

(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 3 月 5 日 (05.03.2020)



(10) 国际公布号
WO 2020/042579 A1

(51) 国际专利分类号:
G06K 9/62 (2006.01)

(21) 国际申请号: PCT/CN2019/077223

(22) 国际申请日: 2019 年 3 月 6 日 (06.03.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810983116.X 2018年8月27日 (27.08.2018) CN

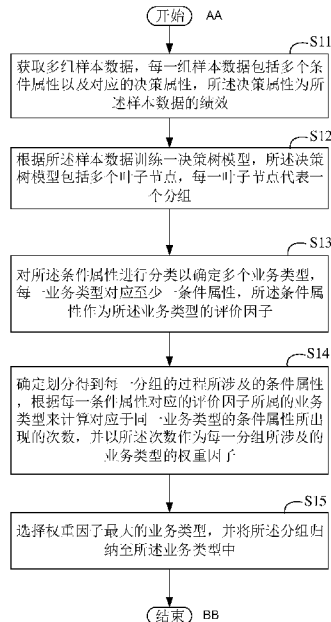
(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 邓悦(DENG, Yue); 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)。 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: 深圳市赛恩倍吉知识产权代理有限公司(SHENZHEN SCIENBIZIP INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国广东省深

(54) Title: GROUP CLASSIFICATION METHOD AND DEVICE, ELECTRONIC DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 分组归纳方法、装置、电子装置及存储介质



S11 Acquire multiple groups of sample data, each group of sample data comprising multiple condition attributes and corresponding decision attributes, and the decision attributes being performance of the sample data

S12 Train a decision tree model according to the sample data, the decision tree model comprising multiple leaf nodes, and each leaf node representing a group

S13 Classify the condition attributes to determine multiple service types, each service type corresponding to at least one condition attribute, and the condition attribute serving as an evaluation factor of the service type

S14 Determine a condition attribute involved in a process of acquiring each group by means of division, calculate, according to the service type to which an evaluation factor corresponding to each condition attribute belongs, the number of times a condition attribute corresponding to the same service type occurs, and using the number of times as a weighting factor of a service type related to each group

S15 Select a service type having the largest weighting factor, and classify the group under the service type

AA Start

BB End

图 1

(57) Abstract: A group classification method comprises: acquiring multiple groups of sample data; training a decision tree model, the decision tree model comprising multiple leaf nodes; classifying condition attributes so as to determine multiple service types, each service type corresponding to at least one condition attribute, and the condition attribute serving as an evaluation factor of the service type; determining a condition attribute related to a process of acquiring each group by means of division, calculating, according to the service type to which an evaluation factor corresponding to each condition attribute belongs, the number of times a condition attribute

圳市龙华新区龙观东路83号荣群大厦9楼, Guangdong 518109 (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, 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))。

corresponding to the same service type occurs, and using the number of times as a weighting factor of a service type related to each group; and selecting a service type having the largest weighting factor, and classifying the group under the service type. A group classification device, an electronic device, and a storage medium are also disclosed. The method has uniform selection criteria, thereby improving the efficiency of data analysis in a sample processing process, and facilitating service analysis.

(57) 摘要: 一种分组归纳方法, 包括: 获取多组样本数据; 训练决策树模型, 所述决策树模型包括多个叶子节点; 对所述条件属性进行分类以确定多个业务类型, 每一业务类型对应至少一条件属性, 所述条件属性作为所述业务类型的评价因子; 确定划分得到每一分组的过程所涉及的条件属性, 根据每一条件属性对应的评价因子所属的业务类型来计算对应于同一业务类型的条件属性所出现的次数, 并以所述次数作为每一分组所涉及的业务类型的权重因子; 选择权重因子最大的业务类型, 并将所述分组归纳至所述业务类型中。还公开一种分组归纳装置、电子装置及存储介质。本方法归纳标准统一, 有利于提高样本处理过程中数据分析的效率, 而且有利于辅助业务分析。

分组归纳方法、装置、电子装置及存储介质

本申请要求于2018年08月27日提交中国专利局,申请号为201810983116.X申请名称为“分组归纳方法及装置、电子装置及计算机可读存储介质”的中国专利申请的优先权,其全部内容通过引用结合在本申请中。

技术领域

本申请涉及一种分组归纳方法、分组归纳装置、电子装置以及存储介质。

背景技术

现有的分群归纳方法基本采用人工基于业务类型(如:勤奋型、资源型等)进行归纳。然而,该方法会受到个人主观的影响,并且当分群数量较多,每个分群涉及的特征量较大时,人工无法有效的完成对分群的归纳。

发明内容

鉴于以上内容,有必要提出一种分组归纳方法、分组归纳装置、电子装置以及存储介质,能够解决以上问题。

本申请一较佳实施方式提供一种分组归纳方法,包括:获取多组样本数据,每一组样本数据包括多个条件属性以及对应的决策属性;根据所述样本数据训练一决策树模型,所述决策树模型包括多个叶子节点,每一叶子节点代表一个分组;对所述条件属性进行分类以确定多个业务类型,每一业务类型对应至少一条件属性,所述条件属性作为所述业务类型的评价因子;确定划分得到每一分组的过程所涉及的条件属性,根据每一条件属性对应的评价因子所属的业务类型来计算对应于同一业务类型的条件属性所出现的次数,并以所述次数作为每一分组所涉及的业务类型的权重因子;以及选择权重因子最大的业务类型,并将所述分组归纳至所述业务类型中。

本申请一较佳实施方式还提供一种分组归纳装置,包括:获取模块,用于获取多组样本数据,每一组样本数据包括多个条件属性以及对应的决策属性;训练模块,用于根据所述样本数据训练一决策树模型,所述决策树模型包括多个叶子节点,每一叶子节点代表一个分组;分类模块,用于对所述条件属性进行分类以确定多个业务类型,每一业务类型对应至少一条件属性,所述条件属性作为所述业务类型的评价因子;计算模块,用于确定划分得到每一分组的过程所涉及的条件属性,根据每一条件属性对应的评价因子所属的业务类型来计算对应于同一业务类型的条件属性所出现的次数,并以所述次数作为每一分组所涉及的业务类型的权重因子;以及归纳模块,用于选择权重因子最大的业务类型,并将所述分组归纳至所述业务类型中。

本申请一较佳实施方式还提供一种电子装置，包括处理器和存储器，所述存储器中存储有分组归纳程序，所述处理器用于执行所述分组归纳程序以实现如前所述的分组归纳方法。

本申请一较佳实施方式还提供一种非易失性可读存储介质，所述非易失性可读存储介质上存储有分组归纳程序，所述分组归纳程序被处理器执行时实现如前所述的分组归纳方法。

本申请实施例基于每一分组涉及的条件属性在同一业务类型中出现的次数对每一分组进行客观归纳，使得归纳标准统一；而且，将分组结果与业务的需求类型对应起来，有利于辅助业务分析。

附图说明

图 1 是本申请一较佳实施例提供的分组归纳方法的流程图。

图 2 是图 1 的分组归纳方法所训练的决策树的示意图。

图 3 是本申请一较佳实施例提供的分组归纳装置的结构示意图。

图 4 为本申请一较佳实施例提供的电子装置的结构示意图。

具体实施方式

为了能够更清楚地理解本申请的上述目的、特征和优点，下面结合附图和具体实施例对本申请进行详细描述。需要说明的是，在不冲突的情况下，本申请的实施例及实施例中的特征可以相互组合。

在下面的描述中阐述了很多具体细节以便于充分理解本申请，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中在本申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请。

图 1 是本申请一较佳实施例提供的分组归纳方法的流程图。所述分组归纳方法应用于一电子装置 1 中。根据不同需求，所述分组归纳方法的步骤顺序可以改变，某些步骤可以省略或合并。所述分组归纳方法包括以下步骤：

步骤 S11，获取多组样本数据，每一组样本数据包括多个条件属性以及对应的决策属性，所述决策属性为所述样本数据的绩效。

其中，所述样本数据需同时包括绩效较优的人员(即，绩优人员)和绩效较差的人员(即，绩差人员)的数据，所述样本数据可存储于所述电子装置中。在另一实施方式中，所述电子装置还可以通过接入网络连接之一外部样本库，进而获取所述样本库存储的样本数据。在其它实施方式中，所述电子装置还可以通过大数据方式来收集并建立所述样本库。

在本实施方式中，所述条件属性可为行为轨迹(如，出差情况)、app 活跃情况、业务扩展情况、消费情况、兴趣爱好、参加培训情况、考勤率等。以所述条件属性包括出差情况、业务扩展情况、参加培训情况以及考勤率为例

进行说明，所述样本数据如表 1 所示。

表 1 样本数据

条件属性				决策属性
出差情况(C)	业务扩展情况(Y)	参加培训情况(P)	考勤率(K)	绩效
每月出差 0 次	每月增加 0 件业务	每月 0 次	<80%	差
每月出差 0 次	每月增加 0 件业务	每月 0 次	$\geq 80\%$	优
每月出差 0 次	每月增加 1 件业务	每月 1 次	$\geq 80\%$	优
每月出差 0 次	每月增加 ≥ 2 件业务	每月 0 次	$\geq 80\%$	优
每月出差 0 次	每月增加 0 件业务	每月 ≥ 2 次	$\geq 80\%$	优
每月出差 1 次	每月增加 0 件业务	每月 0 次	<80%	差
每月出差 1 次	每月增加 1 件业务	每月 0 次	<80%	优
每月出差 1 次	每月增加 0 件业务	每月 1 次	<80%	差
每月出差 1 次	每月增加 1 件业务	每月 0 次	$\geq 80\%$	优
每月出差 ≥ 2 次	每月增加 1 件业务	每月 0 次	<80%	差
每月出差 ≥ 2 次	每月增加 1 件业务	每月 1 次	<80%	优
每月出差 ≥ 2 次	每月增加 1 件业务	每月 1 次	$\geq 80\%$	优
每月出差 ≥ 2 次	每月增加 0 件业务	每月 ≥ 2 次	<80%	优
每月出差 ≥ 2 次	每月增加 0 件业务	每月 1 次	$\geq 80\%$	优
每月出差 ≥ 2 次	每月增加 0 件业务	每月 1 次	$\geq 80\%$	优
每月出差 ≥ 2 次	每月增加 0 件业务	每月 0 次	<80%	差

步骤 S12，根据所述样本数据训练一决策树模型，所述决策树模型包括多个叶子节点，每一叶子节点代表一个分组。

其中，决策树算法属于监督学习分类算法，而决策树模型代表对象属性与对象值之间的一种映射关系。其中，所述决策树模型具有 N 层(N 为自然数， $N>2$)，所述电子装置根据所述条件属性设置所述决策树模型中的每个节点(包括根节点、内部节点和叶子节点)的值。所述决策树模型的第一层为根节点，所述决策树模型的第二层代表以第一级条件属性对所述根节点进行分割得到的多个节点，所述决策树模型的第三层代表以第二级条件属性对上一层节点进行分割得到的多个节点，等等。如图 2 所示，所述决策树模型包括四层，即， $N=4$ 。

在本实施方式中，训练所述决策树模型具体包括：

步骤 S121，以所述样本数据作为训练集计算每一条件属性的信息增益 (information gain)。

步骤 S122，选择信息增益最大的条件属性作为所述决策树模型的根节点以分割所述样本数据以得到下一层节点；其中，某一条件属性的信息增益越大，说明选择所述条件属性对分类提供的信息越多，越有利于确定性，越有利于将所述样本数据进行分类。如图 2 所示，若计算信息增益最大的条件属性为“出差情况”，则选择“出差情况”条件属性作为根节点以分割所述样本数据。

步骤 S123，以每一节点包含的样本数据作为训练集重新计算每一条件属性的信息增益。

步骤 S124，选择信息增益最大的条件属性分割所述节点以得到下一层节点。如图 2 所示，对于包含每月出差情况为 0 次 ($C=0$) 的样本的节点，若计算信息增益最大的条件属性为“业务拓展情况”，则根据“业务拓展情况”条件属性对所述节点进行分割而获得下一层节点。对于包含每月出差情况为 ≥ 2 次 ($C \geq 2$) 的样本的节点，若计算信息增益最大的条件属性为“参加培训情况”，则根据“参加培训情况”条件属性对所述节点进行分割而获得下一层节点。实际训练过程中，对位于同一层的多个节点进行分割所采用的条件属性通常不同。每一次分割后形成的节点比上一层节点的数据纯度更高。

步骤 S125，递归执行步骤 S123 以及 S124，直至分割停止。此时，每一叶节点代表一个分组，每一分组包含的绩优人员数量与绩差人员数量之间具有固定的比值，而多个分组包含的绩优人员数量与绩差人员数量之间的比值可以互不相同。如，叶子节点“分组 1”的所述比值可为 1:8。当然，某一叶节点包含的可均为绩优人员，也可以均为绩差人员。其中，每一组样本数据只能被划分至其中一个叶子节点中，即，每一组样本数据不能同属于两个及两个以上的叶子节点。

在本实例中，在所述决策树的层数达到一预设层数 (如：4 层) 时停止下一次分割。在另一实施例，在所述决策树的当前层的每一节点包含的样本数量小于一预设数量时停止下一次分割。

其中，每一条件属性的信息增益 $Gain(S, A)$ 可通过如下方式计算：

$$Gain(S, A) = Entropy(S) - Entropy(S, A)$$

其中， $Gain(S, A)$ 表示 A 条件属性在训练集 S 上的信息增益， $Entropy(S)$ 表示训练集 S 的信息熵， $Entropy(S, A)$ 表示 A 属性的信息熵。

例如，如表 1 所示，在步骤 S121 中，所述训练集共有样本数据 16 组，绩优的有 11 组，绩差的有 5 组。以“出差情况”条件属性的信息增益计算为例进行说明，其它条件属性的信息增益计算过程同理。对于 $C=0$ 的情况，共有样本数据 5 组，绩优的有 4 组，绩差的有 1 组；对于 $C=1$ 的情况，共有样本数据 4 组，绩优的有 2 组，绩差的有 2 组；对于 $C \geq 2$ 的情况，共有样本数据 7 组，绩优的有 5 组，绩差的有 2 组。因此，“出差情况”条件属性的信息

增益的计算方式如下：

$$\begin{aligned}
 Entropy(S) &= -\frac{11}{16} \log_2 \frac{11}{16} - \frac{5}{16} \log_2 \frac{5}{16} \\
 Entropy(A) &= \frac{4}{16} \times (-\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5}) + \frac{4}{16} \times (-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4}) \\
 &\quad + \frac{7}{16} \times (-\frac{5}{7} \log_2 \frac{5}{7} - \frac{2}{7} \log_2 \frac{2}{7}) \\
 Gain(S, A) &= -\frac{11}{16} \log_2 \frac{11}{16} - \frac{5}{16} \log_2 \frac{5}{16} - \frac{4}{16} \times (-\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5}) \\
 &\quad - \frac{4}{16} \times (-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4}) - \frac{7}{16} \times (-\frac{5}{7} \log_2 \frac{5}{7} - \frac{2}{7} \log_2 \frac{2}{7})
 \end{aligned}$$

步骤 S13，对所述条件属性进行分类以确定多个业务类型，每一业务类型对应至少一条件属性，所述条件属性作为所述业务类型的评价因子。

所述业务类型可包括资源型、勤奋型以及开放型等。资源型可以是指业务能力强、工作能力强的人员。勤奋型可以是指学习时间长、每天工作时长的人员，开放型可以是指性格活跃、乐于社交的人员。其中，所述评价因子是能够表征所述业务类型的重要特征的评价指标。每一业务类型可对应一个评价因子，也可对应至少两个评价因子。

例如，对于“资源型”业务类型来说，对应的评价因子可为业务拓展情况等。对于“勤奋型”业务类型来说，对应的评价因子可为参加培训情况、考勤率等。对于“开放型”业务类型来说，对应的评价因子可为行为轨迹(如，出差情况)、app 活跃情况等。

其中，所述业务类型的数量为 $M(M>1, M$ 为自然数)，所述分组的数量为 $N(N>1, N$ 为自然数)。M 可以等于 N，也可以不等于 N。

步骤 S14，确定划分得到每一分组的过程所涉及的条件属性，根据每一条件属性对应的评价因子所属的业务类型来计算对应于同一业务类型的条件属性所出现的次数，并以所述次数作为每一分组所涉及的业务类型的权重因子。

例如，对于“分组 8”所包含的样本数据来说，所涉及的条件属性为：“出差情况”、“参加培训情况”以及“考勤率”。“出差情况”条件属性对应的评价因子所属的业务类型为“开放型”，“参加培训情况”以及“考勤率”对应的评价因子所属的业务类型为“勤奋型”。因此，对于“分组 8”所涉及的条件属性来说，对应于“开放型”业务类型的条件属性出现 1 次，“开放型”业务类型的权重因子为 1；而对应于“勤奋型”业务类型的条件属性出现 2 次，“勤奋型”业务类型的权重因子为 2。

步骤 S15，选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中。

例如，若“分组 8”所包含的样本数据中，“开放型”业务类型的权重因子为 1，而“勤奋型”业务类型的权重因子为 2，则将所述“分组 8”归纳至“勤奋型”业务类型中。

在本实施方式中，所述选择权重因子最大的业务类型，并将所述分组归

纳至所述业务类型中包括：

步骤 S151，当选择出权重因子最大的业务类型时，判断所述业务类型的个数。

步骤 S152，当权重因子最大的业务类型仅为一个时，直接将所述分组归纳至所述业务类型中；当权重因子最大的业务类型包括至少两个时，随机将所述分组归纳至其中一业务类型中。当然，在其它实施方式中，当权重因子最大的业务类型包括至少两个时，由于所述分组同时符合至少两个业务类型的特征，因此，将所述分组同时分配至所述不同的业务类型中。

图 3 为本申请一较佳实施方式提供的分组归纳装置 300 的结构示意图。在一些实施例中，所述分组归纳装置 300 运行于电子装置中。所述分组归纳装置 300 可以包括多个由程序代码段所组成的功能模块。所述分组归纳装置 300 的各个程序段的程序代码可以存储于电子装置的存储器中，并由所述至少一个处理器所执行，以实现分组归纳功能。

本实施例中，所述分组归纳装置 300 根据其所执行的功能，可以被划分为多个功能模块。如图 3 所示，所述分组归纳装置 300 包括：获取模块 301、训练模块 302、分类模块 303、计算模块 304 以及归纳模块 305。本申请所称的模块是指一种能够被至少一个处理器所执行并且能够完成固定功能的一系列计算机可读指令段，其存储在存储器中。在本实施例中，关于各模块的功能将在后续的实施例中详述。

所述获取模块 301 用于获取多组样本数据，每一组样本数据包括多个条件属性以及对应的决策属性，所述决策属性为所述样本数据的绩效。

其中，所述样本数据需同时包括绩效较优的人员(即，绩优人员)和绩效较差的人员(即，绩差人员)的数据，所述样本数据可存储于所述电子装置中。在另一实施方式中，所述电子装置还可以通过接入网络连接之一外部样本库，进而来获取所述样本库存储的样本数据。在其它实施方式中，所述电子装置还可以通过大数据方式来收集并建立所述样本库。

在本实施方式中，所述条件属性可为行为轨迹(如，出差情况)、app 活跃情况、业务扩展情况、消费情况、兴趣爱好、参加培训情况、考勤率等。以所述条件属性包括出差情况、业务扩展情况、参加培训情况以及考勤率为例进行说明，所述样本数据如上表 1 所示。

所述训练模块 302 用于根据所述样本数据训练一决策树模型，所述决策树模型包括多个叶子节点，每一叶子节点代表一个分组。

其中，决策树算法属于监督学习分类算法，而决策树模型代表对象属性与对象值之间的一种映射关系。其中，所述决策树模型具有 N 层(N 为自然数， $N > 2$)，所述训练模块 302 根据所述条件属性设置所述决策树模型中的每个节点(包括根节点、内部节点和叶子节点)的值。所述决策树模型的第一层为根节点，所述决策树模型的第二层代表以第一级条件属性对所述根节点进行分割得到的多个节点，所述决策树模型的第三层代表以第二级条件属性对上一层节点进行分割得到的多个节点，等等。如图 2 所示，所述决策树模型

包括四层，即， $N=4$ 。

在本实施方式中，所述训练模块 302 以所述样本数据作为训练集计算每一条件属性的信息增益(information gain)，并选择信息增益最大的条件属性作为所述决策树模型的根节点以分割所述样本数据以得到下一层节点。其中，某一条件属性的信息增益越大，说明选择所述条件属性对分类提供的信息越多，越有利于确定性，越有利于将所述样本数据进行分类。如图 2 所示，若计算信息增益最大的条件属性为“出差情况”，则选择“出差情况”条件属性作为根节点以分割所述样本数据。

所述训练模块 302 进一步以每一节点包含的样本数据作为训练集重新计算每一条件属性的信息增益，并选择信息增益最大的条件属性分割所述节点以得到下一层节点。如图 2 所示，对于包含每月出差情况为 0 次($C=0$)的样本的节点，若计算信息增益最大的条件属性为“业务拓展情况”，则根据“业务拓展情况”条件属性对所述节点进行分割而获得下一层节点。对于包含每月出差情况为 ≥ 2 次($C \geq 2$)的样本的节点，若计算信息增益最大的条件属性为“参加培训情况”，则根据“参加培训情况”条件属性对所述节点进行分割而获得下一层节点。实际训练过程中，对位于同一层的多个节点进行分割所采用的条件属性通常不同。每一次分割后形成的节点比上一层节点的数据纯度更高。

所述训练模块 302 进一步递归执行每一条件属性的信息增益的计算步骤以及选择信息增益最大的条件属性分割所述节点以得到下一层节点的步骤，直至分割停止。此时，每一叶节点代表一个分组，每一分组包含的绩优人员数量与绩差人员数量之间具有固定的比值，而多个分组包含的绩优人员数量与绩差人员数量之间的比值可以互不相同。如，叶子节点“分组 1”的所述比值可为 1:8。当然，某一叶节点包含的可均为绩优人员，也可以均为绩差人员。其中，每一组样本数据只能被划分至其中一个叶子节点中，即，每一组样本数据不能同属于两个及两个以上的叶子节点。

在本实例中，在所述决策树的层数达到一预设层数(如：4 层)时停止下一次分割。在另一实施例，在所述决策树的当前层的每一节点包含的样本数量小于一预设数量时停止下一次分割。

其中，每一条件属性的信息增益 $Gain(S, A)$ 可通过如下方式计算：

$$Gain(S, A) = Entropy(S) - Entropy(S, A)$$

其中， $Gain(S, A)$ 表示 A 条件属性在训练集 S 上的信息增益， $Entropy(S)$ 表示训练集 S 的信息熵， $Entropy(S, A)$ 表示 A 属性的信息熵。

例如，如表 1 所示，在所述训练模块 302 选择根节点的过程中，所述训练集共有样本数据 16 组，绩优的有 11 组，绩差的有 5 组。以“出差情况”条件属性的信息增益计算为例进行说明，其它条件属性的信息增益计算过程同理。对于 $C=0$ 的情况，共有样本数据 5 组，绩优的有 4 组，绩差的有 1 组；对于 $C=1$ 的情况，共有样本数据 4 组，绩优的有 2 组，绩差的有 2 组；对于 $C \geq 2$ 的情况，共有样本数据 7 组，绩优的有 5 组，绩差的有 2 组。因此，“出差情况”条件属性的信息增益的计算方式如下：

$$\begin{aligned}
 Entropy(S) &= -\frac{11}{16}\log_2\frac{11}{16} - \frac{5}{16}\log_2\frac{5}{16} \\
 Entropy(A) &= \frac{4}{16}\times(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}) + \frac{4}{16}\times(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}) \\
 &\quad + \frac{7}{16}\times(-\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7}) \\
 Gain(S,A) &= -\frac{11}{16}\log_2\frac{11}{16} - \frac{5}{16}\log_2\frac{5}{16} - \frac{4}{16}\times(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}) \\
 &\quad - \frac{4}{16}\times(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}) - \frac{7}{16}\times(-\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7})
 \end{aligned}$$

所述分类模块 303 用于对所述条件属性进行分类以确定多个业务类型，每一业务类型对应至少一条件属性，所述条件属性作为所述业务类型的评价因子。

所述业务类型可包括资源型、勤奋型以及开放型等。资源型可以是指业务能力强、工作能力强的人员。勤奋型可以是指学习时间长、每天工作时长的人员，开放型可以是指性格活跃、乐于社交的人员。其中，所述评价因子是能够表征所述业务类型的重要特征的评价指标。每一业务类型可对应一个评价因子，也可对应至少两个评价因子。

例如，对于“资源型”业务类型来说，对应的评价因子可为业务拓展情况等。对于“勤奋型”业务类型来说，对应的评价因子可为参加培训情况、考勤率等。对于“开放型”业务类型来说，对应的评价因子可为行为轨迹(如，出差情况)、app 活跃情况等。

其中，所述业务类型的数量为 $M(M>1, M$ 为自然数)，所述分组的数量为 $N(M>1, M$ 为自然数)。M 可以等于 N，也可以不等于 N。

所述计算模块 304 用于确定划分得到每一分组的过程所涉及的条件属性，根据每一条件属性对应的评价因子所属的业务类型来计算对应于同一业务类型的条件属性所出现的次数，并以所述次数作为每一分组所涉及的业务类型的权重因子。

例如，对于“分组 8”所包含的样本数据来说，所涉及的条件属性为：“出差情况”、“参加培训情况”以及“考勤率”。“出差情况”条件属性对应的评价因子所属的业务类型为“开放型”，“参加培训情况”以及“考勤率”对应的评价因子所属的业务类型为“勤奋型”。因此，对于“分组 8”所涉及的条件属性来说，对应于“开放型”业务类型的条件属性出现 1 次，“开放型”业务类型的权重因子为 1；而对应于“勤奋型”业务类型的条件属性出现 2 次，“勤奋型”业务类型的权重因子为 2。

所述归纳模块 305 用于选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中。

例如，若“分组 8”所包含的样本数据中，“开放型”业务类型的权重因子为 1，而“勤奋型”业务类型的权重因子为 2，则将所述“分组 8”归纳至“勤奋型”业务类型中。

在本实施方式中，当所述归纳模块 305 选择出权重因子最大的业务类型

时,判断所述业务类型的个数。当权重因子最大的业务类型仅为一个时,所述归纳模块 305 直接将所述分组归纳至所述业务类型中。当权重因子最大的业务类型包括至少两个时,所述归纳模块 305 随机将所述分组归纳至其中一业务类型中。当然,在其它实施方式中,当权重因子最大的业务类型包括至少两个时,由于所述分组同时符合至少两个业务类型的特征,因此,所述归纳模块 305 将所述分组同时分配至所述不同的业务类型中。

如图 4 所示,图 4 是本申请一较佳实施方式中实现所述分组归纳方法的电子装置 1 的结构示意图。所述电子装置 1 包括存储器 101、处理器 102 以及存储于所述存储器 101 中并可在所述处理器 102 上运行的计算机可读指令 103,例如分组归纳程序。

所述处理器 102 执行所述计算机可读指令 103 时实现上述实施例中分组归纳方法的步骤:

步骤 S11,获取多组样本数据,每一组样本数据包括多个条件属性以及对应的决策属性,所述决策属性为所述样本数据的绩效。

其中,所述样本数据需同时包括绩效较优的人员(即,绩优人员)和绩效较差的人员(即,绩差人员)的数据,所述样本数据可存储于所述电子装置中。在另一实施方式中,所述电子装置还可以通过接入网络连接之一外部样本库,进而来获取所述样本库存储的样本数据。在其它实施方式中,所述电子装置还可以通过大数据方式来收集并建立所述样本库。

在本实施方式中,所述条件属性可为行为轨迹(如,出差情况)、app 活跃情况、业务扩展情况、消费情况、兴趣爱好、参加培训情况、考勤率等。以所述条件属性包括出差情况、业务扩展情况、参加培训情况以及考勤率为例进行说明,所述样本数据如上表 1 所示。

步骤 S12,根据所述样本数据训练一决策树模型,所述决策树模型包括多个叶子节点,每一叶子节点代表一个分组。

其中,决策树算法属于监督学习分类算法,而决策树模型代表对象属性与对象值之间的一种映射关系。其中,所述决策树模型具有 N 层(N 为自然数, $N > 2$),所述电子装置根据所述条件属性设置所述决策树模型中的每个节点(包括根节点、内部节点和叶子节点)的值。所述决策树模型的第一层为根节点,所述决策树模型的第二层代表以第一级条件属性对所述根节点进行分割得到的多个节点,所述决策树模型的第三层代表以第二级条件属性对上一层节点进行分割得到的多个节点,等等。如图 2 所示,所述决策树模型包括四层,即, $N=4$ 。

在本实施方式中,训练所述决策树模型具体包括:

步骤 S121,以所述样本数据作为训练集计算每一条件属性的信息增益(information gain)。

步骤 S122,选择信息增益最大的条件属性作为所述决策树模型的根节点以分割所述样本数据以得到下一层节点;其中,某一条件属性的信息增益越大,说明选择所述条件属性对分类提供的信息越多,越有利于确定性,越有

利于将所述样本数据进行分类。如图 2 所示, 若计算信息增益最大的条件属性为“出差情况”, 则选择“出差情况”条件属性作为根节点以分割所述样本数据。

步骤 S123, 以每一节点包含的样本数据作为训练集重新计算每一条条件属性的信息增益。

步骤 S124, 选择信息增益最大的条件属性分割所述节点以得到下一层节点。如图 2 所示, 对于包含每月出差情况为 0 次($C=0$)的样本的节点, 若计算信息增益最大的条件属性为“业务拓展情况”, 则根据“业务拓展情况”条件属性对所述节点进行分割而获得下一层节点。对于包含每月出差情况为 ≥ 2 次($C \geq 2$)的样本的节点, 若计算信息增益最大的条件属性为“参加培训情况”, 则根据“参加培训情况”条件属性对所述节点进行分割而获得下一层节点。实际训练过程中, 对位于同一层的多个节点进行分割所采用的条件属性通常不同。每一次分割后形成的节点比上一层节点的数据纯度更高。

步骤 S125, 递归执行步骤 S123 以及 S124, 直至分割停止。此时, 每一叶节点代表一个分组, 每一分组包含的绩优人员数量与绩差人员数量之间具有固定的比值, 而多个分组包含的绩优人员数量与绩差人员数量之间的比值可以互不相同。如, 叶子节点“分组 1”的所述比值可为 1:8。当然, 某一叶节点包含的可均为绩优人员, 也可以均为绩差人员。其中, 每一组样本数据只能被划分至其中一个叶子节点中, 即, 每一组样本数据不能同属于两个及两个以上的叶子节点。

在本实例中, 在所述决策树的层数达到一预设层数(如: 4 层)时停止下一次分割。在另一实施例, 在所述决策树的当前层的每一节点包含的样本数量小于一预设数量时停止下一次分割。

其中, 每一条条件属性的信息增益 $Gain(S, A)$ 可通过如下方式计算:

$$Gain(S, A) = Entropy(S) - Entropy(S, A)$$

其中, $Gain(S, A)$ 表示 A 条件属性在训练集 S 上的信息增益, $Entropy(S)$ 表示训练集 S 的信息熵, $Entropy(S, A)$ 表示 A 属性的信息熵。

例如, 如表 1 所示, 在步骤 S121 中, 所述训练集共有样本数据 16 组, 绩优的有 11 组, 绩差的有 5 组。以“出差情况”条件属性的信息增益计算为例进行说明, 其它条件属性的信息增益计算过程同理。对于 $C=0$ 的情况, 共有样本数据 5 组, 绩优的有 4 组, 绩差的有 1 组; 对于 $C=1$ 的情况, 共有样本数据 4 组, 绩优的有 2 组, 绩差的有 2 组; 对于 $C \geq 2$ 的情况, 共有样本数据 7 组, 绩优的有 5 组, 绩差的有 2 组。因此, “出差情况”条件属性的信息增益的计算方式如下:

$$\begin{aligned} Entropy(S) &= -\frac{11}{16} \log_2 \frac{11}{16} - \frac{5}{16} \log_2 \frac{5}{16} \\ Entropy(A) &= \frac{4}{16} \times \left(-\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} \right) + \frac{4}{16} \times \left(-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} \right) \\ &\quad + \frac{7}{16} \times \left(-\frac{5}{7} \log_2 \frac{5}{7} - \frac{2}{7} \log_2 \frac{2}{7} \right) \end{aligned}$$

$$\begin{aligned} Gain(S,A) = & -\frac{11}{16} \log_2 \frac{11}{16} - \frac{5}{16} \log_2 \frac{5}{16} - \frac{4}{16} \times (-\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5}) \\ & - \frac{4}{16} \times (-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4}) - \frac{7}{16} \times (-\frac{5}{7} \log_2 \frac{5}{7} - \frac{2}{7} \log_2 \frac{2}{7}) \end{aligned}$$

步骤 S13, 对所述条件属性进行分类以确定多个业务类型, 每一业务类型对应至少一条条件属性, 所述条件属性作为所述业务类型的评价因子。

所述业务类型可包括资源型、勤奋型以及开放型等。资源型可以是指业务能力强、工作能力强的人员。勤奋型可以是指学习时间长、每天工作时长的人员, 开放型可以是指性格活跃、乐于社交的人员。其中, 所述评价因子是能够表征所述业务类型的重要特征的评价指标。每一业务类型可对应一个评价因子, 也可对应至少两个评价因子。

例如, 对于“资源型”业务类型来说, 对应的评价因子可为业务拓展情况等。对于“勤奋型”业务类型来说, 对应的评价因子可为参加培训情况、考勤率等。对于“开放型”业务类型来说, 对应的评价因子可为行为轨迹(如, 出差情况)、app 活跃情况等。

其中, 所述业务类型的数量为 $M(M>1, M$ 为自然数), 所述分组的数量为 $N(M>1, M$ 为自然数)。 M 可以等于 N , 也可以不等于 N 。

步骤 S14, 确定划分得到每一分组的过程所涉及的条件属性, 根据每一条件属性对应的评价因子所属的业务类型来计算对应于同一业务类型的条件属性所出现的次数, 并以所述次数作为每一分组所涉及的业务类型的权重因子。

例如, 对于“分组 8”所包含的样本数据来说, 所涉及的条件属性为: “出差情况”、“参加培训情况”以及“考勤率”。“出差情况”条件属性对应的评价因子所属的业务类型为“开放型”, “参加培训情况”以及“考勤率”对应的评价因子所属的业务类型为“勤奋型”。因此, 对于“分组 8”所涉及的条件属性来说, 对应于“开放型”业务类型的条件属性出现 1 次, “开放型”业务类型的权重因子为 1; 而对应于“勤奋型”业务类型的条件属性出现 2 次, “勤奋型”业务类型的权重因子为 2。

步骤 S15, 选择权重因子最大的业务类型, 并将所述分组归纳至所述业务类型中。

例如, 若“分组 8”所包含的样本数据中, “开放型”业务类型的权重因子为 1, 而“勤奋型”业务类型的权重因子为 2, 则将所述“分组 8”归纳至“勤奋型”业务类型中。

在本实施方式中, 所述选择权重因子最大的业务类型, 并将所述分组归纳至所述业务类型中包括:

步骤 S151, 当选择出权重因子最大的业务类型时, 判断所述业务类型的个数。

步骤 S152, 当权重因子最大的业务类型仅为一个时, 直接将所述分组归纳至所述业务类型中; 当权重因子最大的业务类型包括至少两个时, 随机将所述分组归纳至其中一业务类型中。当然, 在其它实施方式中, 当权重因子

最大的业务类型包括至少两个时，由于所述分组同时符合至少两个业务类型的特征，因此，将所述分组同时分配至所述不同的业务类型中。

或者，所述处理器 102 执行所述计算机可读指令 103 时实现上述分组归纳装置实施例中各模块/单元的功能，例如图 3 中的单元 301-305。

本申请实施例基于每一分组涉及的条件属性在同一业务类型中出现的次数对每一分组进行客观归纳，使得归纳标准统一；而且，将分组结果与业务的需求类型对应起来，有利于辅助业务分析。

示例性的，所述计算机可读指令 103 可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器 101 中，并由所述处理器 102 执行，以完成本申请。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机可读指令段，该指令段用于描述所述计算机可读指令 103 在所述电子装置 1 中的执行过程。例如，所述计算机可读指令 103 可以被分割成图 3 中的获取模块 301、训练模块 302、分类模块 303、计算模块 304 以及归纳模块 305。

所述电子装置 1 可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。本领域技术人员可以理解，所述示意图仅仅是电子装置 1 的示例，并不构成对电子装置 1 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，例如所述电子装置 1 还可以包括输入输出设备、网络接入设备、总线等。

所称处理器 102 可以是中央处理单元(Central Processing Unit, CPU)，还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现成可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器 30 也可以是任何常规的处理器等，所述处理器 102 是所述电子装置 1 的控制中心，利用各种接口和线路连接整个电子装置 1 的各个部分。

所述存储器 101 可用于存储所述计算机可读指令 103 和/或模块/单元，所述处理器 102 通过运行或执行存储在所述存储器 101 内的计算机可读指令和/或模块/单元，以及调用存储在存储器 101 内的数据，实现所述电子装置 1 的各种功能。所述存储器 101 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（比如声音播放功能、图像播放功能等）等；存储数据区可存储根据电子装置 1 的使用所创建的数据（比如音频数据、电话本等）等。此外，存储器 101 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，例如硬盘、内存、插接式硬盘，智能存储卡 (Smart Media Card, SMC)，安全数字 (Secure Digital, SD) 卡，闪存卡 (Flash Card)、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。

所述电子装置 1 集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个非易失性可读取存储介质中。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，也可以

通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一非易失性可读存储介质中，该计算机可读指令在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机可读指令包括计算机可读指令代码，所述计算机可读指令代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述非易失性可读介质可以包括：能够携带所述计算机可读指令代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述非易失性可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，非易失性可读介质不包括电载波信号和电信信号。

在本申请所提供的几个实施例中，应该理解到，所揭露的电子装置和方法，可以通过其它的方式实现。例如，以上所描述的电子装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

另外，在本申请各个实施例中的各功能单元可以集成在相同处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在相同单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

对于本领域技术人员而言，显然本申请不限于上述示范性实施例的细节，而且在不背离本申请的精神或基本特征的情况下，能够以其他的具体形式实现本申请。因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本申请的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本申请内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。电子装置权利要求中陈述的多个单元或电子装置也可以由同一个单元或电子装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

最后应说明的是，以上实施例仅用以说明本申请的技术方案而非限制，尽管参照较佳实施例对本申请进行了详细说明，本领域的普通技术人员应当理解，可以对本申请的技术方案进行修改或等同替换，而不脱离本申请技术方案的精神和范围。

权 利 要 求 书

1. 一种分组归纳方法，其特征在于，包括：

获取多组样本数据，每一组样本数据包括多个条件属性以及对应的决策属性；

根据所述样本数据训练一决策树模型，所述决策树模型包括多个叶子节点，每一叶子节点代表一个分组；

对所述条件属性进行分类以确定多个业务类型，每一业务类型对应至少一条件属性，所述条件属性作为所述业务类型的评价因子；

确定划分得到每一分组的过程所涉及的条件属性，根据每一条件属性对应的评价因子所属的业务类型来计算对应于同一业务类型的条件属性所出现的次数，并以所述次数作为每一分组所涉及的业务类型的权重因子；以及

选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中。

2. 如权利要求 1 所述的分组归纳方法，其特征在于，训练所述决策树模型包括：

以所述样本数据作为训练集计算每一条件属性的信息增益；

选择信息增益最大的条件属性作为所述决策树模型的根节点以分割所述样本数据以得到下一层节点；

以每一节点包含的样本数据作为训练集重新计算每一条件属性的信息增益；

选择信息增益最大的条件属性分割所述节点以得到下一层节点；以及

递归执行重新计算每一条件属性的信息增益的步骤以及选择信息增益最大的条件属性分割所述节点的步骤，直至分割停止。

3. 如权利要求 2 所述的分组归纳方法，其特征在于，在所述决策树的层数达到一预设层数时停止下一次分割。

4. 如权利要求 2 所述的分组归纳方法，其特征在于，在所述决策树的当前层的每一节点包含的样本数量小于一预设数量时停止下一次分割。

5. 如权利要求 1 所述的分组归纳方法，其特征在于，所述选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中包括：

当选择出权重因子最大的业务类型时，判断所述业务类型的个数；以及

当权重因子最大的业务类型仅为一个时，直接将所述分组归纳至所述业务类型中，当权重因子最大的业务类型包括至少两个时，随机将所述分组归纳至其中一业务类型中。

6. 如权利要求 1 所述的分组归纳方法，其特征在于，所述选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中包括：

当选择出权重因子最大的业务类型时，判断所述业务类型的个数；以及

当权重因子最大的业务类型仅为一个时，直接将所述分组归纳至所述业务类型中，当权重因子最大的业务类型包括至少两个时，将所述分组同时分配至所述不同的业务类型中。

7. 如权利要求 1 所述的分组归纳方法，其特征在于，所述决策属性为所述样本数据的绩效，所述样本数据同时包括绩优人员和绩差人员的数据。

8. 一种分组归纳装置，其特征在于，包括：

获取模块，用于获取多组样本数据，每一组样本数据包括多个条件属性以及对应的决策属性；

训练模块，用于根据所述样本数据训练一决策树模型，所述决策树模型包括多个叶子节点，每一叶子节点代表一个分组；

分类模块，用于对所述条件属性进行分类以确定多个业务类型，每一业务类型对应至少一条件属性，所述条件属性作为所述业务类型的评价因子；

计算模块，用于确定划分得到每一分组的过程所涉及的条件属性，根据每一条件属性对应的评价因子所属的业务类型来计算对应于同一业务类型的条件属性所出现的次数，并以所述次数作为每一分组所涉及的业务类型的权重因子；以及

归纳模块，用于选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中。

9. 一种电子装置，其特征在于，包括处理器和存储器，所述存储器中存储至少一个计算机可读指令，所述处理器用于执行所述计算机可读指令以实现以下步骤：

获取多组样本数据，每一组样本数据包括多个条件属性以及对应的决策属性；

根据所述样本数据训练一决策树模型，所述决策树模型包括多个叶子节点，每一叶子节点代表一个分组；

对所述条件属性进行分类以确定多个业务类型，每一业务类型对应至少一条件属性，所述条件属性作为所述业务类型的评价因子；

确定划分得到每一分组的过程所涉及的条件属性，根据每一条件属性对应的评价因子所属的业务类型来计算对应于同一业务类型的条件属性所出现的次数，并以所述次数作为每一分组所涉及的业务类型的权重因子；以及

选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中。

10. 如权利要求 9 所述的电子装置，其特征在于，所述处理器在训练所述决策树模型时，执行所述计算机可读指令以实现以下步骤：

以所述样本数据作为训练集计算每一条件属性的信息增益；

选择信息增益最大的条件属性作为所述决策树模型的根节点以分割所述样本数据以得到下一层节点；

以每一节点包含的样本数据作为训练集重新计算每一条件属性的信息增益；

选择信息增益最大的条件属性分割所述节点以得到下一层节点；以及

递归执行重新计算每一条件属性的信息增益的步骤以及选择信息增益最大的条件属性分割所述节点的步骤，直至分割停止。

11. 如权利要求 10 所述的分组归纳方法，其特征在于，所述处理器执行所述计算机可读指令还用以实现以下步骤：

在所述决策树的层数达到一预设层数时停止下一次分割。

12. 如权利要求 10 所述的电子装置，其特征在于，所述处理器执行所述计算机可读指令还用以实现以下步骤：

在所述决策树的当前层的每一节点包含的样本数量小于一预设数量时停止下一次分割。

13. 如权利要求 9 所述的电子装置，其特征在于，所述处理器在所述选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中时，执行所述计算机可读指令以实现以下步骤：

当选择出权重因子最大的业务类型时，判断所述业务类型的个数；以及

当权重因子最大的业务类型仅为一个时，直接将所述分组归纳至所述业务类型中，当权重因子最大的业务类型包括至少两个时，随机将所述分组归纳至其中一业务类型中。

14. 如权利要求 9 所述的电子装置，其特征在于，所述处理器在所述选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中时，执行所述计算机可读指令以实现以下步骤：

当选择出权重因子最大的业务类型时，判断所述业务类型的个数；以及

当权重因子最大的业务类型仅为一个时，直接将所述分组归纳至所述业务类型中，当权重因子最大的业务类型包括至少两个时，将所述分组同时分配至所述不同的业务类型中。

15. 一种非易失性可读存储介质，其特征在于，所述非易失性可读存储介质上存储至少一个计算机可读指令，所述至少一个计算机可读指令被处理器执行以实现以下步骤：

获取多组样本数据，每一组样本数据包括多个条件属性以及对应的决策属性；

根据所述样本数据训练一决策树模型，所述决策树模型包括多个叶子节点，每一叶子节点代表一个分组；

对所述条件属性进行分类以确定多个业务类型，每一业务类型对应至少一条件属性，所述条件属性作为所述业务类型的评价因子；

确定划分得到每一分组的过程所涉及的条件属性，根据每一条件属性对应的评价因子所属的业务类型来计算对应于同一业务类型的条件属性所出现的次数，并以所述次数作为每一分组所涉及的业务类型的权重因子；以及

选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中。

16. 如权利要求 15 所述的存储介质，其特征在于，在训练所述决策树模型时，所述计算机可读指令被所述处理器执行以实现以下步骤：

以所述样本数据作为训练集计算每一条件属性的信息增益；

选择信息增益最大的条件属性作为所述决策树模型的根节点以分割所述样本数据以得到下一层节点；

以每一节点包含的样本数据作为训练集重新计算每一条件属性的信息增益；

选择信息增益最大的条件属性分割所述节点以得到下一层节点；以及

递归执行重新计算每一条条件属性的信息增益的步骤以及选择信息增益最大的条件属性分割所述节点的步骤，直至分割停止。

17. 如权利要求 16 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还用以实现以下步骤：

在所述决策树的层数达到一预设层数时停止下一次分割。

18. 如权利要求 16 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还用以实现以下步骤：

在所述决策树的当前层的每一节点包含的样本数量小于一预设数量时停止下一次分割。

19. 如权利要求 15 所述的存储介质，其特征在于，在所述选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中时，所述计算机可读指令被所述处理器执行时以实现以下步骤：

当选择出权重因子最大的业务类型时，判断所述业务类型的个数；以及

当权重因子最大的业务类型仅为一个时，直接将所述分组归纳至所述业务类型中，当权重因子最大的业务类型包括至少两个时，随机将所述分组归纳至其中一业务类型中。

20. 如权利要求 15 所述的存储介质，其特征在于，在所述选择权重因子最大的业务类型，并将所述分组归纳至所述业务类型中时，所述计算机可读指令被所述处理器执行时以实现以下步骤：

当选择出权重因子最大的业务类型时，判断所述业务类型的个数；以及

当权重因子最大的业务类型仅为一个时，直接将所述分组归纳至所述业务类型中，当权重因子最大的业务类型包括至少两个时，将所述分组同时分配至所述不同的业务类型中。

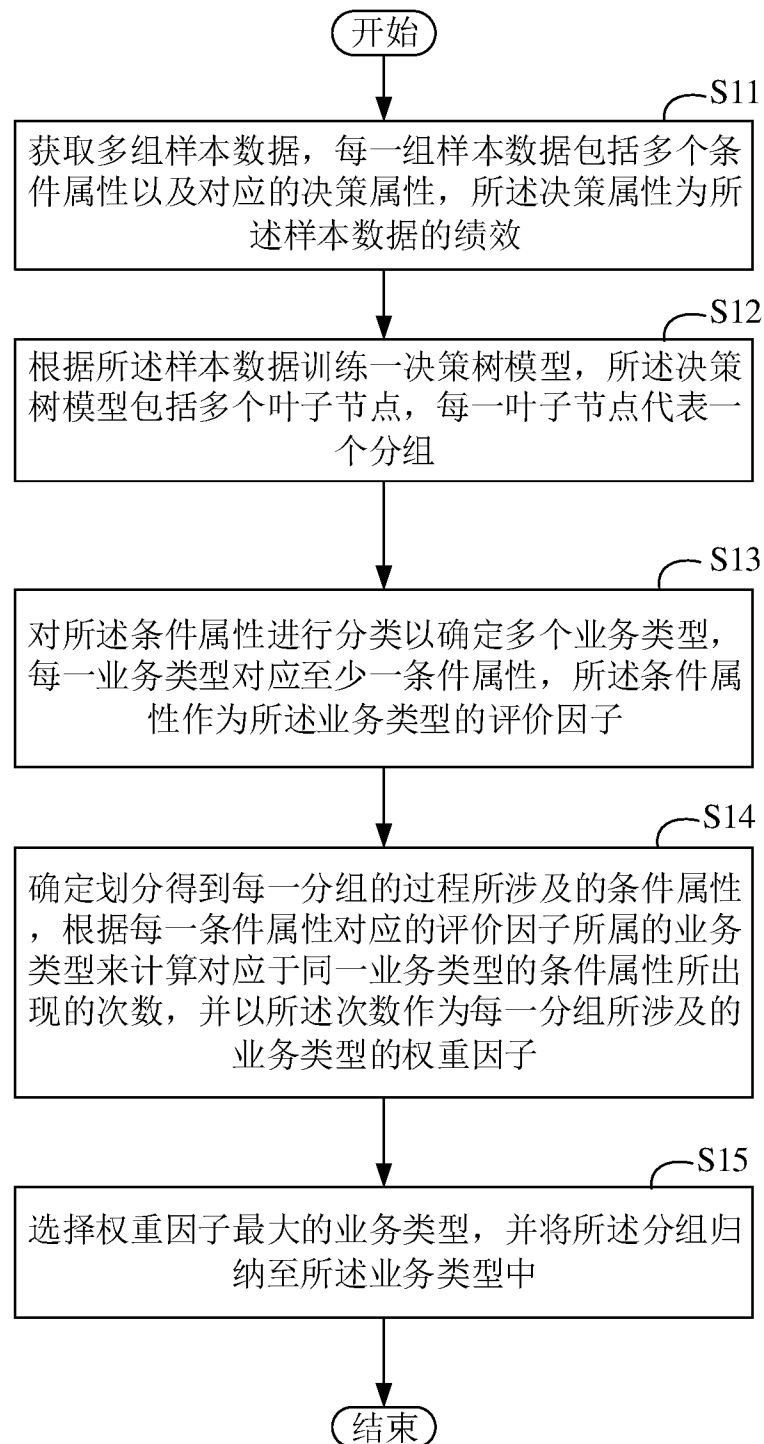


图 1

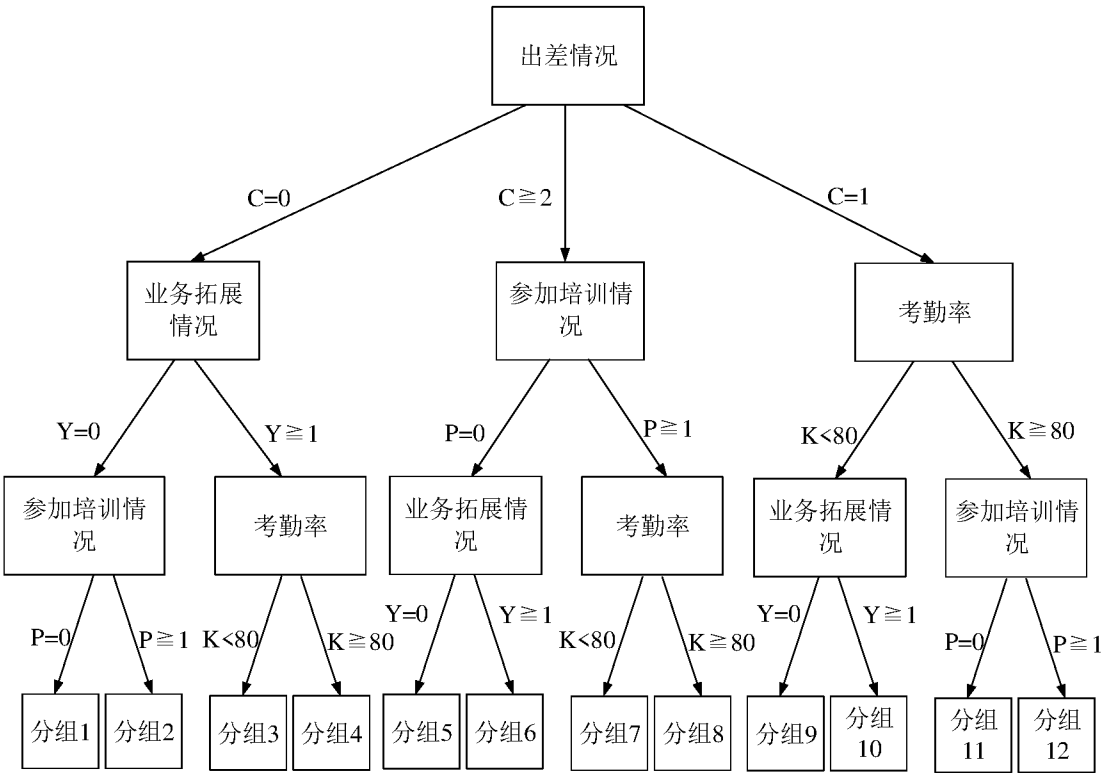


图 2

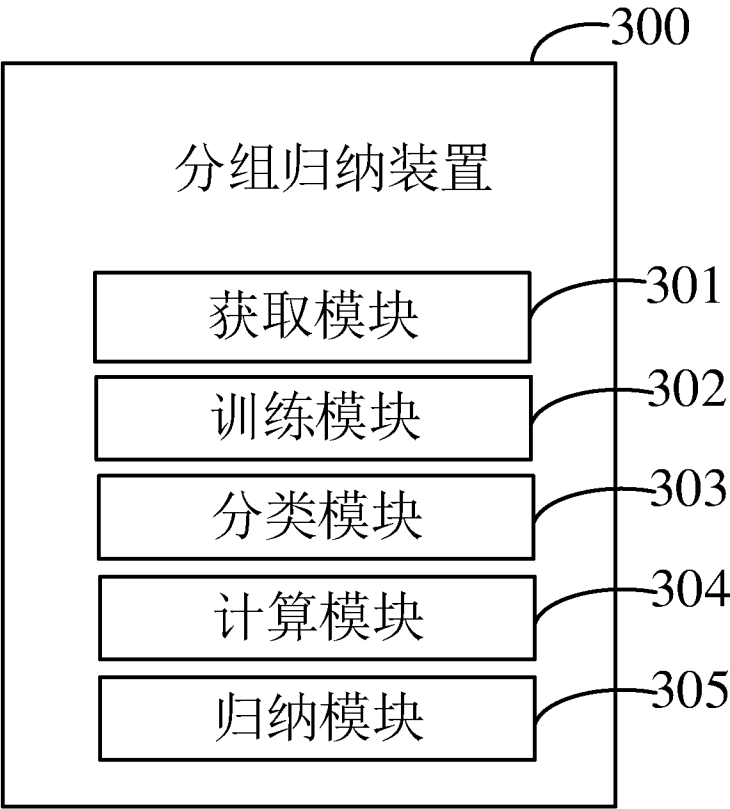


图 3

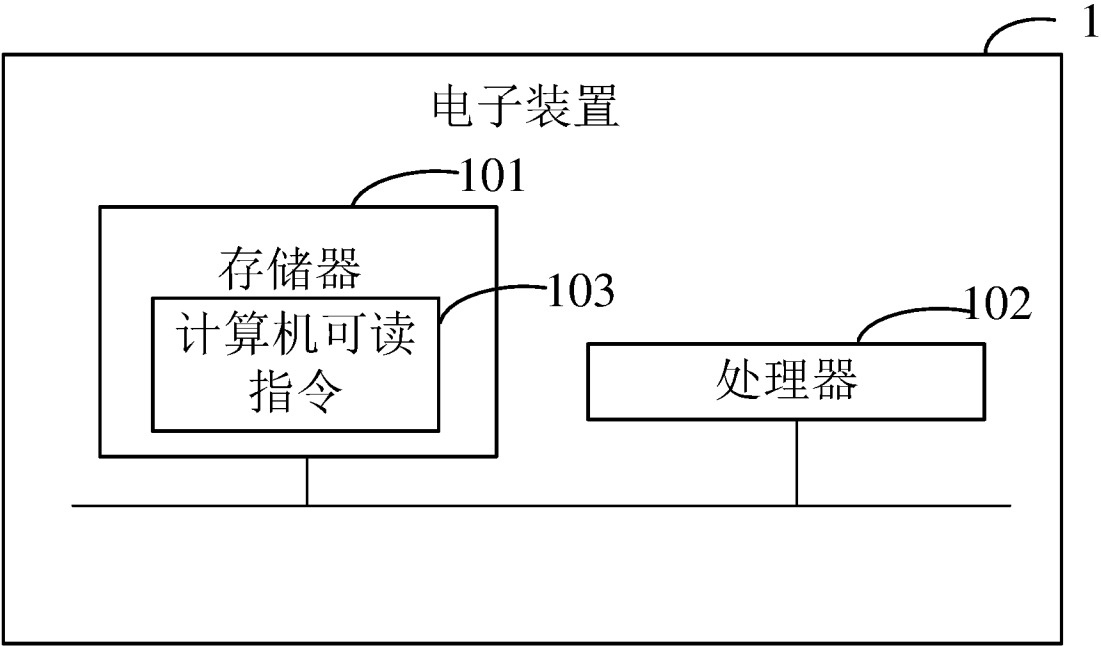


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/077223

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

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)

WPI, EPODOC, CNPAT, CNKI: 分类, 分组, 业务, 决策树, 增益, 条件属性, 模型, 权重, 评价, group, decision tree, model, train+, weight

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109242012 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 18 January 2019 (2019-01-18) claims 1-10, and description, paragraphs [0004]-[0013]	1-20
X	CN 107292186 A (ALIBABA GROUP HOLDING LIMITED) 24 October 2017 (2017-10-24) description, paragraphs [0072]-[0119]	1-20
A	CN 107203774 A (ALIBABA GROUP HOLDING LIMITED) 26 September 2017 (2017-09-26) entire document	1-20
A	CN 108108455 A (GUANGDONG OPPO MOBILE TELECOMMUNICATIONS CORP., LTD.) 01 June 2018 (2018-06-01) entire document	1-20
A	CN 108205570 A (HUAWEI TECHNOLOGIES CO., LTD.) 26 June 2018 (2018-06-26) 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

08 May 2019

Date of mailing of the international search report

31 May 2019

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
Information on patent family members

International application No.

PCT/CN2019/077223

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109242012	A	18 January 2019	None			
CN	107292186	A	24 October 2017	US	2019034834	A1	31 January 2019
				TW	201737058	A	16 October 2017
				WO	2017167097	A1	05 October 2017
CN	107203774	A	26 September 2017	None			
CN	108108455	A	01 June 2018	None			
CN	108205570	A	26 June 2018	None			

国际检索报告

国际申请号

PCT/CN2019/077223

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

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))

WPI, EPDOC, CNPAT, CNKI: 分类, 分组, 业务, 决策树, 增益, 条件属性, 模型, 权重, 评价, group, decision tree, model, train+, weight

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 109242012 A (平安科技深圳有限公司) 2019年 1月 18日 (2019 - 01 - 18) 权利要求1-10, 说明书第[0004]-[0013]段	1-20
X	CN 107292186 A (阿里巴巴集团控股有限公司) 2017年 10月 24日 (2017 - 10 - 24) 说明书第[0072]-[0119]段	1-20
A	CN 107203774 A (阿里巴巴集团控股有限公司) 2017年 9月 26日 (2017 - 09 - 26) 全文	1-20
A	CN 108108455 A (广东欧珀移动通信有限公司) 2018年 6月 1日 (2018 - 06 - 01) 全文	1-20
A	CN 108205570 A (华为技术有限公司) 2018年 6月 26日 (2018 - 06 - 26) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2019年 5月 8日

国际检索报告邮寄日期

2019年 5月 31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

受权官员

丁文勃

电话号码 86-(10)-53961330

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/077223

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	109242012	A	2019年 1月 18日	无	
CN	107292186	A	2017年 10月 24日	US 2019034834 A1	2019年 1月 31日
				TW 201737058 A	2017年 10月 16日
				WO 2017167097 A1	2017年 10月 5日
CN	107203774	A	2017年 9月 26日	无	
CN	108108455	A	2018年 6月 1日	无	
CN	108205570	A	2018年 6月 26日	无	