

(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
2022 年 8 月 11 日 (11.08.2022)



(10) 国际公布号
WO 2022/166830 A1

(51) 国际专利分类号:
G06F 16/35 (2019.01)

(21) 国际申请号: PCT/CN2022/074714

(22) 国际申请日: 2022 年 1 月 28 日 (28.01.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110163603.3 2021年2月5日 (05.02.2021) CN

(71) 申请人: 北京紫光展锐通信技术有限公司(BEIJING UNISOC COMMUNICATIONS TECHNOLOGY CO., LTD.) [CN/CN]; 中国北京市海淀区知春路7号致真大厦B座18层, Beijing 100083 (CN)。

(72) 发明人: 霍小倩 (HUO, Xiaoqian); 中国北京市海淀区知春路7号致真大厦B座18层, Beijing 100083 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司 (SCIHEAD IP LAW FIRM); 中国广东省广州市越秀区先烈中路80号汇华商贸大厦1508室, Guangdong 510070 (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,

(54) **Title:** FEATURE EXTRACTION METHOD AND APPARATUS FOR TEXT CLASSIFICATION

(54) 发明名称: 一种文本分类的特征提取方法及装置

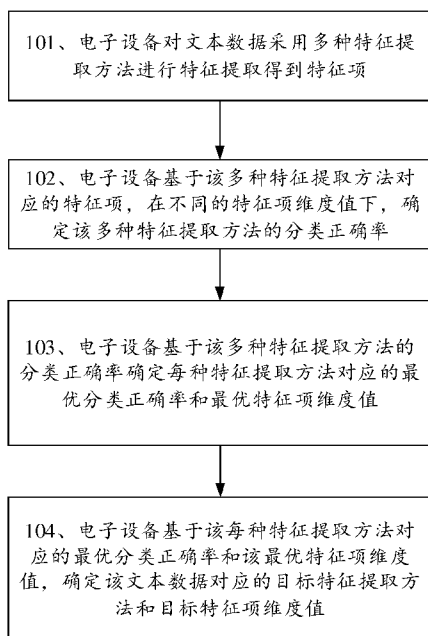


图 1

- 101 An electronic device uses multiple feature extraction methods on text data to extract features so as to obtain feature items
- 102 The electronic device determines the classification accuracy of the multiple feature extraction methods under different feature item dimension values on the basis of the feature items corresponding to the multiple feature extraction methods
- 103 The electronic device determines the optimal classification accuracy and optimal feature dimension value corresponding to each feature extraction method on the basis of the classification accuracy of the multiple feature extraction methods
- 104 The electronic device determines a target feature extraction method and a target feature dimension value corresponding to the text data on the basis of the optimal classification accuracy and optimal feature dimension value corresponding to each feature extraction method

(57) **Abstract:** Provided are a feature extraction method and apparatus for text classification. The method comprises: using multiple feature extraction methods on text data to extract features so as to obtain feature items (101); determining the classification accuracy of the multiple feature extraction methods under different feature item dimension values on the basis of the feature items corresponding to the multiple feature extraction methods, the feature item dimension values being used to represent the number of feature items (102); determining the optimal classification accuracy and optimal feature dimension value corresponding to each feature extraction method on the basis of the classification accuracy of the multiple feature extraction methods (103); determining a target feature extraction method

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))。

and a target feature dimension value corresponding to the text data on the basis of the optimal classification accuracy and optimal feature dimension value corresponding to each feature extraction method (104). The use of the provided method helps to increase the accuracy of text classification.

(57) 摘要: 提供了一种文本分类的特征提取方法及装置, 方法包括: 对文本数据采用多种特征提取方法进行特征提取得到特征项(101); 基于多种特征提取方法对应的特征项, 在不同的特征项维度值下, 确定多种特征提取方法的分类正确率, 特征项维度值用于表示特征项的数量(102); 基于多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值(103); 基于每种特征提取方法对应的最优分类正确率和最优特征项维度值, 确定文本数据对应的目标特征提取方法和目标特征项维度值(104)。采用提供的方法, 有利于提高文本分类的准确性。

一种文本分类的特征提取方法及装置

技术领域

本发明涉及计算机领域，尤其涉及一种文本分类的特征提取方法及装置。

背景技术

自然语言处理（Natural Language Processing, NLP）技术、信息挖掘是目前数据管理的关键技术，而文本分类则是这些技术的操作基础。在文本分类领域，数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限，因此，对文本进行特征提取是直接影响分类效果至关重要的一步。

目前，对于传统的文本分类的特征提取方法，常用的有卡方检验（chi2 Square）、互信息方法（Mutual Information, MI）、信息增益方法（Information Gain, IG）、梯度上升决策树（Gradient Boosting Decision Tree, GBDT）方法、监督式的认为特征提取的方法以及人工方法等。不同的特征提取方法在不同的特征项维度下具有对应的分类效果。但是，传统的特征提取方法存在泛化误差大、无法满足文本分类的高准确率要求的情况。因此，如何提高文本分类的准确性是亟待解决的问题。

发明内容

本申请提供一种文本分类的特征提取方法及装置，有利于提高文本分类的准确性。

第一方面，本申请提供一种文本分类的特征提取方法，该方法包括：对文本数据采用多种特征提取方法进行特征提取得到特征项；基于该多种特征提取方法对应的特征项，在不同的特征项维度值下，确定该多种特征提取方法的分类正确率，该特征项维度值用于表示特征项的数量；基于该多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值；基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值。

基于第一方面描述的方法，通过综合考虑多种特征提取方法的最优分类正确率和最优特征项维度值，针对文本数据，确定出分类效果最优的特征提取方法和该特征提取方法相应的特征项维度值，有利于提高文本分类的准确性。

结合第一方面，在一种可能的实现方式中，该每种特征提取方法对应的最优特征项维度值为该每种特征提取方法的最优分类正确率对应的特征项维度值，一个最优分类正确率对应一个最优特征项维度值。

结合第一方面，在一种可能的实现方式中，该基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值，包括：通过比较该每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率，该数值最大的最优分类正确率包括一个或者多个最优分类正确率；若该数值最大的最优分类正确率包括一个最优分类正确率，则将该数值最大的最优分类正确率对应的特征提取方法作为该目标特征提取方法，并将该数值最大的最优分类正确率对应的最优特征项维度值作为该目标特征项维度值。基于该方式，有利于提高文本分类的准确性。

结合第一方面，在一种可能的实现方式中，该方法还包括：若该数值最大的最优分类正确率包括多个最优分类正确率，则在该多个最优分类正确率对应的多个最优特征项维度值中确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。基于该方式，有利于提高文本分类的准确性。

结合第一方面，在一种可能的实现方式中，该基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值，包括：通过比较该每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率；确定第一最优分类正确率，该第一最优分类正确率与该数值最大的最优分类正确率之间的差值在预设范围内；通过比较该第一最优分类正确率对应的最优特征项维度值与该数值最大的最优分类正确率对应的最优特征项维度值，确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值；其中，该第一最优分类正确率包括一个或者多个最优分类正确率。基于该方式，有利于提高文本分类的准确性。

结合第一方面，在一种可能的实现方式中，该基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值，包括：通过比较该每种特征提取方法对应的最优分类正确率，确定数值大于预设

阈值的最优分类正确率，该数值大于预设阈值的最优分类正确率包括一个或者多个最优分类正确率；通过比较该数值大于预设阈值的最优分类正确率分别对应的最优特征项维度值，确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。

第二方面，本申请提供了一种处理装置，该装置包括处理单元和确定单元，该处理单元和确定单元用于执行上述第一方面或其任一种可能的实现方式中的方法。

第三方面，本申请提供了一种芯片，该芯片包括处理器和通信接口，处理器被配置用于使芯片执行上述第一方面或其任一种可能的实现方式中的方法。

第四方面，本申请提供了一种模组设备，该模组设备包括通信模组、电源模组、存储模组以及芯片，其中：该电源模组用于为该模组设备提供电能；该存储模组用于存储数据和指令；该通信模组用于进行模组设备内部通信，或者用于该模组设备与外部设备进行通信；该芯片用于执行上述第一方面或其任一种可能的实现方式中的方法。

第五方面，本发明实施例公开了一种电子设备，该电子设备包括存储器和处理器，所述存储器用于存储计算机程序，所述计算机程序包括程序指令，所述处理器被配置用于调用所述程序指令，执行上述第一方面或其任一种可能的实现方式中的方法。

第六方面，本申请提供了一种计算机可读存储介质，该计算机存储介质中存储有计算机可读指令，当该计算机可读指令在通信装置上运行时，使得该通信装置执行上述第一方面及其任一种可能的实现方式中的方法。

第七方面，本申请提供一种计算机程序或计算机程序产品，包括代码或指令，当代码或指令在计算机上运行时，使得计算机执行如第一方面或其任一种可能的实现方式中的方法。

附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图1是本申请实施例提供的一种文本分类的特征提取方法的流程图；

图2是本申请实施例提供的一种不同特征提取方法的分类效果随特征项维度值的变化规律的折线图；

图3是本申请实施例提供的一种不同特征提取方法的最优分类效果比较的条形图；

图4是本申请实施例提供的又一种文本分类的特征提取方法的流程图；

图5是本申请实施例提供的又一种文本分类的特征提取方法的流程图；

图6是本申请实施例提供的一种处理装置的结构示意图；

图7是本申请实施例提供的一种电子设备的结构示意图；

图8是本申请实施例提供的一种模组设备的结构示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

本申请以下实施例中所使用的术语只是为了描述特定实施例的目的，而并非旨在作为对本申请的限制。如在本申请的说明书和所附权利要求书中所使用的那样，单数表达形式“一个”、“一种”、“所述”、“上述”、“该”和“这一”旨在也包括复数表达形式，除非其上下文中明确地有相反指示。还应当理解，本申请中使用的术语“和/或”是指并包含一个或多个所列出项目的任何或所有可能组合。

需要说明的是，本申请的说明书和权利要求书中及上述附图中的属于“第一”、“第二”、“第三”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，以便这里描述的本申请的实施例能够以除了在这里图示或描述以外的顺序实施。此外，术语“包括”及其任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或服务不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或服务固有的其它步骤或单元。

本申请提出的方法的执行主体可以是电子设备，该电子设备可以是终端设备，也称为终端；可以是一种具有无线收发功能的设备，其可以部署在陆地上，包括室内或室外、手持或车载；也可以部署在水面上（如轮船等）；还可以部署在空中（例如飞机、气球和卫星上等）。电子设备可以是用户设备（user equipment, UE），其中，UE包括具有无线通信功能的手持式设备、车载设备、可穿戴设备或计算设备。示例性地，UE可以是手机（mobile phone）、平板电脑或带无线收发功能的电脑。电子设备还可以是虚拟现实（virtual reality, VR）电子设备、增强现实（augmented reality, AR）电子设备、工业控制中的无线终端、无人驾驶中的无线终端、远程医疗中的无线终端、智能电网中的无线终端、智慧城市（smart city）中的无线终端、智慧家庭（smart home）中的无线终端等等。本申请实施例中，用于实现电子设备的功能的装置可以是终端；也可以是能够支持电子设备实现该功能的装置，例如芯片系统，该装置可以被安装在电子设备中。本申请实施例中，芯片系统可以由芯片构成，也可以包括芯片和其他分立器件。

需要说明的是，自然语言处理（Natural Language Processing, NLP）技术、信息挖掘是目前数据管理的关键技术，而文本分类则是这些技术的操作基础。在文本分类领域，数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限，因此，对文本进行特征提取是直接影响分类效果至关重要的一步。

目前，对于传统的文本分类的特征提取方法，常用的有卡方检验（chi2 Square）、互信息方法（Mutual Information, MI）、信息增益方法（Information Gain, IG）、梯度上升决策树（Gradient Boosting Decision Tree, GBDT）方法、监督式的认为特征提取的方法以及人工方法等。不同的特征提取方法在不同的特征项维度下具有对应的分类效果。但是，传统的特征提取方法存在泛化误差大、无法满足文本分类的高准确率要求的情况。因此，如何提高文本分类的准确性是亟待解决的问题。

为了能够提高文本分类的准确性，本申请实施例提供了一种文本分类的特征提取方法及装置。为了更好地理解本申请实施例提供的文本分类的特征提取方法，下面对该文本分类的特征提取方法进行详细描述。

请参阅图1，图1是本申请实施例提供的一种文本分类的特征提取方法的流程图，该文本分类的特征提取方法包括步骤101~步骤104。图1所示的方法执行主体可以为电子设备，

或主体可以为电子设备中的芯片。图1所示的方法执行主体以电子设备为例。其中：

101、电子设备对文本数据采用多种特征提取方法进行特征提取得到特征项。

本申请实施例中，常用的特征提取方法有卡方检验、互信息方法、信息增益方法、梯度上升决策树方法、监督式的认为特征提取的方法等。需要说明的是，还可以使用其他特征提取方法对文本数据进行特征提取，本申请实施例不作限制。基于该方式，便于后续对不同特征提取方法得到的特征项进行处理和分析。

示例性的，以下将主要以电子设备选取卡方检验、互信息方法、信息增益方法、梯度上升决策树方法对文本数据进行特征提取，得到每种特征提取方法提取的特征项为例进行介绍。

可选的，该文本数据可以是一个文本数据，也可以是一个文本数据分为多个部分文本数据中的一个部分文本数据。例如，将一个文本数据分为3个部分文本数据，对其中一个部分文本数据采用多种特征提取方法进行特征提取得到特征项。基于该方式，便于后续确定出每个部分文本数据的最优特征提取方法，以提高整个文本数据分类的准确性。

102、电子设备基于该多种特征提取方法对应的特征项，在不同的特征项维度值下，确定该多种特征提取方法的分类正确率。

本申请实施例中，该特征项维度值用于表示特征项的数量，不同的特征提取方法在不同的特征项维度值下具有对应的分类正确率。其中，可以采用分类器对特征项进行评估和优化。基于该方式，能够确定采用的多种特征提取方法在不同的特征项维度值下的分类正确率。

例如，电子设备选取卡方检验(chi2)、互信息方法(MI)、信息增益方法(IG)、梯度上升决策树方法(GBDT)对文本数据进行特征提取，将随机森林(Random Forest, RF)作为检验分类效果的分类器，特征项维度值在之间变化，变化的步长为5，不同的特征提取方法在不同的特征项维度值下具有对应的分类正确率，如图2所示。例如，当值为25时，卡方检验的分类正确率为92%，互信息方法的分类正确率为76%，信息增益方法的分类正确率为86%，梯度上升决策树方法的分类正确率为84%。

可选的，不同的特征项维度值可以设定在任意范围，选取的多种特征提取方法可以为任意特征提取方法，特征项维度值变化的步长可以为任意数值，在此不作限定。

可选的，检验分类效果的分类器可以为任意分类器，在此不作限定。

103、电子设备基于该多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值。

在一种可能的实现方式中，每种特征提取方法对应的最优特征项维度值为该每种特征提取方法的最优分类正确率对应的特征项维度值，一个最优分类正确率对应一个最优特征项维度值。

例如，电子设备选取卡方检验（chi2）、互信息方法（MI）、信息增益方法（IG）、梯度上升决策树方法（GBDT）对文本数据进行特征提取，将随机森林作为检验分类效果的分类器，从图2可以确定出每种特征提取方法对应的最优分类正确率和最优特征项维度值，如图3所示。其中，卡方检验的最优分类正确率为92%，最优特征项维度值为25；互信息方法的最优分类正确率为90%，最优特征项维度值为55；信息增益方法的最优分类正确率为92%，最优特征项维度值为55；梯度上升决策树方法的最优分类正确率为92%，最优特征项维度值为50。另外，图3中也包含了人工提取方法（byman）的最优分类正确率为84%，最优特征项维度值为80。由此可以得知，卡方检验、互信息方法、信息增益方法、梯度上升决策树方法这四种方法的最优分类正确率都比人工提取方法的最优分类正确率高。

104、电子设备基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值。

本申请实施例中，确定该文本数据对应的目标特征提取方法和目标特征项维度值的方式可以是：电子设备通过比较每种特征提取方法对应的最优分类正确率和该最优特征项维度值，将数值最大的最优分类正确率对应的特征提取方法作为该文本数据对应的目标特征提取方法，将数值最大的最优分类正确率对应的最优特征项维度值作为目标特征项维度值。

可选的，确定该文本数据对应的目标特征提取方法和目标特征项维度值的方式也可以是：电子设备通过比较每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定出大于预设阈值的最优分类正确率，通过比较这些大于预设阈值的最优分类正确率分别对应的最优特征项维度值，将数值最小的最优特征项维度值对应的特征提取方法作为该文本数据对应的目标特征提取方法，将数值最小的最优特征项维度值作为最优特征项维度值。

可选的，确定该文本数据对应的目标特征提取方法和目标特征项维度值的方式还可以为其他方式，在此不作限定。

在图1所描述的方法中，电子设备通过综合考虑多种特征提取方法的最优分类正确率和最优特征项维度值，针对文本数据，确定出分类效果最优的特征提取方法和该特征提取方法相应的特征项维度值。因此，基于图1所描述的方法，有利于提高文本分类的准确性。

请参见图4，图4是本申请实施例提供的又一种文本分类的特征提取方法的流程示意图。该文本分类的特征提取方法包括步骤401~步骤407。步骤404~步骤407为上述步骤104的一种具体的实现方式。图4所示的方法执行主体可以为电子设备，或主体可以为电子设备中的芯片。图4所示的方法执行主体以电子设备为例。其中：

401、电子设备对文本数据采用多种特征提取方法进行特征提取得到特征项。

402、电子设备基于该多种特征提取方法对应的特征项，在不同的特征项维度值下，确定该多种特征提取方法的分类正确率。

403、电子设备基于该多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值。

其中，步骤401~步骤403的具体实现方式与上述步骤101~步骤103的具体实现方式相同，在此不赘述。

404、电子设备通过比较每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率。

其中，该数值最大的最优分类正确率包括一个或者多个最优分类正确率。若该数值最大的最优分类正确率包括一个最优分类正确率，则执行步骤405。若该数值最大的最优分类正确率包括多个最优分类正确率（示例性的，可以参照图3所示的实施例），则执行步骤406和步骤407。

例如，电子设备选取卡方检验（chi2）、互信息方法（MI）、信息增益方法（IG）、梯度上升决策树方法（GBDT）对文本数据进行特征提取，如图4所示，电子设备通过比较每种特征提取方法对应的最优分类正确率，确定出的数值最大的最优分类正确率有3个，均为92%。

405、电子设备将该数值最大的最优分类正确率对应的特征提取方法作为该目标特征提取方法，并将该数值最大的最优分类正确率对应的最优特征项维度值作为该目标特征项维度值。

例如，电子设备选取卡方检验（chi2）、互信息方法（MI）、信息增益方法（IG）、梯度上升决策树方法（GBDT）对文本数据进行特征提取，其中卡方检验的最优分类正确率为92%，最优特征项维度值为25；互信息方法的最优分类正确率为90%，最优特征项维度值为55；信息增益方法的最优分类正确率为91%，最优特征项维度值为50。通过比较每种特征提取方法对应的最优分类正确率，电子设备确定数值最大的最优分类正确率有1个，为92%。该数值最大的最优分类正确率对应的特征提取方法为卡方检验，该数值最大的最优分类正确率对应的最优特征项维度值为25，因此电子设备将卡方检验作为目标特征提取方法，将25作为目标特征项维度值。

406、电子设备在该多个最优分类正确率对应的多个最优特征项维度值中确定数值最小的最优特征项维度值。

407、电子设备将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。

本申请实施例中，确定数值最小的最优特征项维度值是因为降低维度更便于计算和可视化，有利于有效信息的提取综合及无用信息的摒弃。基于该方式，有利于提高文本分类的准确性。

例如，电子设备选取卡方检验（chi2）、互信息方法（MI）、信息增益方法（IG）、梯度上升决策树方法（GBDT）对文本数据进行特征提取，其中卡方检验的最优分类正确率为92%，最优特征项维度值为25；互信息方法的最优分类正确率为90%，最优特征项维度值为55；信息增益方法的最优分类正确率为92%，最优特征项维度值为50。通过比较每种特征提取方法对应的最优分类正确率，电子设备确定数值最大的最优分类正确率有2个，均为92%。2个最优分类正确率92%对应分别对应的最优特征项维度值为25和50，因此数值最小的最优特征项维度值为25。该数值最小的最优特征项维度值对应的特征提取方法为卡方检验，因此电子设备将卡方检验作为目标特征提取方法，将25作为目标特征项维度值。

在图4所描述的方法中，电子设备通过考虑多种特征提取方法的最优分类正确率和最

优特征项维度值，针对文本数据，确定出数值最大的最优分类正确率，从而确定出分类效果最优的特征提取方法和该特征提取方法相应的特征项维度值。因此，基于图4所描述的方法，有利于提高文本分类的准确性。

请参见图5，图5是本申请实施例提供的又一种文本分类的特征提取方法的流程示意图。该文本分类的特征提取方法包括步骤501~步骤507。步骤504~步骤507为上述步骤104的一种具体的实现方式。图5所示的方法执行主体可以为电子设备，或主体可以为电子设备中的芯片。图5所示的方法执行主体以电子设备为例。其中：

501、电子设备对文本数据采用多种特征提取方法进行特征提取得到特征项。

502、电子设备基于该多种特征提取方法对应的特征项，在不同的特征项维度值下，确定该多种特征提取方法的分类正确率。

503、电子设备基于该多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值。

其中，步骤501~步骤503的具体实现方式与上述步骤401~步骤403的具体实现方式相同，在此不赘述。

504、电子设备通过比较每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率。

505、电子设备确定第一最优分类正确率。

本申请实施例中，该第一最优分类正确率与该数值最大的最优分类正确率之间的差值在预设范围内。该第一最优分类正确率包括一个或者多个最优分类正确率。

可选的，该预设范围可以为任意范围，在此不作限定。

506、电子设备通过比较该第一最优分类正确率对应的最优特征项维度值与该数值最大的最优分类正确率对应的最优特征项维度值，确定数值最小的最优特征项维度值。

507、电子设备将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。

例如，假设该第一最优分类正确率与该数值最大的最优分类正确率之间的差值的预设范围为0%~5%，电子设备选取卡方检验（chi2）、互信息方法（MI）、信息增益方法（IG）、

梯度上升决策树方法（GBDT）对文本数据进行特征提取，其中卡方检验的最优分类正确率为92%，最优特征项维度值为25；互信息方法的最优分类正确率为86%，最优特征项维度值为55；信息增益方法的最优分类正确率为90%，最优特征项维度值为15。通过比较每种特征提取方法对应的最优分类正确率，电子设备确定数值最大的最优分类正确率为92%，与该数值最大的最优分类正确率之间的差值在预设范围内的最优分类正确率有1个，为90%，因此电子设备确定第一最优分类正确率有1个，为90%。该第一最优分类正确率对应的最优特征项维度值为15，该数值最大的最优分类正确率对应的最优特征项维度值为25，通过比较，确定数值最小的最优特征项维度值为15。该数值最小的最优特征项维度值对应的特征提取方法为信息增益方法，因此电子设备将信息增益方法作为目标特征提取方法，将15作为目标特征项维度值。

又例如，假设该第一最优分类正确率与该数值最大的最优分类正确率之间的差值的预设范围为0%~5%，电子设备选取卡方检验（chi2）、互信息方法（MI）、信息增益方法（IG）、梯度上升决策树方法（GBDT）对文本数据进行特征提取，其中卡方检验的最优分类正确率为92%，最优特征项维度值为25；互信息方法的最优分类正确率为90%，最优特征项维度值为55；信息增益方法的最优分类正确率为90%，最优特征项维度值为15。通过比较每种特征提取方法对应的最优分类正确率，电子设备确定数值最大的最优分类正确率为92%，与该数值最大的最优分类正确率之间的差值在预设范围内的最优分类正确率有2个，均为90%，因此电子设备确定第一最优分类正确率有2个，均为90%。该第一最优分类正确率对应的最优特征项维度值分别为55和15，该数值最大的最优分类正确率对应的最优特征项维度值为25，通过比较，确定数值最小的最优特征项维度值为15。该数值最小的最优特征项维度值对应的特征提取方法为信息增益方法，因此电子设备将信息增益方法作为目标特征提取方法，将15作为目标特征项维度值。

在图5所描述的方法中，电子设备通过综合考虑多种特征提取方法的最优分类正确率和最优特征项维度值，针对文本数据，确定出分类效果最优的特征提取方法和该特征提取方法相应的特征项维度值。因此，基于图5所描述的方法，有利于提高文本分类的准确性。

请参见图6，图6是本发明实施例提供的一种处理装置的结构示意图，该候选同步信号

块的处理装置可以为电子设备或具有电子设备功能的装置（例如芯片）。具体的，如图6所示，该候选同步信号块的处理装置60，可以包括：

处理单元601，用于对文本数据采用多种特征提取方法进行特征提取得到特征项；

确定单元602，用于基于该多种特征提取方法对应的特征项，在不同的特征项维度值下，确定该多种特征提取方法的分类正确率，该特征项维度值用于表示该特征项的数量；

该确定单元602，还用于基于该多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值；

该确定单元602，还用于基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值。

可选的，该每种特征提取方法对应的最优特征项维度值为该每种特征提取方法的最优分类正确率对应的特征项维度值，一个最优分类正确率对应一个最优特征项维度值。

可选的，该确定单元602基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值的具体实现方式为：通过比较该每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率，该数值最大的最优分类正确率包括一个或者多个最优分类正确率；若该数值最大的最优分类正确率包括一个最优分类正确率，则将该数值最大的最优分类正确率对应的特征提取方法作为该目标特征提取方法，并将该数值最大的最优分类正确率对应的最优特征项维度值作为该目标特征项维度值。

可选的，该确定单元602还用于：若该数值最大的最优分类正确率包括多个最优分类正确率，则在该多个最优分类正确率对应的多个最优特征项维度值中确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。

可选的，该确定单元602基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值的具体实现方式为：通过比较该每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率；确定第一最优分类正确率，该第一最优分类正确率与该数值最大的最优分类正确率之间的差值在预设范围内；通过比较该第一最优分类正确率对应的最优特征项维度值与该

数值最大的最优分类正确率对应的最优特征项维度值，确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值；其中，该第一最优分类正确率包括一个或者多个最优分类正确率。

可选的，该确定单元602基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值的具体实现方式为：通过比较该每种特征提取方法对应的最优分类正确率，确定数值大于预设阈值的最优分类正确率，该数值大于预设阈值的最优分类正确率包括一个或者多个最优分类正确率；通过比较该数值大于预设阈值的最优分类正确率分别对应的最优特征项维度值，确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。

本发明实施例和图1、图4和图5所示方法实施例基于同一构思，其带来的技术效果也相同，具体原理请参照图1、图4和图5所示实施例的描述，在此不赘述。

请参阅图7，图7是本发明实施例提供的一种电子设备的结构示意图。该电子设备70可以包括存储器701、处理器702和通信接口703，存储器701、处理器702和通信接口703通过一条或多条通信总线连接。其中，通信接口703受处理器702的控制用于收发信息。

存储器701可以包括只读存储器和随机存取存储器，并向处理器702提供指令和数据。存储器701的一部分还可以包括非易失性随机存取存储器。

处理器702可以是中央处理单元（Central Processing Unit，CPU），该处理器702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，可选的，该处理器702也可以是任何常规的处理器等。其中：

存储器701，用于存储程序指令。

处理器702，用于调用存储器701中存储的程序指令。

处理器702调用存储器701中存储的程序指令，使该电子设备70执行以下操作：

对文本数据采用多种特征提取方法进行特征提取得到特征项；

基于该多种特征提取方法对应的特征项，在不同的特征项维度值下，确定该多种特征提取方法的分类正确率，该特征项维度值用于表示特征项的数量；

基于该多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值；

基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值。

在一种实现方式中，该每种特征提取方法对应的最优特征项维度值为该每种特征提取方法的最优分类正确率对应的特征项维度值，一个最优分类正确率对应一个最优特征项维度值。

在一种实现方式中，处理器702基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值的操作具体为：通过比较该每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率，该数值最大的最优分类正确率包括一个或者多个最优分类正确率；若该数值最大的最优分类正确率包括一个最优分类正确率，则将该数值最大的最优分类正确率对应的特征提取方法作为该目标特征提取方法，并将该数值最大的最优分类正确率对应的最优特征项维度值作为该目标特征项维度值。

在一种实现方式中，处理器702还用于：若该数值最大的最优分类正确率包括多个最优分类正确率，则在该多个最优分类正确率对应的多个最优特征项维度值中确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。基于该方式，有利于提高文本分类的准确性。

在一种实现方式中，处理器702基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值的操作具体为：通过比较该每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率；确定第一最优分类正确率，该第一最优分类正确率与该数值最大的最优分类正确率

率之间的差值在预设范围内；通过比较该第一最优分类正确率对应的最优特征项维度值与该数值最大的最优分类正确率对应的最优特征项维度值，确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值；其中，该第一最优分类正确率包括一个或者多个最优分类正确率。

在一种实现方式中，处理器702基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值的具体实现方式为：通过比较该每种特征提取方法对应的最优分类正确率，确定数值大于预设阈值的最优分类正确率，该数值大于预设阈值的最优分类正确率包括一个或者多个最优分类正确率；通过比较该数值大于预设阈值的最优分类正确率分别对应的最优特征项维度值，确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。

需要说明的是，图7对应的实施例中未提及的内容以及各个步骤的具体实现方式可参见图1、图4和图5所示实施例以及前述内容，这里不再赘述。

本申请实施例还提供了一种芯片，该芯片可以执行前述方法实施例中电子设备的相关步骤。该芯片，包括处理器和通信接口，该处理器被配置用于使该芯片执行如下操作：对文本数据采用多种特征提取方法进行特征提取得到特征项；基于该多种特征提取方法对应的特征项，在不同的特征项维度值下，确定该多种特征提取方法的分类正确率，该特征项维度值用于表示该特征项的数量；基于该多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值；基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值。

可选的，该每种特征提取方法对应的最优特征项维度值为该每种特征提取方法的最优分类正确率对应的特征项维度值，一个最优分类正确率对应一个最优特征项维度值。

可选的，该基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，

确定该文本数据对应的目标特征提取方法和目标特征项维度值时，该处理器被配置用于使所述芯片具体执行如下操作：通过比较该每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率，该数值最大的最优分类正确率包括一个或者多个最优分类正确率；若该数值最大的最优分类正确率包括一个最优分类正确率，则将该数值最大的最优分类正确率对应的特征提取方法作为该目标特征提取方法，并将该数值最大的最优分类正确率对应的最优特征项维度值作为该目标特征项维度值。

可选的，该处理器被配置还用于使所述芯片执行如下操作：若该数值最大的最优分类正确率包括多个最优分类正确率，则在该多个最优分类正确率对应的多个最优特征项维度值中确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。

可选的，该基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值时，该处理器被配置用于使所述芯片具体执行如下操作：通过比较该每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率；确定第一最优分类正确率，该第一最优分类正确率与该数值最大的最优分类正确率之间的差值在预设范围内；通过比较该第一最优分类正确率对应的最优特征项维度值与该数值最大的最优分类正确率对应的最优特征项维度值，确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值；其中，该第一最优分类正确率包括一个或者多个最优分类正确率。

可选的，该基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值时，该处理器被配置用于使该芯片具体执行如下操作：通过比较该每种特征提取方法对应的最优分类正确率，确定数值大于预设阈值的最优分类正确率，该数值大于预设阈值的最优分类正确率包括一个或者多个最优分类正确率；通过比较该数值大于预设阈值的最优分类正确率分别对应的最优特征项维度值，确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目

标特征项维度值。

在一种可能的实现方式中，上述芯片包括至少一个处理器、至少一个第一存储器和至少一个第二存储器；其中，前述至少一个第一存储器和前述至少一个处理器通过线路互联，前述第一存储器中存储有指令；前述至少一个第二存储器和前述至少一个处理器通过线路互联，前述第二存储器中存储前述方法实施例中需要存储的数据。

对于应用于或集成于芯片的各个装置、产品，其包含的各个模块可以都采用电路等硬件的方式实现，或者，至少部分模块可以采用软件程序的方式实现，该软件程序运行于芯片内部集成的处理器，剩余的（如果有）部分模块可以采用电路等硬件方式实现。

如图8所示，图8是本申请实施例提供的一种模组设备的结构示意图。该模组设备80可以执行前述方法实施例中电子设备的相关步骤，该模组设备80包括：通信模组801、电源模组802、存储模组803以及芯片804。

其中，所述电源模组802用于为所述模组设备提供电能；所述存储模组803用于存储数据和指令；所述通信模组801用于进行模组设备内部通信，或者用于所述模组设备与外部设备进行通信；所述芯片804用于执行如下操作：对文本数据采用多种特征提取方法进行特征提取得到特征项；基于该多种特征提取方法对应的特征项，在不同的特征项维度值下，确定该多种特征提取方法的分类正确率，该特征项维度值用于表示特征项的数量；基于该多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值；基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值。

可选的，该每种特征提取方法对应的最优特征项维度值为该每种特征提取方法的最优分类正确率对应的特征项维度值，一个最优分类正确率对应一个最优特征项维度值。

可选的，该芯片804基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值的具体实现方式为：通过比较该每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率，该数值最大的最优分类正确率包括一个或者多个最优分类正确率；若该数值最大的最优分类正确率包括一个最优分类正确率，则将该数值最大的最优分类正确率对应的特征提取方

法作为该目标特征提取方法，并将该数值最大的最优分类正确率对应的最优特征项维度值作为该目标特征项维度值。

可选的，该芯片804，还用于若该数值最大的最优分类正确率包括多个最优分类正确率，则在该多个最优分类正确率对应的多个最优特征项维度值中确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。基于该方式，有利于提高文本分类的准确性。

可选的，该芯片804基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值的具体实现方式为：通过比较该每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率；确定第一最优分类正确率，该第一最优分类正确率与该数值最大的最优分类正确率之间的差值在预设范围内；通过比较该第一最优分类正确率对应的最优特征项维度值与该数值最大的最优分类正确率对应的最优特征项维度值，确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值；其中，该第一最优分类正确率包括一个或者多个最优分类正确率。

可选的，该芯片804基于该每种特征提取方法对应的最优分类正确率和该最优特征项维度值，确定该文本数据对应的目标特征提取方法和目标特征项维度值的具体实现方式为：通过比较该每种特征提取方法对应的最优分类正确率，确定数值大于预设阈值的最优分类正确率，该数值大于预设阈值的最优分类正确率包括一个或者多个最优分类正确率；通过比较该数值大于预设阈值的最优分类正确率分别对应的最优特征项维度值，确定数值最小的最优特征项维度值；将该数值最小的最优特征项维度值对应的特征提取方法作为该目标特征提取方法，并将该数值最小的最优特征项维度值作为该目标特征项维度值。

对于应用于或集成于模组设备的各个装置、产品，其包含的各个模块可以都采用电路等硬件的方式实现，不同的模块可以位于模组设备的同一组件（例如芯片、电路模块等）或者不同组件中，或者，至少部分模块可以采用软件程序的方式实现，该软件程序运行于模组设备内部集成的处理器，剩余的（如果有）部分模块可以采用电路等硬件方式实现。

本申请实施例还提供一种计算机可读存储介质，该计算机可读存储介质中存储有指令，当其在处理器上运行时，上述方法实施例的方法流程得以实现。

本申请实施例还提供一种计算机程序产品，当所述计算机程序产品在处理器上运行时，上述方法实施例的方法流程得以实现。

关于上述实施例中描述的各个装置、产品包含的各个模块/单元，其可以是软件模块/单元，也可以是硬件模块/单元，或者也可以部分是软件模块/单元，部分是硬件模块/单元。例如，对于应用于或集成于芯片的各个装置、产品其包含的各个模块/单元可以都采用电路等硬件的方式实现，或者，至少部分模块/单元可以采用软件程序的方式实现，该软件程序运行于芯片内部集成处理器，剩余的（如果有）部分模块/单元可以采用电路等硬件方式实现；对于应用于或集成于芯片模组的各个装置、产品，其包含的各个模块/单元可以都采用电路等硬件的方式实现，不同模块/单元可以位于芯片模组的同一件（例如芯片、电路模块等）或者不同组件中，或者，至少部分模块/单元可以采用软件程序的方式实现，该软件程序运行于芯片模组内部集成的处理器，剩余的（如果有）部分模块/单元可以采用电路等硬件方式实现；对于应用于或集成于终端的各个装置、产品，其包含的模块/单元可以都采用电路等硬件的方式实现，不同的模块/单元可以位于终端内同一组件（例如，芯片、电路模块等）或者不同组件中，或者，至少部分模块/单元可以采用软件程序的方式实现，该软件程序运行于终端内部集成的处理器，剩余的（如果有）部分模块/单元可以采用电路等硬件方式实现。

需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些操作可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本申请所必须的。

本申请提供的各实施例的描述可以相互参照，对各个实施例的描述都各有侧重，某个实施例中未详述的部分，可以参见其他实施例的相关描述。为描述的方便和简洁，例如关于本申请实施例提供的各装置、设备的功能以及执行的操作可以参照本申请方法实施例

的相关描述，各方法实施例之间、各装置实施例之间也可以互相参考、结合或引用。

最后应说明的是：以上各实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述各实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分或者全部技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的范围。

权 利 要 求

1.一种文本分类的特征提取方法，其特征在于，所述方法包括：

对文本数据采用多种特征提取方法进行特征提取得到特征项；

基于所述多种特征提取方法对应的特征项，在不同的特征项维度值下，确定所述多种特征提取方法的分类正确率，所述特征项维度值用于表示所述特征项的数量；

基于所述多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值；

基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值。

2.根据权利要求1所述的方法，其特征在于，所述每种特征提取方法对应的最优特征项维度值为所述每种特征提取方法的最优分类正确率对应的特征项维度值，一个最优分类正确率对应一个最优特征项维度值。

3.根据权利要求2所述的方法，其特征在于，所述基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值，包括：

通过比较所述每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率，所述数值最大的最优分类正确率包括一个或者多个最优分类正确率；

若所述数值最大的最优分类正确率包括一个最优分类正确率，则将所述数值最大的最优分类正确率对应的特征提取方法作为所述目标特征提取方法，并将所述数值最大的最优分类正确率对应的最优特征项维度值作为所述目标特征项维度值。

4.根据权利要求3所述的方法，其特征在于，所述方法还包括：

若所述数值最大的最优分类正确率包括多个最优分类正确率，则在所述多个最优分类正确率对应的多个最优特征项维度值中确定数值最小的最优特征项维度值；

将所述数值最小的最优特征项维度值对应的特征提取方法作为所述目标特征提取方法，并将所述数值最小的最优特征项维度值作为所述目标特征项维度值。

5.根据权利要求2所述的方法，其特征在于，所述基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目

标特征项维度值，包括：

通过比较所述每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率；

确定第一最优分类正确率，所述第一最优分类正确率与所述数值最大的最优分类正确率之间的差值在预设范围内；

通过比较所述第一最优分类正确率对应的最优特征项维度值与所述数值最大的最优分类正确率对应的最优特征项维度值，确定数值最小的最优特征项维度值；

将所述数值最小的最优特征项维度值对应的特征提取方法作为所述目标特征提取方法，并将所述数值最小的最优特征项维度值作为所述目标特征项维度值；

其中，所述第一最优分类正确率包括一个或者多个最优分类正确率。

6.根据权利要求2所述的方法，其特征在于，所述基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值，包括：

通过比较所述每种特征提取方法对应的最优分类正确率，确定数值大于预设阈值的最优分类正确率，所述数值大于预设阈值的最优分类正确率包括一个或者多个最优分类正确率；

通过比较所述数值大于预设阈值的最优分类正确率分别对应的最优特征项维度值，确定数值最小的最优特征项维度值；

将所述数值最小的最优特征项维度值对应的特征提取方法作为所述目标特征提取方法，并将所述数值最小的最优特征项维度值作为所述目标特征项维度值。

7.一种处理装置，其特征在于，所述装置包括：

处理单元，用于对文本数据采用多种特征提取方法进行特征提取得到特征项；

确定单元，用于基于所述多种特征提取方法对应的特征项，在不同的特征项维度值下，确定所述多种特征提取方法的分类正确率，所述特征项维度值用于表示所述特征项的数量；

所述确定单元，还用于基于所述多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值；

所述确定单元，还用于基于所述每种特征提取方法对应的最优分类正确率和所述最优

特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值。

8.根据权利要求7所述的装置，其特征在于，所述每种特征提取方法对应的最优特征项维度值为所述每种特征提取方法的最优分类正确率对应的特征项维度值，一个最优分类正确率对应一个最优特征项维度值。

9.根据权利要求8所述的装置，其特征在于，所述确定单元基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值的具体实现方式为：

通过比较所述每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率，所述数值最大的最优分类正确率包括一个或者多个最优分类正确率；

若所述数值最大的最优分类正确率包括一个最优分类正确率，则将所述数值最大的最优分类正确率对应的特征提取方法作为所述目标特征提取方法，并将所述数值最大的最优分类正确率对应的最优特征项维度值作为所述目标特征项维度值。

10.根据权利要求9所述的装置，其特征在于，所述确定单元还用于：

若所述数值最大的最优分类正确率包括多个最优分类正确率，则在所述多个最优分类正确率对应的多个最优特征项维度值中确定数值最小的最优特征项维度值；

将所述数值最小的最优特征项维度值对应的特征提取方法作为所述目标特征提取方法，并将所述数值最小的最优特征项维度值作为所述目标特征项维度值。

11.根据权利要求8所述的装置，其特征在于，所述确定单元基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值的具体实现方式为：

通过比较所述每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率；

确定第一最优分类正确率，所述第一最优分类正确率与所述数值最大的最优分类正确率之间的差值在预设范围内；

通过比较所述第一最优分类正确率对应的最优特征项维度值与所述数值最大的最优分类正确率对应的最优特征项维度值，确定数值最小的最优特征项维度值；

将所述数值最小的最优特征项维度值对应的特征提取方法作为所述目标特征提取方

法，并将所述数值最小的最优特征项维度值作为所述目标特征项维度值；

其中，所述第一最优分类正确率包括一个或者多个最优分类正确率。

12.根据权利要求8所述的装置，其特征在于，所述确定单元基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值的具体实现方式为：

通过比较所述每种特征提取方法对应的最优分类正确率，确定数值大于预设阈值的最优分类正确率，所述数值大于预设阈值的最优分类正确率包括一个或者多个最优分类正确率；

通过比较所述数值大于预设阈值的最优分类正确率分别对应的最优特征项维度值，确定数值最小的最优特征项维度值；

将所述数值最小的最优特征项维度值对应的特征提取方法作为所述目标特征提取方法，并将所述数值最小的最优特征项维度值作为所述目标特征项维度值。

13.一种芯片，其特征在于，包括处理器和通信接口，所述处理器被配置用于使所述芯片执行如下操作：

对文本数据采用多种特征提取方法进行特征提取得到特征项；

基于所述多种特征提取方法对应的特征项，在不同的特征项维度值下，确定所述多种特征提取方法的分类正确率，所述特征项维度值用于表示所述特征项的数量；

基于所述多种特征提取方法的分类正确率确定每种特征提取方法对应的最优分类正确率和最优特征项维度值；

基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值。

14.根据权利要求13所述的芯片，其特征在于，所述每种特征提取方法对应的最优特征项维度值为所述每种特征提取方法的最优分类正确率对应的特征项维度值，一个最优分类正确率对应一个最优特征项维度值。

15.根据权利要求14所述的芯片，其特征在于，所述基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值时，所述处理器被配置用于使所述芯片具体执行如下操作：

通过比较所述每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率，所述数值最大的最优分类正确率包括一个或者多个最优分类正确率；

若所述数值最大的最优分类正确率包括一个最优分类正确率，则将所述数值最大的最优分类正确率对应的特征提取方法作为所述目标特征提取方法，并将所述数值最大的最优分类正确率对应的最优特征项维度值作为所述目标特征项维度值。

16.根据权利要求15所述的芯片，其特征在于，所述处理器被配置还用于使所述芯片执行如下操作：

若所述数值最大的最优分类正确率包括多个最优分类正确率，则在所述多个最优分类正确率对应的多个最优特征项维度值中确定数值最小的最优特征项维度值；

将所述数值最小的最优特征项维度值对应的特征提取方法作为所述目标特征提取方法，并将所述数值最小的最优特征项维度值作为所述目标特征项维度值。

17.根据权利要求14所述的芯片，其特征在于，所述基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值时，所述处理器被配置用于使所述芯片具体执行如下操作：

通过比较所述每种特征提取方法对应的最优分类正确率，确定数值最大的最优分类正确率；

确定第一最优分类正确率，所述第一最优分类正确率与所述数值最大的最优分类正确率之间的差值在预设范围内；

通过比较所述第一最优分类正确率对应的最优特征项维度值与所述数值最大的最优分类正确率对应的最优特征项维度值，确定数值最小的最优特征项维度值；

将所述数值最小的最优特征项维度值对应的特征提取方法作为所述目标特征提取方法，并将所述数值最小的最优特征项维度值作为所述目标特征项维度值；

其中，所述第一最优分类正确率包括一个或者多个最优分类正确率。

18.根据权利要求14所述的芯片，其特征在于，所述基于所述每种特征提取方法对应的最优分类正确率和所述最优特征项维度值，确定所述文本数据对应的目标特征提取方法和目标特征项维度值时，所述处理器被配置用于使所述芯片具体执行如下操作：

通过比较所述每种特征提取方法对应的最优分类正确率，确定数值大于预设阈值的最

优分类正确率，所述数值大于预设阈值的最优分类正确率包括一个或者多个最优分类正确率；

通过比较所述数值大于预设阈值的最优分类正确率分别对应的最优特征项维度值，确定数值最小的最优特征项维度值；

将所述数值最小的最优特征项维度值对应的特征提取方法作为所述目标特征提取方法，并将所述数值最小的最优特征项维度值作为所述目标特征项维度值。

19.一种模组设备，其特征在于，所述模组设备包括通信模组、电源模组、存储模组以及芯片，其中：

所述电源模组用于为所述模组设备提供电能；

所述存储模组用于存储数据和指令；

所述通信模组用于进行模组设备内部通信，或者用于所述模组设备与外部设备进行通信；

所述芯片用于执行如权利要求1~6中任一项所述的方法。

20.一种电子设备，其特征在于，包括存储器和处理器，所述存储器用于存储计算机程序，所述计算机程序包括程序指令，所述处理器被配置用于调用所述程序指令，执行如权利要求1~6任一项所述的方法。

21.一种计算机可读存储介质，其特征在于，所述计算机存储介质中存储有计算机可读指令，当所述计算机可读指令在通信装置上运行时，使得所述通信装置执行权利要求1~6中任一项所述的方法。

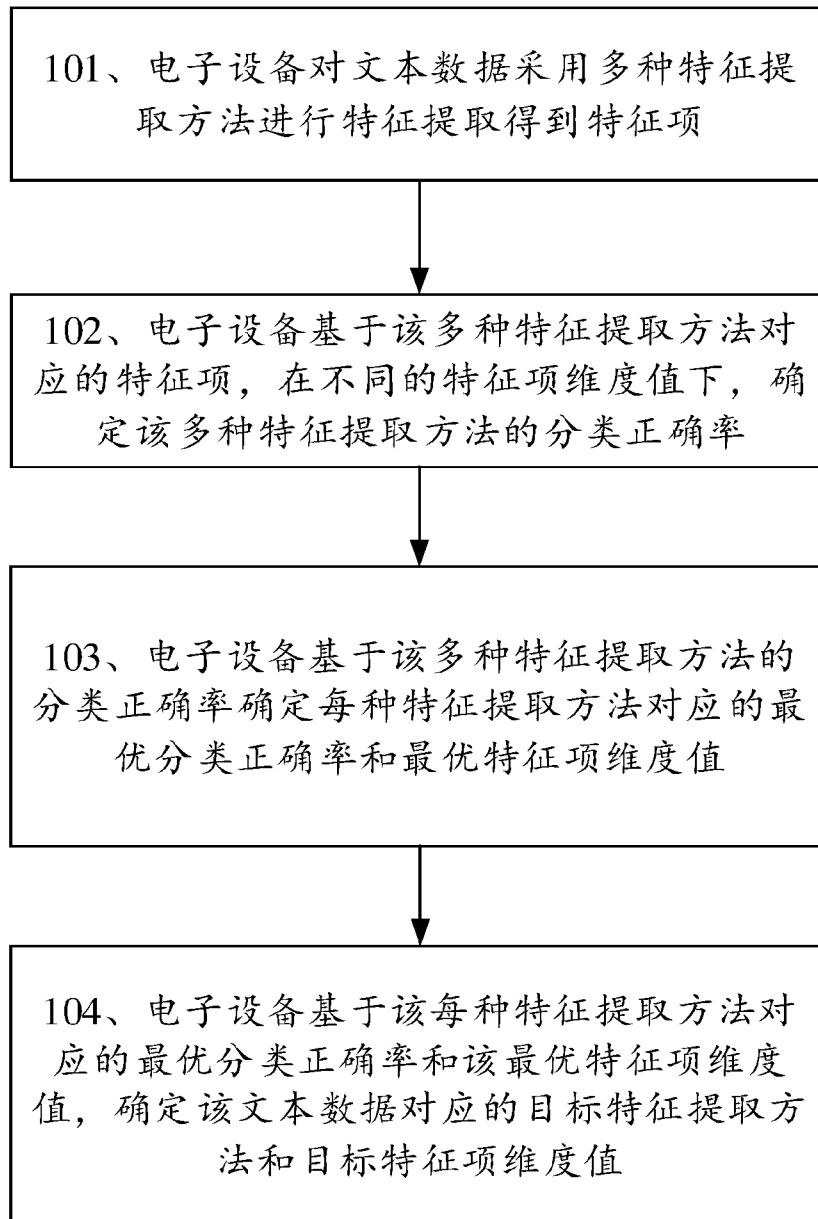


图 1

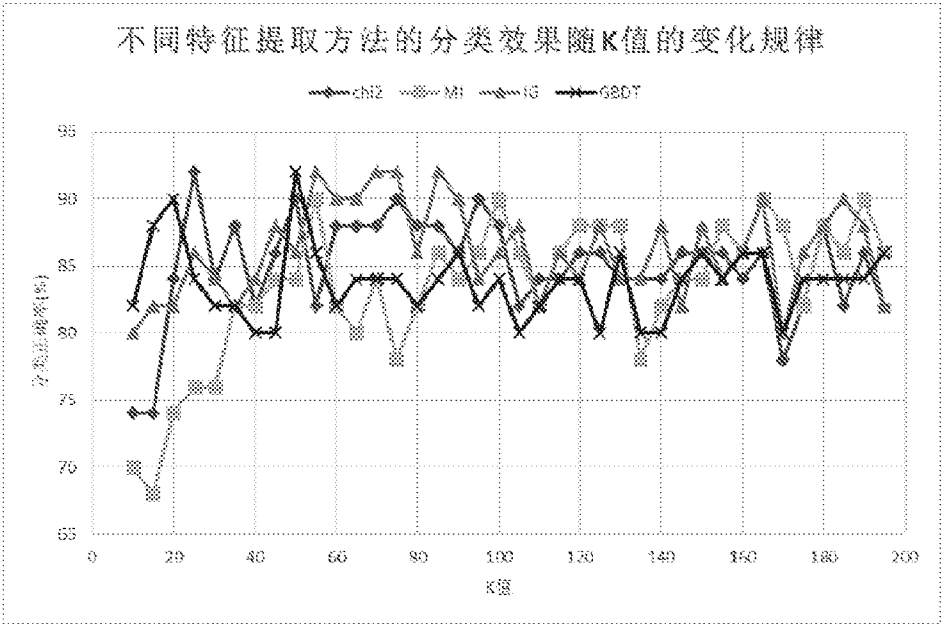


图 2

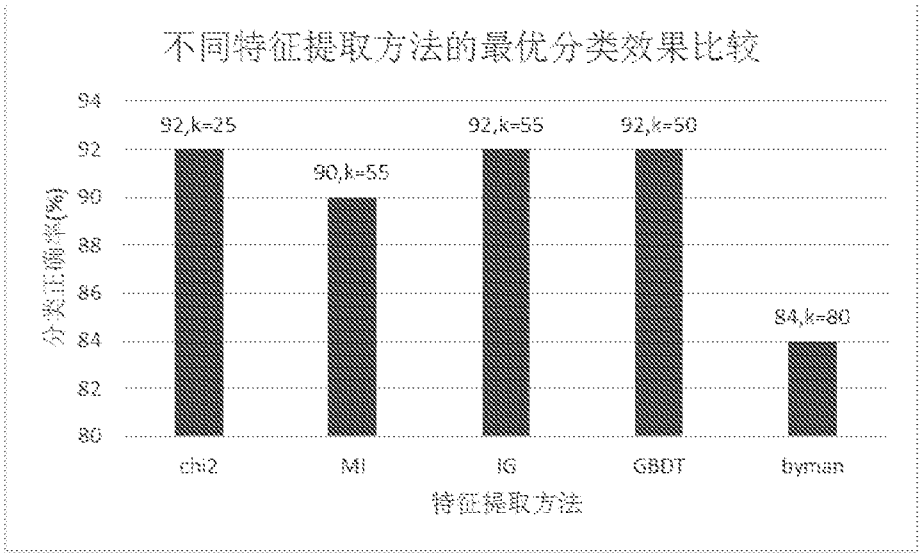


图 3

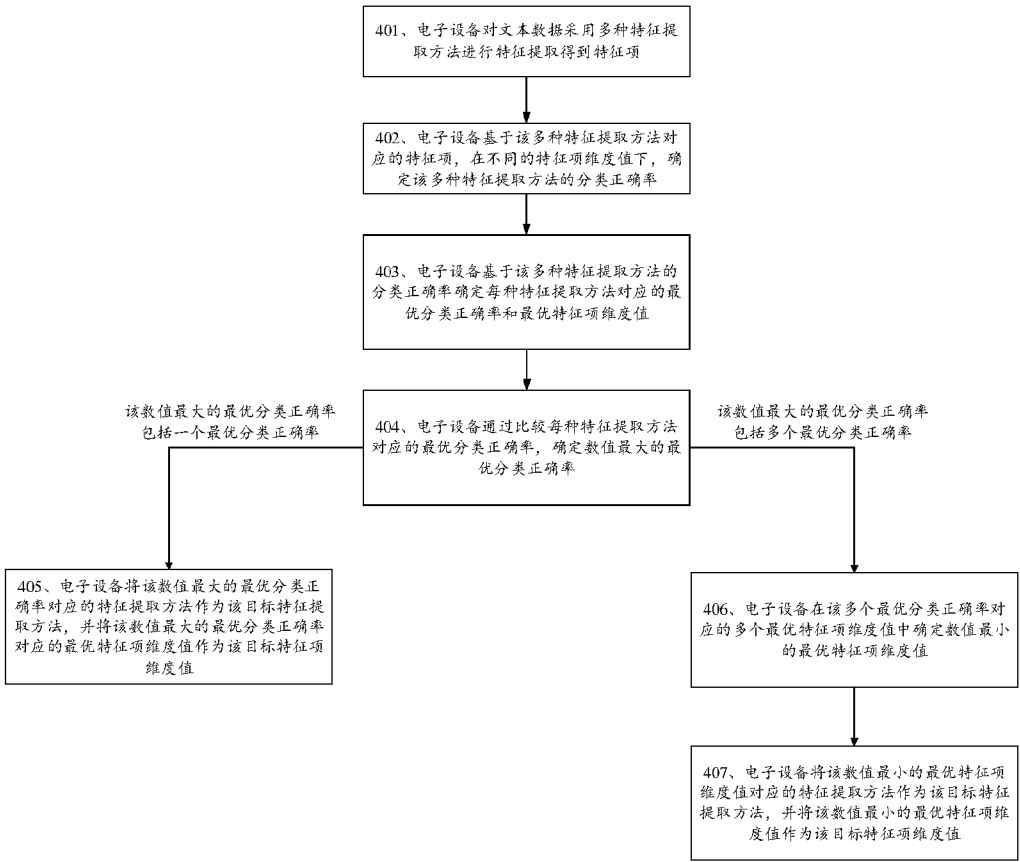


图 4

— 4/5 —

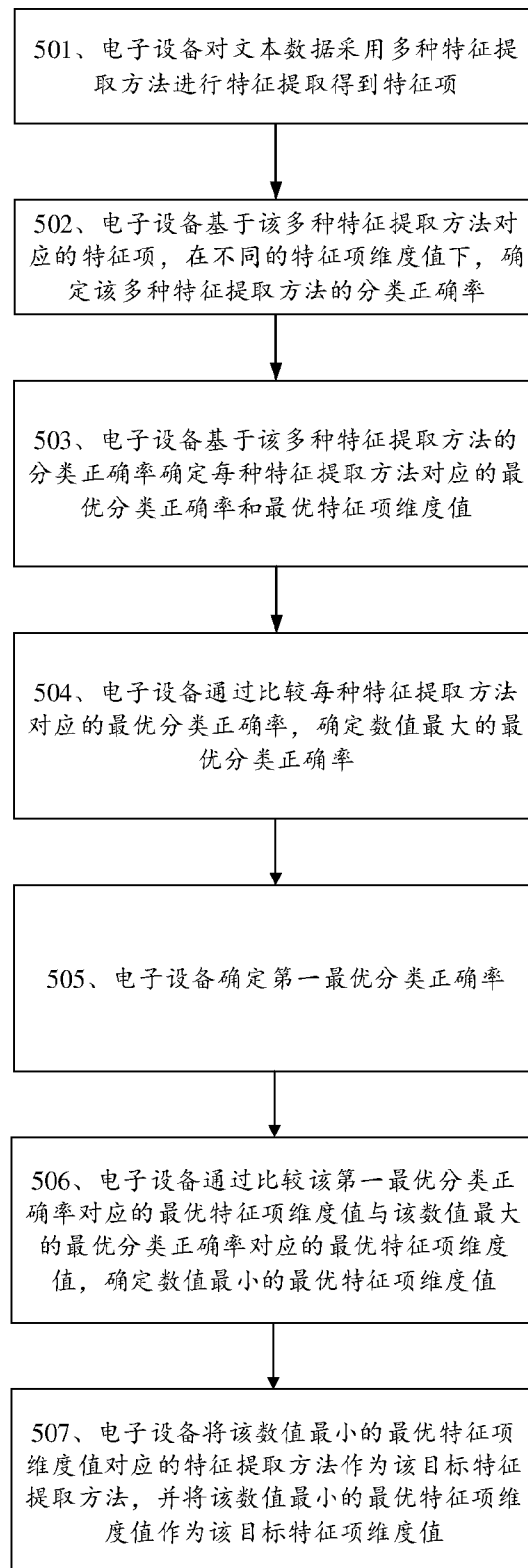


图 5

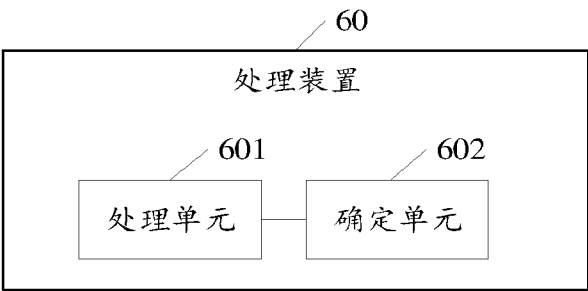


图 6

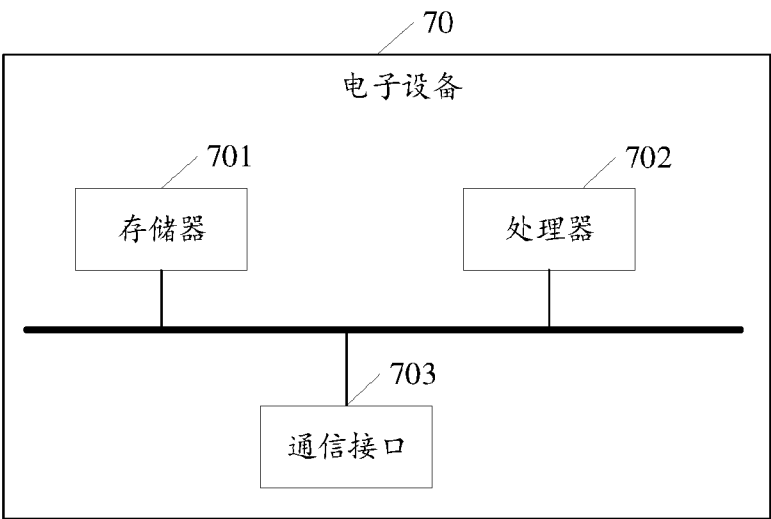


图 7

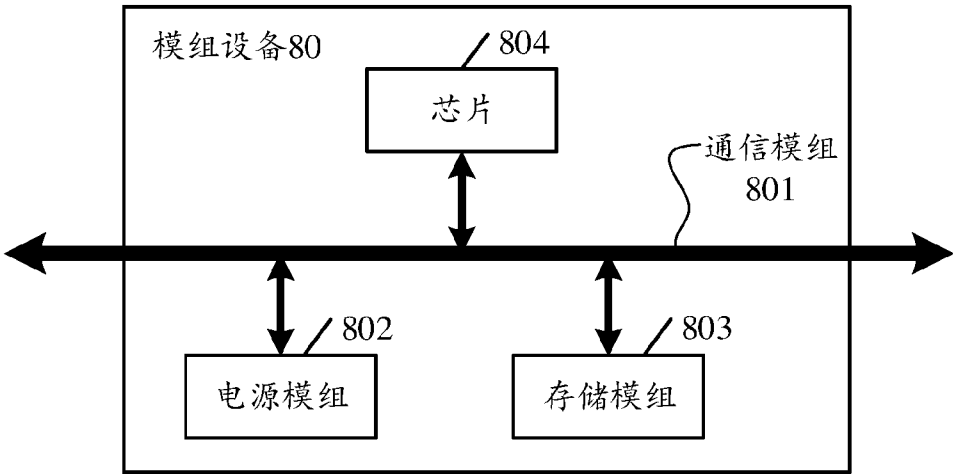


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/074714

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/35(2019.01)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)

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)

CNABS, CNTXT, VEN, SIPOABS, DWPI, ENTXT, CNKI: 特征, 提取, 获取, 多种, 几种, 若干, 目标, 最优, 最好, 最佳, 维度, 数量, 对应, 分类, 准确率, 正确率, 误差, feature, extract, obtain, multiple, various, several, target, optimal, best, dimension, number, correspond, classify, correct, accuracy, rate, error

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112989036 A (BEIJING UNISOC COMMUNICATION TECHNOLOGY CO., LTD.) 18 June 2021 (2021-06-18) description, paragraphs 4-122, and figures 1-8	1-21
A	CN 110597878 A (GUANGDONG UNIVERSITY OF TECHNOLOGY) 20 December 2019 (2019-12-20) description, paragraphs 5-161, and figures 1-7	1-21
A	CN 111666748 A (WUHAN UNIVERSITY et al.) 15 September 2020 (2020-09-15) entire document	1-21
A	CN 107766792 A (BEIJING INSTITUTE OF TECHNOLOGY) 06 March 2018 (2018-03-06) entire document	1-21
A	US 2020394557 A1 (BOULT TERRANCE E) 17 December 2020 (2020-12-17) entire document	1-21
A	US 2019377823 A1 (ELEMENT AI INC) 12 December 2019 (2019-12-12) entire document	1-21



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance
 “E” earlier application or patent but published on or after the international filing date
 “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
 “O” document referring to an oral disclosure, use, exhibition or other means
 “P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
 “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
 “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
 “&” document member of the same patent family

Date of the actual completion of the international search

22 March 2022

Date of mailing of the international search report

13 April 2022

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/CN2022/074714

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	112989036	A	18 June 2021	None			
CN	110597878	A	20 December 2019	None			
CN	111666748	A	15 September 2020	None			
CN	107766792	A	06 March 2018	CN	107766792	B	26 February 2021
US	2020394557	A1	17 December 2020	None			
US	2019377823	A1	12 December 2019	WO	2019232645	A1	12 December 2019

国际检索报告

国际申请号

PCT/CN2022/074714

A. 主题的分类 G06F 16/35 (2019.01) 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) 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)) CNABS, CNTXT, VEN, SIPOABS, DWPI, ENTXT, CNKI: 特征, 提取, 获取, 多种, 几种, 若干, 目标, 最优, 最好, 最佳, 维度, 数量, 对应, 分类, 准确率, 正确率, 误差, feature, extract, obtain, multiple, various, several, target, optimal, best, dimension, number, correspond, classify, correct, accuracy, rate, error																							
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 112989036 A (北京紫光展锐通信技术有限公司) 2021年6月18日 (2021 - 06 - 18) 说明书第4-122段, 图1-8</td> <td>1-21</td> </tr> <tr> <td>A</td> <td>CN 110597878 A (广东工业大学) 2019年12月20日 (2019 - 12 - 20) 说明书第5-161段, 图1-7</td> <td>1-21</td> </tr> <tr> <td>A</td> <td>CN 111666748 A (武汉大学等) 2020年9月15日 (2020 - 09 - 15) 全文</td> <td>1-21</td> </tr> <tr> <td>A</td> <td>CN 107766792 A (北京理工大学) 2018年3月6日 (2018 - 03 - 06) 全文</td> <td>1-21</td> </tr> <tr> <td>A</td> <td>US 2020394557 A1 (BOULT TERRANCE E) 2020年12月17日 (2020 - 12 - 17) 全文</td> <td>1-21</td> </tr> <tr> <td>A</td> <td>US 2019377823 A1 (ELEMENT AI INC) 2019年12月12日 (2019 - 12 - 12) 全文</td> <td>1-21</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 112989036 A (北京紫光展锐通信技术有限公司) 2021年6月18日 (2021 - 06 - 18) 说明书第4-122段, 图1-8	1-21	A	CN 110597878 A (广东工业大学) 2019年12月20日 (2019 - 12 - 20) 说明书第5-161段, 图1-7	1-21	A	CN 111666748 A (武汉大学等) 2020年9月15日 (2020 - 09 - 15) 全文	1-21	A	CN 107766792 A (北京理工大学) 2018年3月6日 (2018 - 03 - 06) 全文	1-21	A	US 2020394557 A1 (BOULT TERRANCE E) 2020年12月17日 (2020 - 12 - 17) 全文	1-21	A	US 2019377823 A1 (ELEMENT AI INC) 2019年12月12日 (2019 - 12 - 12) 全文	1-21
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
PX	CN 112989036 A (北京紫光展锐通信技术有限公司) 2021年6月18日 (2021 - 06 - 18) 说明书第4-122段, 图1-8	1-21																					
A	CN 110597878 A (广东工业大学) 2019年12月20日 (2019 - 12 - 20) 说明书第5-161段, 图1-7	1-21																					
A	CN 111666748 A (武汉大学等) 2020年9月15日 (2020 - 09 - 15) 全文	1-21																					
A	CN 107766792 A (北京理工大学) 2018年3月6日 (2018 - 03 - 06) 全文	1-21																					
A	US 2020394557 A1 (BOULT TERRANCE E) 2020年12月17日 (2020 - 12 - 17) 全文	1-21																					
A	US 2019377823 A1 (ELEMENT AI INC) 2019年12月12日 (2019 - 12 - 12) 全文	1-21																					
☐ 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> * 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> “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 2022年3月22日		国际检索报告邮寄日期 2022年4月13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 潘芳芳 电话号码 (86-27) 59371929																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/074714

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	112989036	A	2021年6月18日	无	
CN	110597878	A	2019年12月20日	无	
CN	111666748	A	2020年9月15日	无	
CN	107766792	A	2018年3月6日	CN	107766792 B 2021年2月26日
US	2020394557	A1	2020年12月17日	无	
US	2019377823	A1	2019年12月12日	WO	2019232645 A1 2019年12月12日