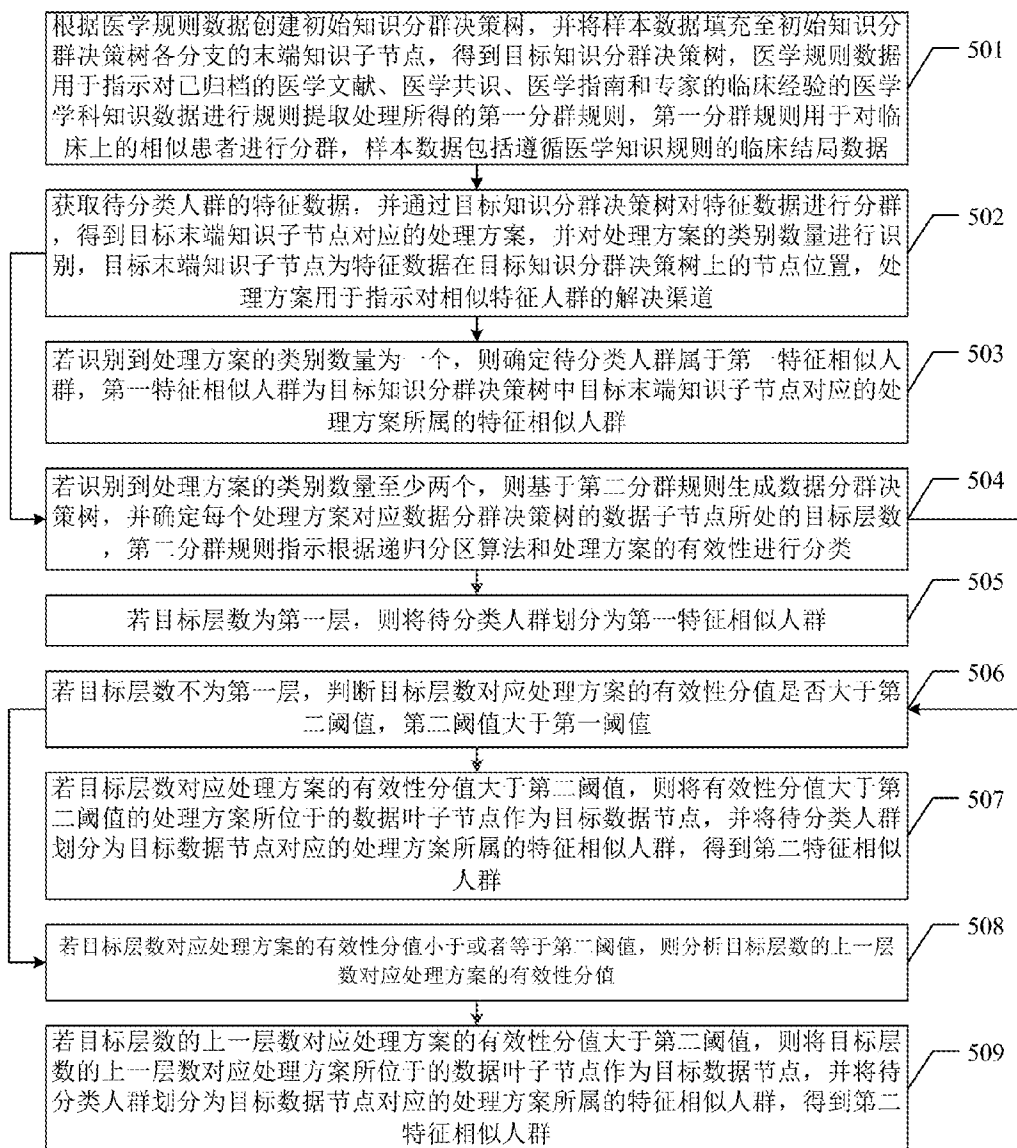


图 5

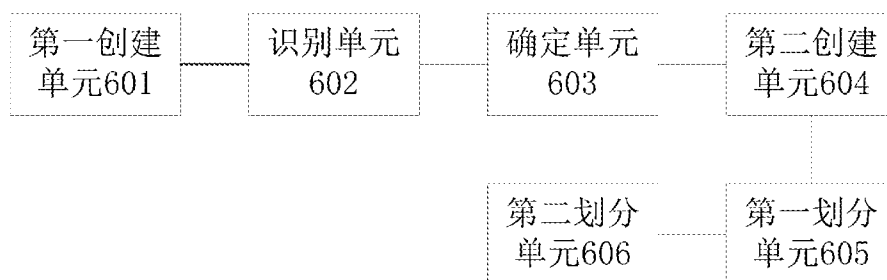


图 6



图 7

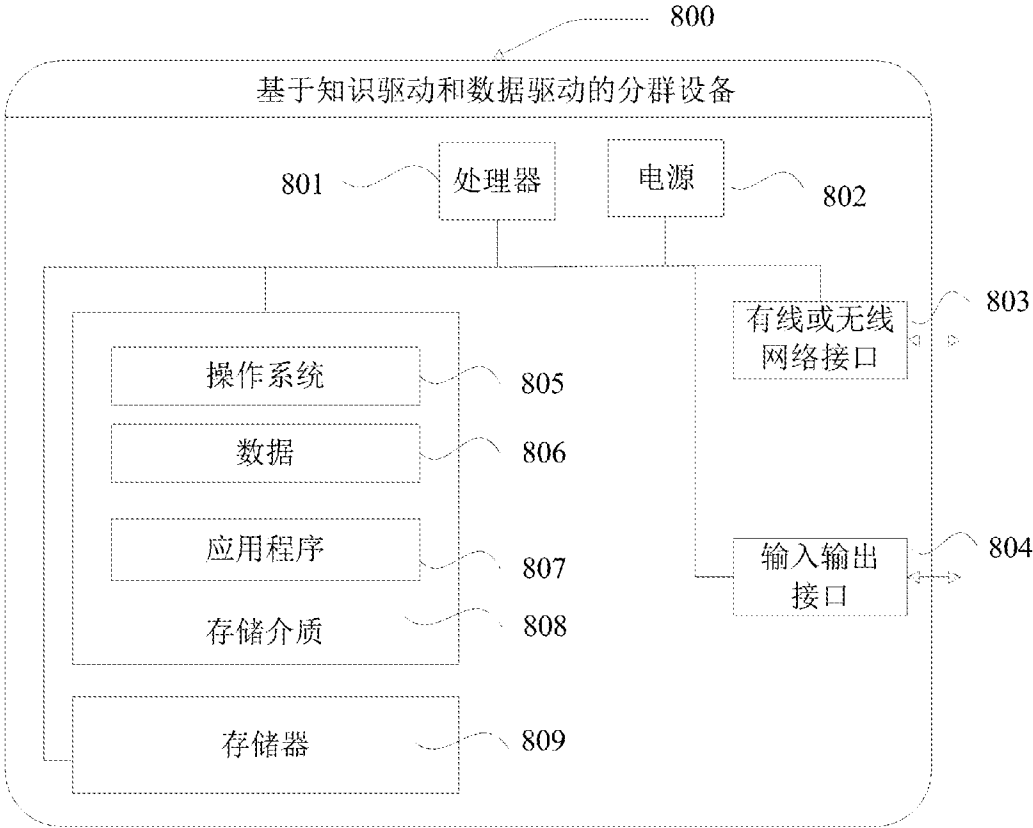


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/111413

A. CLASSIFICATION OF SUBJECT MATTER G06K 9/62(2006.01)i; G06N 5/02(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 9/, G06N 5/ 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) DWPI, CNABS, SIPOABS, CNTXT, USTXT, CNKI: 分群, 分类, 划分, 类别, 知识驱动, 数据驱动, 知识, 数据, 驱动, 医学, 医疗, 诊断, 规则, 决策树, 决策, 样本, 末端, 节点, 根节点, 子节点, 分支, 提取, 临床, 人群, 相似, 特征, 识别, 数量, 层, 递归, 有效性, 阈值, 因素因子, 关联度, 指数变换 clustering, grouping, divide, class, kind, sort, knowledge-driven, data-driven, knowledge, data, datum, drive, medical, iatrical, diagnosis, rule, decision tree, decision, sample, terminal, end, node, root node, son node, sub-node, branch, extract, abstract, clinical, crowd, similar, character, features, identify, number, layer, recursion, validity, threshold, factor, facil, degree of association, exponent transform																	
C. DOCUMENTS CONSIDERED TO BE RELEVANT <table border="1"> <thead> <tr> <th>Category*</th> <th>Citation of document, with indication, where appropriate, of the relevant passages</th> <th>Relevant to claim No.</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 110929752 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 27 March 2020 (2020-03-27) entire document</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 104699939 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 10 June 2015 (2015-06-10) entire document</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 108492887 A (HEFEI UNIVERSITY OF TECHNOLOGY) 04 September 2018 (2018-09-04) entire document</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 108984761 A (NANCHANG INSTITUTE OF TECHNOLOGY) 11 December 2018 (2018-12-11) entire document</td> <td>1-20</td> </tr> </tbody> </table>			Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.	PX	CN 110929752 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 27 March 2020 (2020-03-27) entire document	1-20	A	CN 104699939 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 10 June 2015 (2015-06-10) entire document	1-20	A	CN 108492887 A (HEFEI UNIVERSITY OF TECHNOLOGY) 04 September 2018 (2018-09-04) entire document	1-20	A	CN 108984761 A (NANCHANG INSTITUTE OF TECHNOLOGY) 11 December 2018 (2018-12-11) entire document	1-20
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.															
PX	CN 110929752 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 27 March 2020 (2020-03-27) entire document	1-20															
A	CN 104699939 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 10 June 2015 (2015-06-10) entire document	1-20															
A	CN 108492887 A (HEFEI UNIVERSITY OF TECHNOLOGY) 04 September 2018 (2018-09-04) entire document	1-20															
A	CN 108984761 A (NANCHANG INSTITUTE OF TECHNOLOGY) 11 December 2018 (2018-12-11) 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 24 November 2020		Date of mailing of the international search report 01 December 2020															
Name and mailing address of the ISA/CN China National Intellectual Property Administration (ISA/CN) 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/CN2020/111413**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2019272654 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 05 September 2019 (2019-09-05) entire document	1-20
A	US 2013226616 A1 (THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY) 29 August 2013 (2013-08-29) entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/111413

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110929752	A	27 March 2020	None			
CN	104699939	A	10 June 2015	US	2015161346	A1	11 June 2015
				DE	102014116177	A1	11 June 2015
CN	108492887	A	04 September 2018	CN	108492887	B	22 September 2020
CN	108984761	A	11 December 2018	None			
US	2019272654	A1	05 September 2019	US	10692254	B2	23 June 2020
US	2013226616	A1	29 August 2013	None			

国际检索报告

国际申请号

PCT/CN2020/111413

A. 主题的分类

G06K 9/62 (2006.01) i; G06N 5/02 (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 9/, G06N 5/

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))

DWPI, CNABS, SIPOABS, CNTXT, USTXT, CNKI: 分群, 分类, 划分, 类别, 知识驱动, 数据驱动, 知识, 数据, 驱动, 医学, 医疗, 诊断, 规则, 决策树, 决策, 样本, 末端, 节点, 根节点, 子节点, 分支, 提取, 临床, 人群, 相似, 特征, 识别, 数量, 层, 递归, 有效性, 阈值, 因素因子, 关联度, 指数变换 clustering, grouping, divide, class, kind, sort, knowledge-driven, data-driven, knowledge, data, datum, drive, medical, iatrical, diagnosis, rule, decision tree, decision, sample, terminal, end, node, root node, son node, sub-node, branch, extract, abstract, clinical, crowd, similar, character, features, identify, number, layer, recursion, validity, threshold, factor, facil, degree of association, exponent transform

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 110929752 A (平安科技深圳有限公司) 2020年 3月 27日 (2020 - 03 - 27) 全文	1-20
A	CN 104699939 A (国际商业机器公司) 2015年 6月 10日 (2015 - 06 - 10) 全文	1-20
A	CN 108492887 A (合肥工业大学) 2018年 9月 4日 (2018 - 09 - 04) 全文	1-20
A	CN 108984761 A (南昌工程学院) 2018年 12月 11日 (2018 - 12 - 11) 全文	1-20
A	US 2019272654 A1 (INT BUSINESS MACHINES CORP) 2019年 9月 5日 (2019 - 09 - 05) 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

2020年 11月 24日

国际检索报告邮寄日期

2020年 12月 1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

谭毅

电话号码 62412070

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
A	US 2013226616 A1 (UNIV LELAND STANFORD JUNIOR) 2013年 8月 29日 (2013 - 08 - 29) 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/CN2020/111413

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	110929752	A	2020年 3月 27日	无			
CN	104699939	A	2015年 6月 10日	US	2015161346	A1	2015年 6月 11日
				DE	102014116177	A1	2015年 6月 11日
CN	108492887	A	2018年 9月 4日	CN	108492887	B	2020年 9月 22日
CN	108984761	A	2018年 12月 11日	无			
US	2019272654	A1	2019年 9月 5日	US	10692254	B2	2020年 6月 23日
US	2013226616	A1	2013年 8月 29日	无			