(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
2017 年 10 月 26 日 (26.10.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/181665 A1

(51) 国际专利分类号:
G06F 19/24 (2011.01)

(21) 国际申请号: PCT/CN2016/106255

(22) 国际申请日: 2016 年 11 月 17 日 (17.11.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610246971.3 2016年4月20日 (20.04.2016) CN

(71) 申请人: 苏州大学(SOOCHOW UNIVERSITY) [CN/CN]; 中国江苏省苏州市工业园区仁爱路199号, Jiangsu 215123 (CN)。

(72) 发明人: 张莉(ZHANG, Li); 中国江苏省苏州市张家港市长泾路10号, Jiangsu 215600 (CN)。 黄晓娟(HUANG, Xiaojuan); 中国江苏省苏州市张家港市长泾路10号, Jiangsu 215600 (CN)。 王邦军(WANG, Bangjun); 中国江苏省苏州市张家港市长泾路10号, Jiangsu 215600 (CN)。 张召(ZHANG, Zhao); 中国江苏省苏州市张家港市长泾路10号, Jiangsu 215600 (CN)。 李凡长(LI, Fanzhang); 中国江苏省苏州市张家港市长泾路10号, Jiangsu 215600 (CN)。

(74) 代理人: 北京集佳知识产权代理有限公司(UNITALEN ATTORNEYS AT LAW); 中国北京市朝阳区建国门外大街22号赛特广场7层, Beijing 100004 (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,

(54) **Title:** GENE EXPRESSION DATA CLASSIFICATION METHOD AND CLASSIFICATION SYSTEM

(54) **发明名称**: 一种基因表达数据分类方法及分类系统

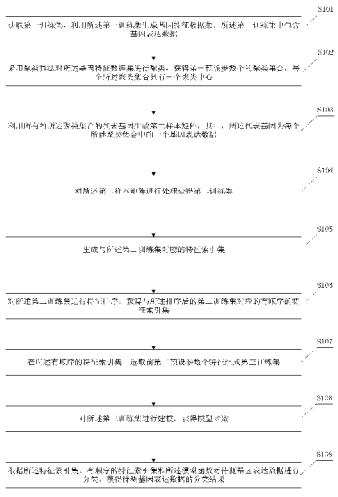


图 1

S101 Acquire a first training set and use the first training set to generate a gene feature data set, the first training set comprising gene expression data
S102 Use a clustering algorithm to perform clustering of the gene feature data set and acquire a number of first preset parameter cluster sets, each cluster sets having one cluster centre
S103 Use a representative gene of all of the cluster sets to generate a second sample matrix, the representative gene being gene expression data present in each cluster set
S104 Perform processing of the second sample matrix to acquire a second training set
S105 Generate a feature index set corresponding to the second training set
S106 Perform feature sorting of the second training set to acquire an ordered feature index set corresponding to the sorted second training set
S107 Select from the ordered feature index set a third training set composed of a number of second preset parameter features
S108 Perform modelling of the third training set to acquire a model function
S109 On the basis of the feature index set, the ordered feature index set, and the model function, perform classification of the gene expression data to be measured, and acquire classification results of the gene expression data to be measured

(57) Abstract: Disclosed in the present application are a gene expression data classification method and classification system, the gene expression data classification method comprising: after acquiring a gene feature data set, using a clustering algorithm to perform clustering of the gene feature data set, and acquiring a number of first preset parameter cluster sets; performing processing of the cluster sets to acquire a second sample matrix, a second training set, and a feature index set in order to perform dimensionality reduction on the gene expression data, and thereby reduce the redundancies in the gene expression data, greatly reducing the calculation resources and the calculation time taken up during the subsequent process of performing feature selection of the second training set; using a clustering algorithm to perform a clustering operation on the gene feature data set takes up few calculation resources and little calculation time, and therefore using the present gene expression data classification method to implement classification of gene expression data to be measured also takes up few calculation resources and little calculation time.

(57) 摘要: 本申请公开了一种基因表达数据分类方法及分类系统，其中，所述基因表达数据分类方法中在获得所述基因特征数据集之后，采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，然后对所述聚类集合进行处理获得第二样本矩阵、第二训练集以及特征索引集以对基因表达数据进行降维，从而降低基因表达数据之间的冗余度，进而在很大程度上降低了在之后对所述第二训练集进行特征选择的过程中占用的计算资源以及耗费的计算时间；而采用聚类算法对所述基因特征数据集进行聚类操作所占用的计算资源以及耗费的计算时间都很少，因此采用所述基因表达数据分类方法对待测基因表达数据进行分类占用的计算资源和耗费的计算时间都较少。

WO 2017/181665 A1

[见续页]

GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, 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))。

一种基因表达数据分类方法及分类系统

本申请要求于2016年4月20日提交中国专利局、申请号为201610246971.3、发明名称为“一种基因表达数据分类方法及分类系统”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

本申请涉及基因分类技术领域，更具体地说，涉及一种基因表达数据分类方法及分类系统。

背景技术

利用DNA微阵列计数可以同时测定成千上万维基因的表达数据，这些表达数据能够帮助研究人员研究生物的本质。但是在大量的基因表达数据中，只有很少量的基因表达数据是研究人员的研究客体，以癌症基因的研究为例，癌症基因的表达数据样本通常少于一百，而在大量的基因表达数据中对癌症基因与其他基因进行分类就需要耗费大量的计算资源和计算时间。

有研究人员利用支持向量机递归特征消除（Support Vector Machine Recursive Feature Elimination, SVM-RFE）算法可以自动消除大量的基因表达数据中的冗余基因（即对基因分类没用贡献的基因），以实现从大量的基因表达数据中找出目标基因的目的。

但是SVM-RFE算法仍然需要对大量的基因表达数据进行特征选择处理，需要占用大量的计算资源并且耗费大量的计算时间。

发明内容

为解决上述技术问题，本发明提供了一种基因表达数据分类方法及分类系统，以解决对基因表达数据进行分类需要占用大量的计算资源并且耗费大量的计算时间的问题。

为解决上述技术问题，本发明实施例提供了如下技术方案:

一种基因表达数据分类方法，包括:

获取第一训练集，利用所述第一训练集生成基因特征数据集，所述第一训练集中包含基因表达数据;

采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心;

利用所有的所述聚类集合的代表基因生成第二样本矩阵，其中，所述 代表基因为每个所述聚类集合中的一个基因表达数据;

对所述第二样本矩阵进行处理获得第二训练集;

生成与所述第二训练集对应的特征索引集;

对所述第二训练集进行特征排序，获得与所述排序后的第二训练集对应的有顺序的特征索引集;

在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集;

对所述第三训练集进行建模，获得模型函数;

根据所述特征索引集、有顺序的特征索引集和所述模型函数对待测基因表达数据进行分类，获得待测基因表达数据的分类结果。

优选的，所述第一预设参数的设定方式为:

采用 N 折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数，其中 N 为 5、10 或 20。

优选的，所述采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心，包括:

采用 K-means 聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心。

优选的，所述代表基因通过公式 $\overline{\mathbf{g}}_k = \min_{\mathbf{g}_i \in G_k} \|\mathbf{g}_i - \mathbf{m}_k\|_2^2, k = 1, \cdots, K$ 生成;

其中，G_k表示第k个聚类集合，$\overline{\mathbf{g}}_k$ 表示所述第k个聚类集合的代表基因，m_k表示第k个聚类中心，K表示所述第一预设参数；g_i表示所述聚类集合中的基因表达数据;

其中，所述第二样本矩阵为 $\mathbf{X}' = [\overline{\mathbf{g}}_1, \cdots, \overline{\mathbf{g}}_K]^T \in R^{K \times N}$ ，其中， R 表示实数集，

N 表示所述第一训练集中的样本总个数。

优选的，根据所述特征索引集、有顺序的特征索引集和所述模型函数对待测基因表达数据进行分类，获得待测基因表达数据的诊断结果包括:

根据所述特征索引集对所述待测基因的表达数据进行特征选择，获得第一次特征选择后的样本;

根据所述有顺序的特征索引集，选择所述第一次特征选择后的样本中的前第二预设参数个特征组成第二次特征选择后的样本;

将所述第二次特征选择后的样本输入所述模型函数中，获得所述模型函数的输出结果，根据所述输出结果获得所述待测基因表达数据的分类结果。

一种基因表达数据分类系统，包括:

特征选择模块，用于获取第一训练集，利用所述第一训练集生成基因特征数据集，所述第一训练集包含基因表达数据; 采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心; 利用所有的所述聚类集合的代表基因生成第二样本矩阵，其中，所述代表基因为每个所述聚类集合中的一个基因表达数据; 对所述第二样本矩阵进行处理获得第二训练集，生成与所述第二训练集对应的特征索引集; 对所述第二训练集进行特征排序，获得与所述排序后的第二训练集对应的有顺序的特征索引集; 在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集;

训练模块，用于对所述第三训练集进行建模，获得模型函数;

诊断模块，用于根据所述特征索引集、有顺序的特征索引集和所述模型函数对待测基因表达数据进行分类，获得待测基因表达数据的分类结果。

优选的，所述特征选择模块包括:

预处理单元，用于获取基因样本的第一训练集，对所述第一训练集进行预处理，生成第一样本矩阵，并根据所述样本矩阵生成基因特征数据集;

第一特征选择单元，用于采用 N 折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数，其中 N 为 5 或 10 或 20; 采用 K-means 聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心，在每个聚类集合

中选取一个基因作为该聚类集合的代表基因，并利用所有的聚类集合的代表基因生成第二样本矩阵，并对所述第二样本矩阵进行处理获得第二训练集，同时生成与所述第二训练集对应的特征索引集；

第二特征选择单元，用于对所述第二训练集进行特征排序，获得有顺序的特征索引集，确定保留的特征个数为第二预设参数，在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集。

优选的，所述第一特征选择单元用于采用 N 折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数，其中 N 为 5 或 10 或 20；采用 K-means 聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个聚类中心和第一预设参数个聚类集合，在每个聚类集合中选取一个基因作为该聚类集合的代表基因，并利用所有的聚类集合的代表基因生成第二样本矩阵，并取所述第二样本矩阵的每一列 $\mathbf{x}'_i$ 组成所述第二训练集 $\{\mathbf{x}'_i, y_i\}_{i=1}^{N}$，其中，$\mathbf{x}'_i \in R^K$。

优选的，所述代表基因通过公式 $\overline{\mathbf{g}}_k = \min_{\mathbf{g}_i \in G_k} \|\mathbf{g}_i - \mathbf{m}_k\|_2^2, k = 1, \cdots, K$ 生成，其中，G_k表示第k个聚类集合，$\overline{\mathbf{g}}_k$表示所述第k个聚类集合的代表基因，m_k表示第k个聚类中心，K表示所述第一预设参数；g_i表示所述聚类集合中的基因表达数据；

其中，所述第二样本矩阵为 $\mathbf{X}' = [\overline{\mathbf{g}}_1, \cdots, \overline{\mathbf{g}}_K]^T \in R^{K \times N}$，其中 R 表示实数集，N 表示生成所述第一训练集中的样本总个数。

优选的，所述诊断模块包括：

第一选择单元，用于根据所述特征索引集对所述待测基因表达数据进行特征选择，获得第一次特征选择后的样本；

第二选择单元，用于根据所述有顺序的特征索引集选择所述第一次特征选择后的样本中的前第二预设参数个特征组成第二次特征选择后的样本；

诊断单元，用于将所述第二次特征选择后的样本输入所述模型函数中，获得所述模型函数的输出结果，根据所述输出结果获得所述待测基因表达数据的分类结果。

从上述技术方案可以看出，本发明实施例提供了一种基因表达数据分类方法及分类系统，其中，所述基因表达数据分类方法中在获得所述基因特征数据集之后，采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个

的聚类集合，然后对所述聚类集合进行处理获得第二样本矩阵、第二训练集以及特征索引集以对基因表达数据进行降维，从而降低基因表达数据之间的冗余度，进而在很大程度上降低了在之后对所述第二训练集进行特征选择的过程中占用的计算资源以及耗费的计算时间；而采用聚类算法对所述基因特征数据集进行聚类操作所占用的计算资源以及耗费的计算时间都很少，因此采用所述基因表达数据分类方法对待测基因表达数据进行分类占用的计算资源和耗费的计算时间都较少。

附图说明

为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术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，还可以根据提供的附图获得其他的附图。

图1为本申请的一个实施例提供的一种基因表达数据分类方法的流程示意图；

图2为本申请的另一个实施例提供的一种基因表达数据分类方法的流程示意图；

图3为本申请的一个实施例提供的一种基因表达数据分类系统的结构示意图；

图4为本申请的一个实施例提供的一种特征选择模块的结构示意图；

图5为本申请的一个实施例提供的一种诊断模块的结构示意图。

具体实施方式

下面将结合本发明实施例中的附图，对本发明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，都属于本发明保护的范围。

本申请的一个实施例提供了一种基因表达数据分类方法，如图1所示，包括:

S101: 获取第一训练集，利用所述第一训练集生成基因特征数据集，所述第一训练集中包含基因表达数据。

需要说明的是，所述第一训练集中的基因表达数据通过DNA微阵列技术获取。

但在本申请的其他实施例中，还可以通过其他技术或者设备获取基因表达数据，本申请对获取基因表达数据所采用的方法或装置并不做限定，具体视实际情况而定。

在本申请的一个具体实施例中，通过DNA微阵列技术获取基因表达数据构成第一训练集$\{\mathbf{x}_i, y_i\}_{i=1}^{N}$，其中，$\mathbf{x}_i$是所述第一训练集中的基因表达数据，$\mathbf{x}_i \in R^D$，$\mathbf{x}_i$为所述第一训练集中的基因表达数据，$y_i$是$\mathbf{x}_i$的标签，表明$\mathbf{x}_i$的类别，$y_i \in \{-1,+1\}$，$N$代表所述第一训练集中样本的总个数，$D$代表所述第一训练集中样本的维数，R代表实数集。对所述第一训练集进行预处理，生成所述第一样本矩阵$\mathbf{X}=[\mathbf{x}_1,\cdots,\mathbf{x}_N]$，该矩阵的每一列为所述第一训练集的一个样本，取所述第一样本矩阵的每一行，即$\mathbf{g}_j=(X_{j\cdot})^T$，$j=1,\cdots,D$，形成所述基因特征数据集$\{\mathbf{g}_j\}_{j=1}^{D}$，$\mathbf{g}_j \in R^N$。

S102: 采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心。

其中，每个聚类集合中包含所述基因特征数据集中相似的基因表达数据，每个聚类集合都具有一个聚类中心，每个聚类集合的聚类中心由该聚类集合中的所有基因表达数据计算得来。在本申请的一个实施例中，每个聚类集合的聚类中心为该聚类集合中所有基因表达数据的平均值。但本申请对此并不做限定，每个聚类集合的聚类中心还可以通过其他方式确定，具体视实际情况而定。

需要说明的是，采用聚类算法对所述基因特征数据集进行聚类的目的是对基因表达数据进行降维，以降低基因表达数据之间的冗余度。

S103: 利用所有的所述聚类集合的代表基因生成第二样本矩阵，其中，所述代表基因为每个所述聚类集合中的一个基因表达数据。

S104: 对所述第二样本矩阵进行处理获得第二训练集。

其中在获得所有的所述聚类集合的代表基因后，利用所有的所述聚类集合的代表基因组成所述第二样本矩阵，取所述第二样本矩阵的每一列组成所述第二训练集。

S105: 生成与所述第二训练集对应的特征索引集。

S106: 对所述第二训练集进行特征排序，获得与所述排序后的第二训练集对应的有顺序的特征索引集。

需要说明的是，在本申请的一个实施例中，采用 SVM-RFE 算法对所述第二训练集进行特征排序，获得与所述排序后的第二训练集对应的有顺序的特征索引集。

S107: 在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集。

其中，所述第二预设参数的值小于所述第一预设参数的值。

S108: 对所述第三训练集进行建模，获得模型函数。

在本申请的一个实施例中，采用支持向量机分类器对所述第三训练集进行建模，获得模型函数。

S109: 根据所述特征索引集、有顺序的特征索引集和所述模型函数对待测基因表达数据进行分类，获得待测基因表达数据的分类结果。

需要说明的是，所述待测基因与所述第一训练集在同一次的基因表达数据的采集过程中获得。

在上述实施例的基础上，在本申请的一个优选实施例中，如图 2 所示，所述基因表达数据分类方法包括:

S201: 获取基因表达数据构成的第一训练集，对所述第一训练集进行预处理，生成所述第一样本矩阵，取所述第一样本矩阵的每一行形成所述基因特征数据集。

S202: 采用 N 折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数，其中 N 为 5 或 10 或 20; 利用 K-means 聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心。

在本实施例中，所述第一预设参数的设定方式为:

采用 N 折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数，其中 N 为 5、10 或 20。在本申请的一个优选实施例中，N 优选为 10。

S203: 利用所有的所述聚类集合的代表基因生成第二样本矩阵，所述代表基因通过公式 $\overline{\mathbf{g}}_k = \min_{\mathbf{g}_i \in G_k} \|\mathbf{g}_i - \mathbf{m}_k\|_2^2, k = 1, \cdots, K$ 获得，所述第二样本矩阵为 $\mathbf{X}' = [\overline{\mathbf{g}}_1, \cdots, \overline{\mathbf{g}}_K]^T \in R^{K \times N}$ 。

其中R表示实数集，N表示所述第一训练集中的样本总个数，G_k表示第k个聚类集合，$\overline{\mathbf{g}}_k$表示所述第k个聚类集合的代表基因，$\| \|_2$表示取范数运算，下标表示范数类型为欧几里得范数；m_k表示第k个聚类中心，K表示所述第一预设参数；g_i表示所述聚类集合中的基因表达数据。

S204: 取所述第二样本矩阵的每一列组成第二训练集。

S205: 生成与所述第二训练集对应的特征索引集。

S206: 选取 N 折交叉验证法对所述基因特征数据集进行处理的过程中最高识别率时特征基因的集合的大小作为所述第二预设参数的值，采用 SVM-RFE 方法对所述第二训练集进行特征排序，得到一个与所述排序后的第二训练集对应的有顺序的特征索引集;

S207: 在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集。

S208: 采用支持向量机分类器对所述第三训练集进行建模，获得模型函数。

S209: 根据所述特征索引集对所述待测基因的表达数据进行特征选择，获得第一次特征选择后的样本。

其中，所述待测基因与所述第一训练集中的基因表达数据通过同一次 DNA 微阵列技术采集获得。

S210: 根据所述有顺序的特征索引集，选择所述第一次特征选择后的样本中的前第二预设参数个特征组成第二次特征选择后的样本。

S211: 将所述第二次特征选择后的样本输入所述模型函数中，获得所述模型函数的输出结果，根据所述输出结果获得所述待测基因表达数据的分类结果。

在上述实施例的基础上，在本申请的一个具体实施例中，在乳腺癌数据集

上对本申请实施例提供的基因表达数据分类方法进行了测试，所述乳腺癌数据集包括97名患者样本，属于两个类别。每个样本都有24481个基因表达数据。所述第一训练集包括78个患者样本，其中34个是在至少5年内的癌细胞转移的患者（标记为“旧病复发”），其余44个样本是从患者初步诊断为至少5年后仍健康的患者（标记为“非复发”）。相应地，待测基因样本中包括12个“旧病复发”患者样本和7个“非复发”患者样本。

具体测试步骤如下：

通过DNA微阵列技术获取所述乳腺癌数据集中的基因表达数据构成第一训练集$\{\mathbf{x}_i, y_i\}_{i=1}^{N}$，其中，$\mathbf{x}_i$为构成所述第一训练集的基因表达数据，$\mathbf{x}_i \in R^D$，$y_i$是$\mathbf{x}_i$的标签，表明$\mathbf{x}_i$的类别，$y_i \in \{-1,+1\}$，$N$代表所述第一训练集中样本的总个数，$D$代表所述第一训练集中样本的维数，R 代表实数集。这里 N=97，D=24481。

对所述第一训练集进行预处理，生成所述第一样本矩阵$\mathbf{X}=[\mathbf{x}_1,\cdots,\mathbf{x}_N]$，该矩阵的每一列为所述第一训练集的一个样本，取所述第一样本矩阵的每一行，即$\mathbf{g}_j=\left(X_{j\cdot}\right)^T$，$j=1,\cdots,24481$，形成所述基因特征数据集$\{\mathbf{g}_j\}_{j=1}^{24481}$，$\mathbf{g}_j \in R^{97}$。

确定聚类中心数目为第一预设参数K，此处K=80（通过10折交叉验证法选定），采用K-means聚类算法对所述基因特征数据集$\{\mathbf{g}_j\}_{j=1}^{24481}$进行聚类，获得80个聚类中心$\mathrm{m_k}$和80个聚类集合$\mathrm{G_k}$，$k=1,\cdots,K$。

在每个聚类集合中选取一个基因表达数据作为该聚类集合的代表基因，所述代表基因的选取公式为：$\bar{\mathbf{g}}_k=\min_{\mathbf{g}_i\in G_k}\|\mathbf{g}_i-\mathbf{m}_k\|_2^2, k=1,\cdots,80$；其中，$\|\ \|_2$表示取范数运算，下标表示范数类型为欧几里得范数，$\mathrm{G_k}$表示第k个聚类集合，$\bar{\mathbf{g}}_k$表示所述第k个聚类集合的代表基因，$\mathrm{m_k}$表示第k个聚类中心。生成所述样本矩阵$\mathbf{X}'=[\bar{\mathbf{g}}_1,\cdots,\bar{\mathbf{g}}_{80}]^T \in R^{80\times 97}$，其中N表示训练集中训练样本的总个数。取所述样本矩阵$\mathbf{X}'$的每一列$\bar{\mathbf{g}}_i$作为$\mathbf{x}_i'$，组成所述第二训练集$\{\mathbf{x}_i', y_i\}_{i=1}^{97}$，其中$\mathbf{x}_i' \in R^{80}$，同时生成与所述第二训练集对应的特征索引集$F \subset \{1,\cdots,24481\}$，$|F|=80$。

确定第二预设参数的值d，所述第二预设参数d（d＜80）的值为采用10折交叉验证法对所述基因特征数据集进行处理的过程中，最高识别率时特征基因的集合的大小。采用SVM-RFE方法对所述第二训练集$\{\mathbf{x}_i', y_i\}_{i=1}^{N}$进行特征排序，得到一个有顺序的特征索引集$F' \subset \{1,\cdots,24481\}$，$|F'|=80$。在所述有顺序的特征

索引集中选取前d个特征组成第三训练集$\{\mathbf{x}_i'', y_i\}_{i=1}^{97}$，其中$\mathbf{x}_i'' \in R^d$，在本实施例中，d=37。

采用支持向量机分类器对所述第三训练集$\{\mathbf{x}_i'', y_i\}_{i=1}^{97}$进行建模，获得模型函数$f(\mathbf{x}'')$。

令待测基因表达数据（在本实施例中为癌症基因表达数据）为 x，其中$\mathbf{x} \in R^{24481}$。

根据所述特征索引集 F 对所述待测基因$\mathbf{x}$（$\mathbf{x} \in R^D$）的表达数据进行特征选择，获得第一次特征选择后的样本$\mathbf{x}'$（$\mathbf{x}' \in R^K$）;

根据所述有顺序的特征索引集F'选择所述第一次特征选择后的样本$\mathbf{x}'$中的前第二预设参数个特征组成第二次特征选择后的样本$\mathbf{x}''$（$\mathbf{x}'' \in R^d$）;

将所述第二次特征选择后的样本$\mathbf{x}''$输入所述模型函数$f(\mathbf{x}'')$中，获得所述模型函数的输出结果，根据所述输出结果获得所述待测基因表达数据的分类结果。

采用相同的上述测试过程对本申请实施例提供的基因表达数据分类方法与 SVM-RFE（SVM-Recursive Feature Elimination）算法以及 MRMR+SVM-RFE(minimal redundancy-maximal relevance + SVM-Recursive Feature Elimination)算法在相同的乳腺癌数据集上进行对比实验。随机选取 78 个训练样本 10 次，表 1 给出了上述三种方法各自获得的最好平均分类性能时的对比。

表 1 SVM-RFE, MRMR+SVM-RFE 和本申请提供的基因表达数据分类方法最好分类性能的对比

方法	保留特征数	正类识别率(%)	负类识别率(%)	平均识别率(%)	时间(s)
本申请提供的基因表达数据分类方法	37	**85.71**	**100.00**	**94.74**	**17.175**

SVM-RFE	29	57.14	**100.00**	84.21	4764.271
MRMR+SVM-RFE	14	53.57	85.42	73.68	791.034

通过表1的对比可以发现，本申请实施例提供的基因表达数据分类方法比其他两种算法在各项参数上都有较大提升，且大大降低了时间消耗。

相应的，本申请实施例还提供了一种基因表达数据分类系统，如图3所示，包括：

特征选择模块A10，用于获取第一训练集，利用所述第一训练集生成基因特征数据集，所述第一训练集包含基因表达数据；采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心；利用所有的所述聚类集合的代表基因生成第二样本矩阵，其中，所述代表基因为每个所述聚类集合中的一个基因表达数据；对所述第二样本矩阵进行处理获得第二训练集，生成与所述第二训练集对应的特征索引集；对所述第二训练集进行特征排序，获得与所述排序后的第二训练集对应的有顺序的特征索引集；在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集；

训练模块A20，用于对所述第三训练集进行建模，获得模型函数；

诊断模块A30，用于根据所述特征索引集、有顺序的特征索引集和所述模型函数对待测基因表达数据进行分类，获得待测基因表达数据的分类结果。

需要说明的是，在本实施例中，所述待测基因表达数据与所述第一训练集采用生物微阵列技术在同一次采集中获得。

所述基因表达数据分类系统在获得所述基因特征数据集之后，采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个聚类集合，每个所述聚类集合具有一个聚类中心，然后对所述聚类集合进行处理获得第二样本矩阵、第二训练集以及特征索引集以对基因表达数据进行降维，从而降低基因表达数据之间的冗余度，进而在很大程度上降低了在之后对所述第二训练集进行特征选择的过程中占用的计算资源以及耗费的计算时间；而采用聚类算法对所述基因特征数据集进行聚类操作所占用的计算资源以及耗费的计算时间都很少，从而大大降低了对待测基因表达数据进行分类占用的计算资源和耗费的计

算时间。

在上述实施例的基础上，在本申请的一个实施例中，如图4所示，所述特征选择模块A10包括:

预处理单元 A11，用于获取基因样本的第一训练集，对所述第一训练集进行预处理，生成第一样本矩阵，并根据所述样本矩阵生成基因特征数据集;

第一特征选择单元 A12，用于采用 N 折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数，其中 N 为 5 或 10 或 20; 采用 K-means 聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心，在每个聚类集合中选取一个基因作为该聚类集合的代表基因，并利用所有的聚类集合的代表基因生成第二样本矩阵，并对所述第二样本矩阵进行处理获得第二训练集，同时生成与所述第二训练集对应的特征索引集;

第二特征选择单元 A13，用于对所述第二训练集进行特征排序，获得有顺序的特征索引集，确定保留的特征个数为第二预设参数，在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集。

需要说明的是，在本实施例中，所述预处理单元A11通过DNA微阵列技术获取基因表达数据的第一训练集$\{\mathbf{x}_i, y_i\}_{i=1}^{N}$，其中，$\mathbf{x}_i$表示所述第一训练集中的基因表达数据，$\mathbf{x}_i \in R^D$，$y_i$是$\mathbf{x}_i$的标签，表明$\mathbf{x}_i$的类别，$y_i \in \{-1,+1\}$，$N$代表所述第一训练集中样本的总个数，$D$代表所述第一训练集中样本的维数，R代表实数集。对所述第一训练集进行预处理，生成所述第一样本矩阵$\mathbf{X}=[\mathbf{x}_1,\cdots,\mathbf{x}_N]$，该矩阵的每一列为所述第一训练集的一个样本，取所述第一样本矩阵的每一行，即$\mathbf{g}_j=\left(X_{j\cdot}\right)^T$，$j=1,\cdots,D$，形成所述基因特征数据集$\{\mathbf{g}_j\}_{j=1}^{D}$，$\mathbf{g}_j \in R^N$。

在本实施例中，采用K-means聚类算法对所述基因特征数据集进行聚类的目的是对基因表达数据进行降维，以降低基因表达数据之间的冗余度。

在上述实施例的基础上，在本申请的另一个实施例中，所述第一特征选择单元 A12 用于采用 N 折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数 K，其中 N 为 5 或 10 或 20;

采用K-means聚类算法对所述基因特征数据集进行聚类，获得K个聚类中心m_k和K个聚类集合G_k，$k=1,\cdots,K$。

在每个聚类集合中选取一个基因作为该聚类集合的代表基因，所述代表基因的生成公式为：$\bar{\mathbf{g}}_k = \min_{\mathbf{g}_i \in G_k} \|\mathbf{g}_i - \mathbf{m}_k\|_2^2, k = 1, \cdots, K$ ；其中，G_k表示第k个聚类集合，$\bar{\mathbf{g}}_k$表示所述第k个聚类集合的代表基因，m_k表示第k个聚类中心，K表示所述第一预设参数。生成所述样本矩阵 $\mathbf{X}' = [\bar{\mathbf{g}}_1, \cdots, \bar{\mathbf{g}}_K]^T \in R^{K \times N}$ ，其中N表示训练集中训练样本的总个数。取所述样本矩阵$\mathbf{X}'$的每一列为$\mathbf{x}_i'$，组成所述第二训练集 $\{\mathbf{x}_i', y_i\}_{i=1}^N$ ，其中 $\mathbf{x}_i' \in R^K$ ，同时生成与所述第二训练集对应的特征索引集 $F \subset \{1, \cdots, D\}$ ， $|F| = K$ 。

需要说明的是，在本申请的一个优选实施例中，采用10折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数。但本申请对此并不做限定，具体视实际情况而定。

在上述实施例的基础上，在本申请的又一个实施例中，所述第二预设参数d（d＜K）的值为采用10折交叉验证法对所述基因特征数据集进行处理的过程中，最高识别率时特征基因的集合的大小。在本申请的一个优选实施例中，采用SVM-RFE方法对所述第二训练集合$\{\mathbf{x}_i', y_i\}_{i=1}^N$进行特征排序，得到一个有顺序的特征索引集 $F' \subset \{1, \cdots, D\}$ ， $|F'| = K$ 。在所述有顺序的特征索引集中选取前d个特征组成第三训练集$\{\mathbf{x}_i'', y_i\}_{i=1}^N$，其中 $\mathbf{x}_i'' \in R^d$ 。

在上述实施例的基础上，在本申请的再一个实施例中，采用支持向量机分类器对所述第三训练集进行建模，获得模型函数 $f(\mathbf{x}'')$ 。但本申请对所述第三训练集进行建模所采用的具体方法并不做限定，具体视实际情况而定。

在上述实施例的基础上，在本申请的一个具体实施例中，如图5所示，所述诊断模块A30包括:

第一选择单元A31，用于 根据所述特征索引集F对所述待测基因$\mathbf{x}$（ $\mathbf{x} \in R^D$ ）的表达数据进行特征选择，获得第一次特征选择后的样本$\mathbf{x}'$（ $\mathbf{x}' \in R^K$ ）;

第二选择单元A32，用于 根据所述有顺序的特征索引集F'选择所述第一次特征选择后的样本$\mathbf{x}'$中的前第二预设参数个特征组成第二次特征选择后的样本$\mathbf{x}''$（ $\mathbf{x}'' \in R^d$ ）;

诊断单元A33，用于将所述第二次特征选择后的样本$\mathbf{x}''$输入所述模型函数 $f(\mathbf{x}'')$ 中，获得所述模型函数的输出结果，根据所述输出结果获得所述待测基因表达数据的分类结果。

综上所述，本申请实施例提供了一种基因表达数据分类方法及分类系统，其中，所述基因表达数据分类方法中在获得所述基因特征数据集之后，采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，然后对所述聚类集合进行处理获得第二样本矩阵、第二训练集以及特征索引集以对基因表达数据进行降维，从而降低基因表达数据之间的冗余度，进而在很大程度上降低了在之后对所述第二训练集进行特征选择的过程中占用的计算资源以及耗费的计算时间；而采用聚类算法对所述基因特征数据集进行聚类操作所占用的计算资源以及耗费的计算时间都很少，因此采用所述基因表达数据分类方法对待测基因表达数据进行分类占用的计算资源和耗费的计算时间都较少。

本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。

对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

权 利 要 求

1、一种基因表达数据分类方法，其特征在于，包括:

获取第一训练集，利用所述第一训练集生成基因特征数据集，所述第一训练集中包含基因表达数据;

采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心;

利用所有的所述聚类集合的代表基因生成第二样本矩阵，其中，所述 代表基因为每个所述聚类集合中的一个基因表达数据;

对所述第二样本矩阵进行处理获得第二训练集;

生成与所述第二训练集对应的特征索引集;

对所述第二训练集进行特征排序，获得与所述排序后的第二训练集对应的有顺序的特征索引集;

在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集;

对所述第三训练集进行建模，获得模型函数;

根据所述特征索引集、有顺序的特征索引集和所述模型函数对待测基因表达数据进行分类，获得待测基因表达数据的分类结果。

2、根据权利要求 1 所述的基因表达数据分类方法，其特征在于，所述第一预设参数的设定方式为:

采用 N 折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数，其中 N 为 5、10 或 20。

3、根据权利要求 1 所述的基因表达数据分类方法，其特征在于，所述采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心，包括:

采用 K-means 聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心。

4、根据权利要求 1 所述的基因表达数据分类方法，其特征在于，所述代表基因通过公式 $\overline{\mathbf{g}}_k = \min_{\mathbf{g}_i \in G_k} \|\mathbf{g}_i - \mathbf{m}_k\|_2^2, k = 1, \cdots, K$ 生成;

其中，G_k表示第k个聚类集合，$\bar{\mathbf{g}}_k$表示所述第k个聚类集合的代表基因，m_k表示第k个聚类中心，K表示所述第一预设参数；g_i表示所述聚类集合中的基因表达数据；

其中，所述第二样本矩阵为 $\mathbf{X}'=[\bar{\mathbf{g}}_1,\cdots,\bar{\mathbf{g}}_K]^T \in R^{K\times N}$，其中，R表示实数集，N表示所述第一训练集中的样本总个数。

5、根据权利要求1所述的基因表达数据分类方法，其特征在于，根据所述特征索引集、有顺序的特征索引集和所述模型函数对待测基因表达数据进行分类，获得待测基因表达数据的诊断结果包括：

根据所述特征索引集对所述待测基因的表达数据进行特征选择，获得第一次特征选择后的样本；

根据所述有顺序的特征索引集，选择所述第一次特征选择后的样本中的前第二预设参数个特征组成第二次特征选择后的样本；

将所述第二次特征选择后的样本输入所述模型函数中，获得所述模型函数的输出结果，根据所述输出结果获得所述待测基因表达数据的分类结果。

6、一种基因表达数据分类系统，其特征在于，包括：

特征选择模块，用于获取第一训练集，利用所述第一训练集生成基因特征数据集，所述第一训练集包含基因表达数据；采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心；利用所有的所述聚类集合的代表基因生成第二样本矩阵，其中，所述代表基因为每个所述聚类集合中的一个基因表达数据；对所述第二样本矩阵进行处理获得第二训练集，生成与所述第二训练集对应的特征索引集；对所述第二训练集进行特征排序，获得与所述排序后的第二训练集对应的有顺序的特征索引集；在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集；

训练模块，用于对所述第三训练集进行建模，获得模型函数；

诊断模块，用于根据所述特征索引集、有顺序的特征索引集和所述模型函数对待测基因表达数据进行分类，获得待测基因表达数据的分类结果。

7、根据权利要求6所述的基因表达数据分类系统，其特征在于，所述特征选择模块包括：

预处理单元，用于获取基因样本的第一训练集，对所述第一训练集进行预处理，生成第一样本矩阵，并根据所述样本矩阵生成基因特征数据集；

第一特征选择单元，用于采用 N 折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数，其中 N 为 5 或 10 或 20；采用 K-means 聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心，在每个聚类集合中选取一个基因作为该聚类集合的代表基因，并利用所有的聚类集合的代表基因生成第二样本矩阵，并对所述第二样本矩阵进行处理获得第二训练集，同时生成与所述第二训练集对应的特征索引集；

第二特征选择单元，用于对所述第二训练集进行特征排序，获得有顺序的特征索引集，确定保留的特征个数为第二预设参数，在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集。

8、根据权利要求 7 所述的基因表达数据分类系统，其特征在于，所述第一特征选择单元用于采用 N 折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数，其中 N 为 5 或 10 或 20；采用 K-means 聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个聚类中心和第一预设参数个聚类集合，在每个聚类集合中选取一个基因作为该聚类集合的代表基因，并利用所有的聚类集合的代表基因生成第二样本矩阵，并取所述第二样本矩阵的每一列 $\mathbf{x}'_i$ 组成所述第二训练集 $\{\mathbf{x}'_i, y_i\}_{i=1}^{N}$，其中，$\mathbf{x}'_i \in R^K$。

9、根据权利要求 6 所述的基因表达数据分类系统，其特征在于，所述代表基因通过公式 $\overline{\mathbf{g}}_k = \min_{\mathbf{g}_i \in G_k} \|\mathbf{g}_i - \mathbf{m}_k\|_2^2, k = 1, \cdots, K$ 生成，其中，G_k表示第k个聚类集合，$\overline{\mathbf{g}}_k$ 表示所述第k个聚类集合的代表基因，m_k表示第k个聚类中心，K表示所述第一预设参数；g_i表示所述聚类集合中的基因表达数据；

其中，所述第二样本矩阵为 $\mathbf{X}' = [\overline{\mathbf{g}}_1, \cdots, \overline{\mathbf{g}}_K]^T \in R^{K \times N}$，其中 R 表示实数集，N 表示生成所述第一训练集中的样本总个数。

10、根据权利要求 6 所述的基因表达数据分类系统，其特征在于，所述诊断模块包括：

第一选择单元，用于根据所述特征索引集对所述待测基因表达数据进行特征选择，获得第一次特征选择后的样本；

第二选择单元，用于根据所述有顺序的特征索引集选择所述第一次特征选择后的样本中的前第二预设参数个特征组成第二次特征选择后的样本；

诊断单元，用于将所述第二次特征选择后的样本输入所述模型函数中，获得所述模型函数的输出结果，根据所述输出结果获得所述待测基因表达数据的分类结果。

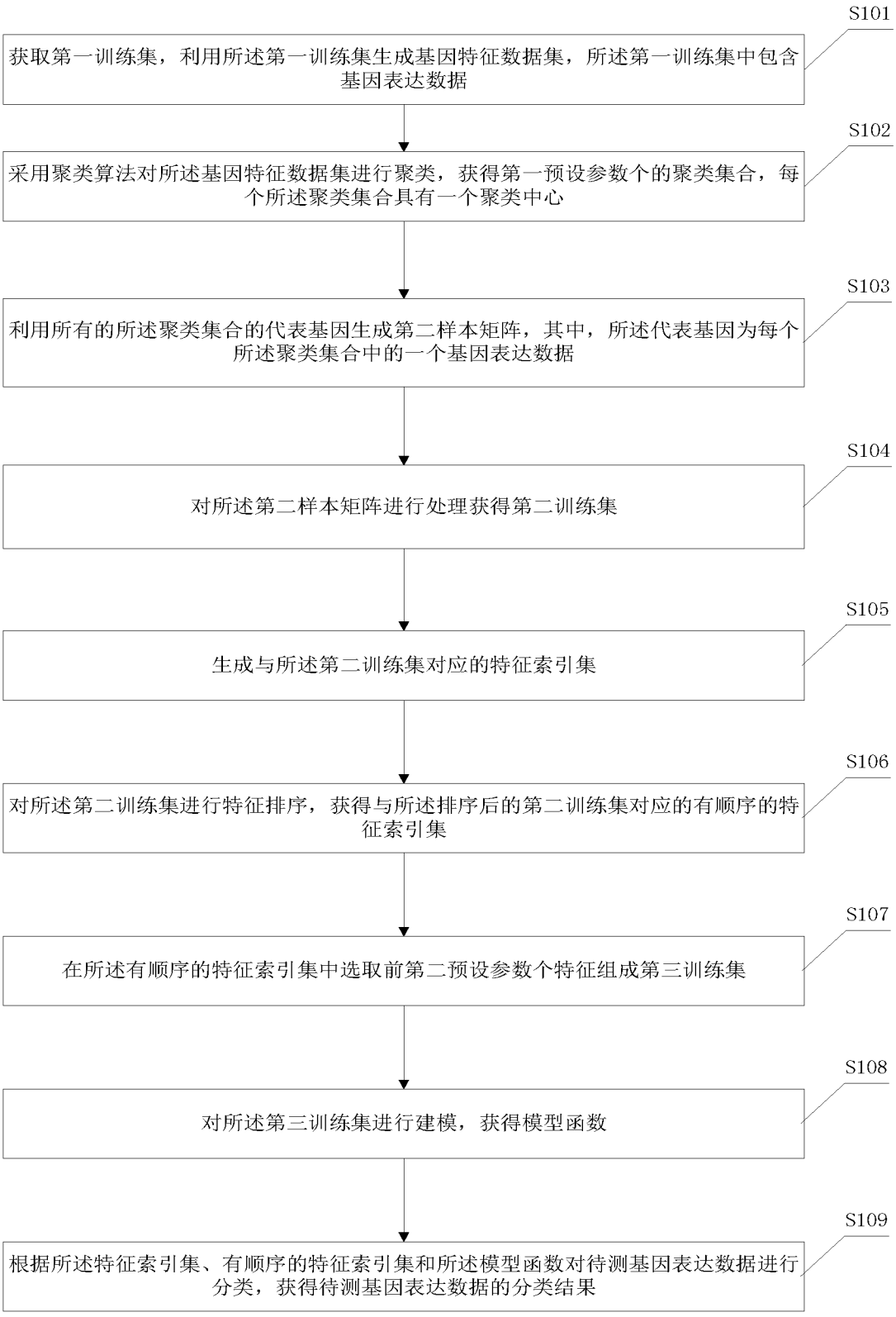

S101
获取第一训练集，利用所述第一训练集生成基因特征数据集，所述第一训练集中包含基因表达数据
S102
采用聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心
S103
利用所有的所述聚类集合的代表基因生成第二样本矩阵，其中，所述代表基因为每个所述聚类集合中的一个基因表达数据
S104
对所述第二样本矩阵进行处理获得第二训练集
S105
生成与所述第二训练集对应的特征索引集
S106
对所述第二训练集进行特征排序，获得与所述排序后的第二训练集对应的有顺序的特征索引集
S107
在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集
S108
对所述第三训练集进行建模，获得模型函数
S109
根据所述特征索引集、有顺序的特征索引集和所述模型函数对待测基因表达数据进行分类，获得待测基因表达数据的分类结果

图 1

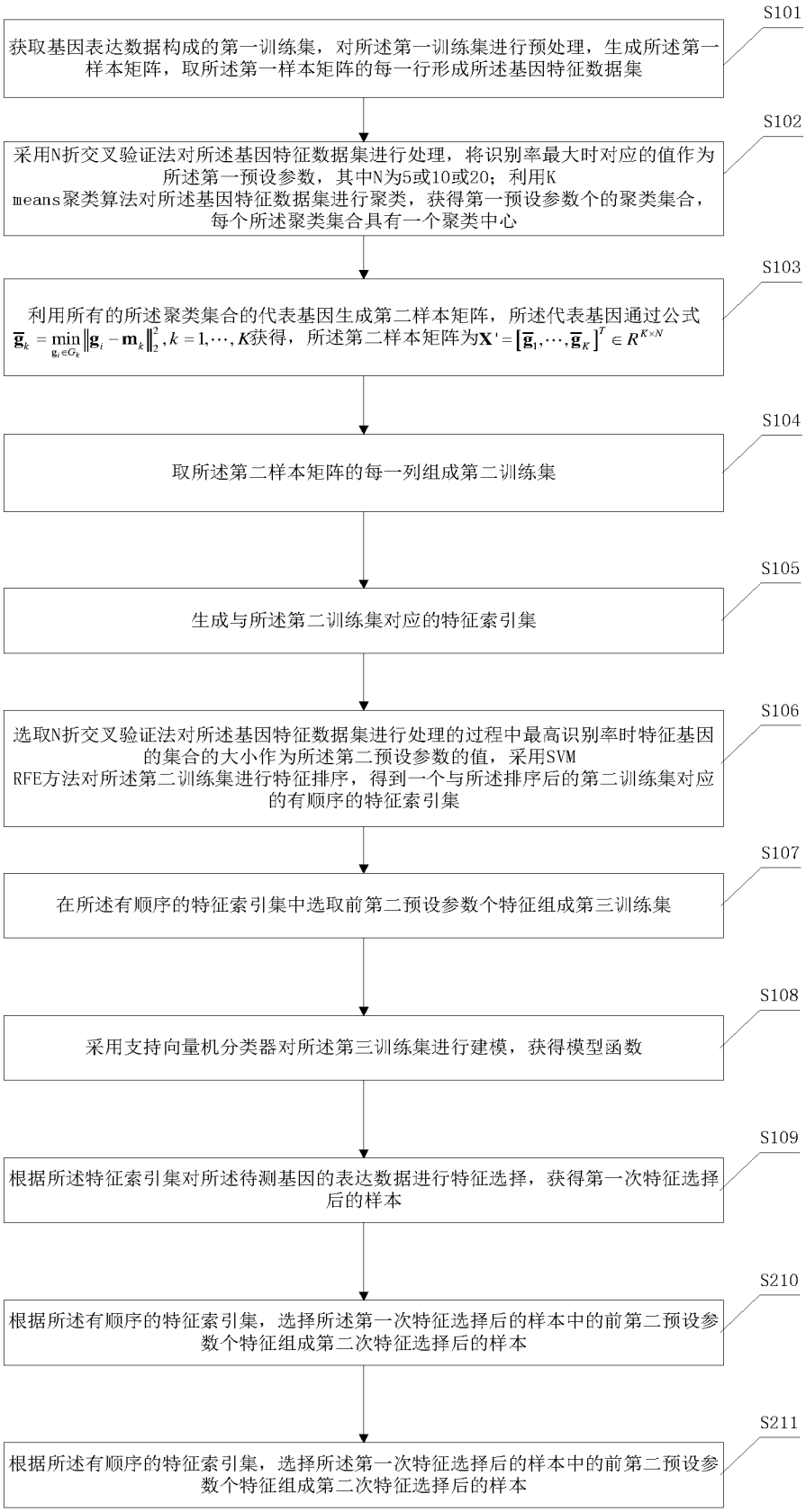

S101
获取基因表达数据构成的第一训练集，对所述第一训练集进行预处理，生成所述第一样本矩阵，取所述第一样本矩阵的每一行形成所述基因特征数据集
S102
采用N折交叉验证法对所述基因特征数据集进行处理，将识别率最大时对应的值作为所述第一预设参数，其中N为5或10或20；利用K means聚类算法对所述基因特征数据集进行聚类，获得第一预设参数个的聚类集合，每个所述聚类集合具有一个聚类中心
S103
利用所有的所述聚类集合的代表基因生成第二样本矩阵，所述代表基因通过公式 $\overline{\mathbf{g}}_k = \min_{\mathbf{g}_i \in G_k} \|\mathbf{g}_i - \mathbf{m}_k\|_2^2, k = 1, \cdots, K$ 获得，所述第二样本矩阵为 $\mathbf{X}' = [\overline{\mathbf{g}}_1, \cdots, \overline{\mathbf{g}}_K]^T \in R^{K \times N}$
S104
取所述第二样本矩阵的每一列组成第二训练集
S105
生成与所述第二训练集对应的特征索引集
S106
选取N折交叉验证法对所述基因特征数据集进行处理的过程中最高识别率时特征基因的集合的大小作为所述第二预设参数的值，采用SVM RFE方法对所述第二训练集进行特征排序，得到一个与所述排序后的第二训练集对应的有顺序的特征索引集
S107
在所述有顺序的特征索引集中选取前第二预设参数个特征组成第三训练集
S108
采用支持向量机分类器对所述第三训练集进行建模，获得模型函数
S109
根据所述特征索引集对所述待测基因的表达数据进行特征选择，获得第一次特征选择后的样本
S210
根据所述有顺序的特征索引集，选择所述第一次特征选择后的样本中的前第二预设参数个特征组成第二次特征选择后的样本
S211
根据所述有顺序的特征索引集，选择所述第一次特征选择后的样本中的前第二预设参数个特征组成第二次特征选择后的样本

图 2

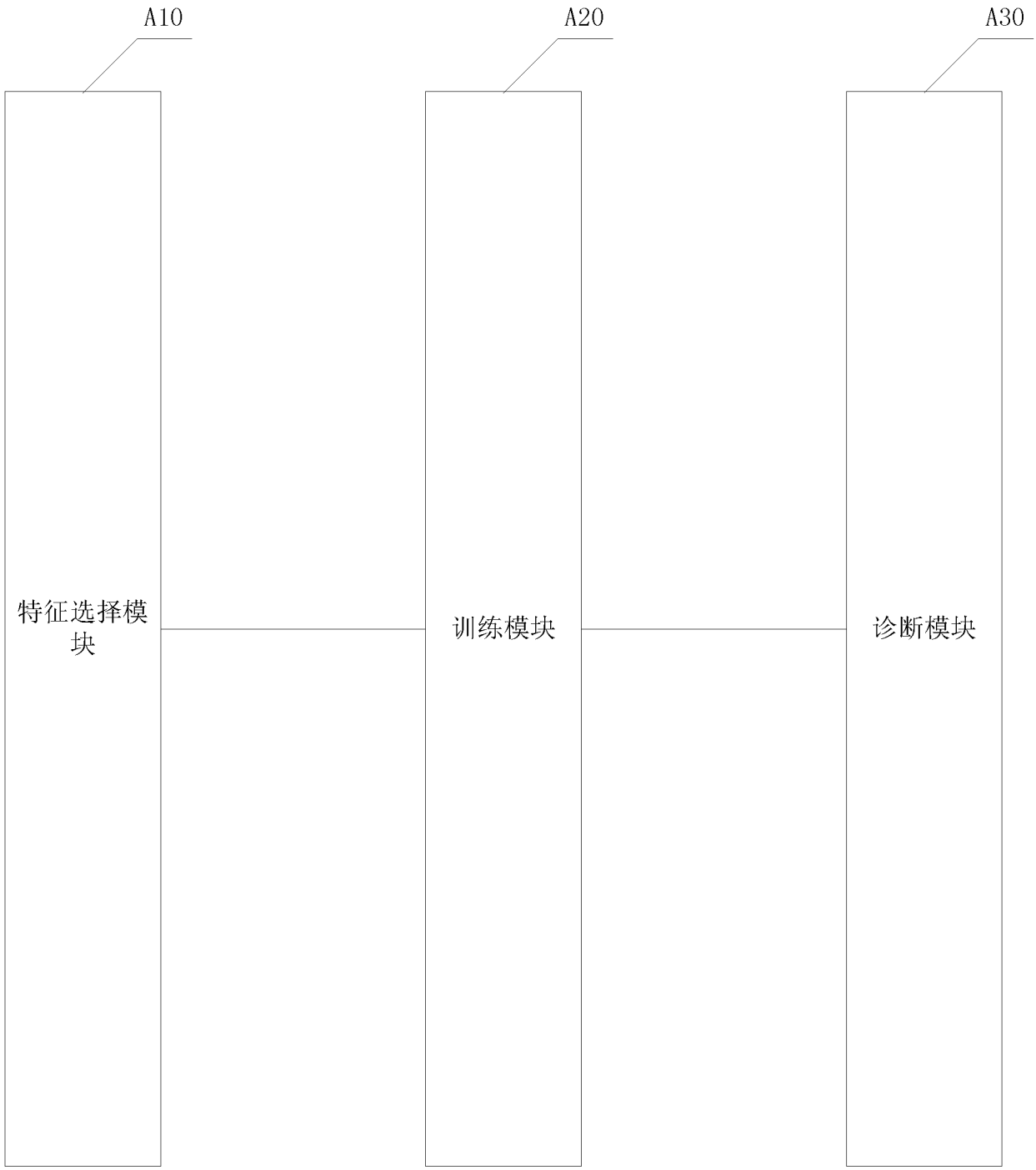


图 3

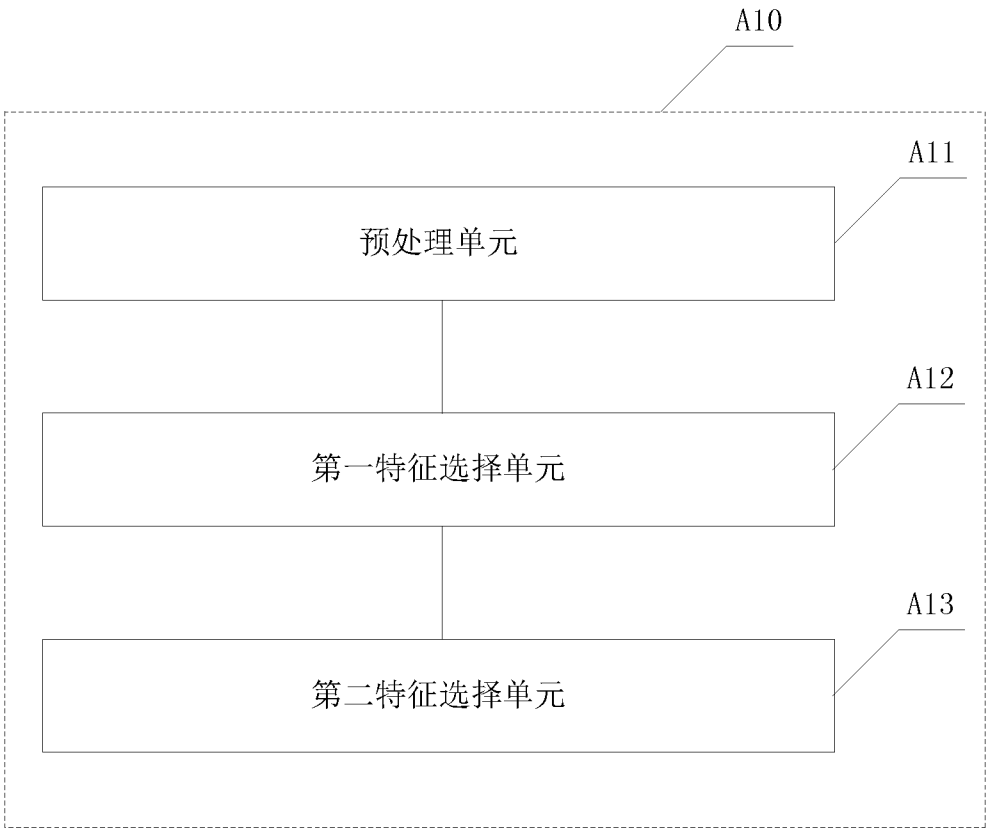


图 4

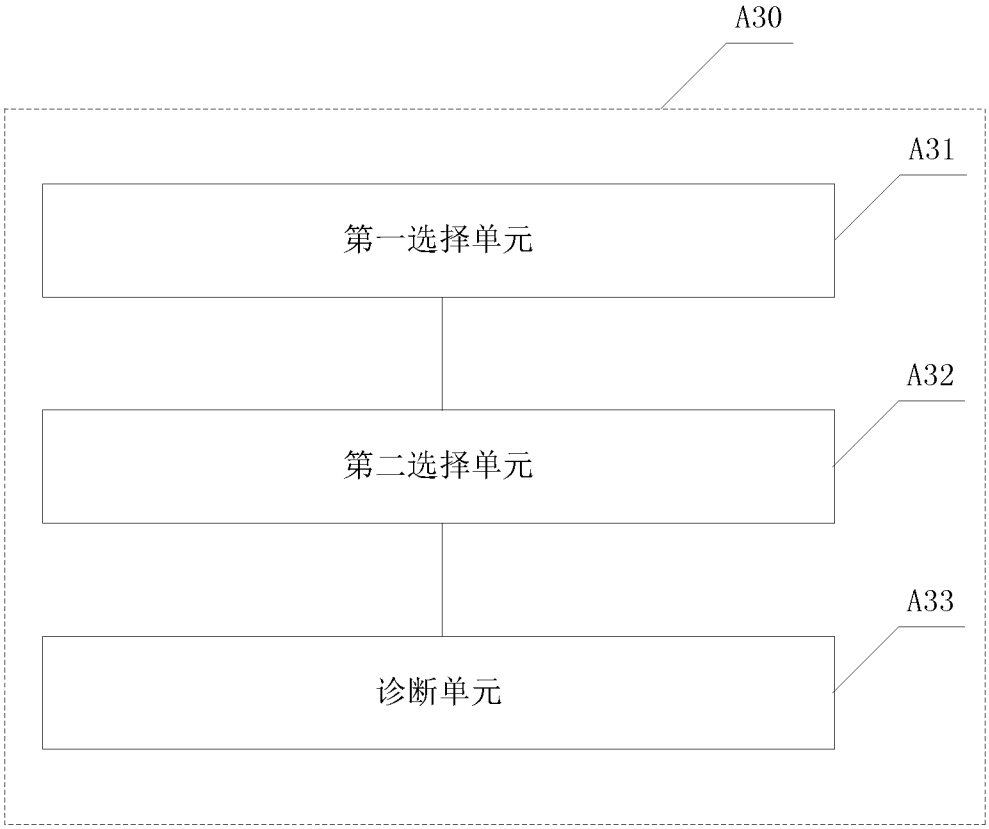


图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/106255

A. CLASSIFICATION OF SUBJECT MATTER

G06F 19/24 (2011.01) 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, CNKI, DWPI, IEEE: gather, sequencing, gene, cluster, classify, sort, model

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 105825081 A (SOOCHOW UNIVERSITY), 03 August 2016 (03.08.2016), claims 1-10	1-10
A	CN 102272764 A (ROYAL DUTCH PHILIPS ELECTRONICS LTD.), 07 December 2011 (07.12.2011), the whole document	1-10
A	CN 104200134 A (BEIJING UNIVERSITY OF TECHNOLOGY), 10 December 2014 (10.12.2014), the whole document	1-10
A	WO 2001073428 A1 (SHAMIR, R. et al.), 04 October 2001 (04.10.2001), the whole document	1-10

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 February 2017 (08.02.2017)	**15 February 2017 (15.02.2017)**
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer **TANG, Yuxi** Telephone No.: (86-10) **62411854**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/106255

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105825081 A	03 August 2016	None	
CN 102272764 A	07 December 2011	US 2012016826 A1	19 January 2012
		KR 101642270 B1	26 July 2016
		EP 2387758 B1	29 May 2013
		RU 2011133091 A	20 February 2013
		US 8712935 B2	29 April 2014
		EP 2207119 A1	14 July 2010
		KR 20110112833 A	13 October 2011
		WO 2010079402 A1	15 July 2010
		EP 2387758 A1	23 November 2011
		JP 5674679 B2	25 February 2015
		CN 102272764 B	28 January 2015
		JP 2012514783 A	28 June 2012
		BR PI0918672 A2	23 August 2016
		IN 201105636 P4	30 November 2012
CN 104200134 A	10 December 2014	None	
WO 2001073428 A1	04 October 2001	None	

国际检索报告

国际申请号
PCT/CN2016/106255

A. 主题的分类

G06F 19/24(2011.01)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,CNKI,DWPI,IEEE:基因，聚类，聚集，分类，建模，排序，模型，gene， cluster， classify， sort， model

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 105825081 A (苏州大学) 2016年 8月 3日 (2016-08-03) 权利要求1-10	1-10
A	CN 102272764 A (皇家飞利浦电子股份有限公司) 2011年 12月 7日 (2011-12-07) 全文	1-10
A	CN 104200134 A (北京工业大学) 2014年 12月 10日 (2014-12-10) 全文	1-10
A	WO 2001073428 A1 (SHAMIR， RON等) 2001年 10月 4日 (2001-10-04) 全文	1-10

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

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

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

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

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

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

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2017年 2月 8日	2017年 2月 15日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(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 唐宇希 电话号码 (86-10)62411854

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/106255

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	105825081	A	2016年 8月 3日	无			
CN	102272764	A	2011年 12月 7日	US	2012016826	A1	2012年 1月 19日
				KR	101642270	B1	2016年 7月 26日
				EP	2387758	B1	2013年 5月 29日
				RU	2011133091	A	2013年 2月 20日
				US	8712935	B2	2014年 4月 29日
				EP	2207119	A1	2010年 7月 14日
				KR	20110112833	A	2011年 10月 13日
				WO	2010079402	A1	2010年 7月 15日
				EP	2387758	A1	2011年 11月 23日
				JP	5674679	B2	2015年 2月 25日
				CN	102272764	B	2015年 1月 28日
				JP	2012514783	A	2012年 6月 28日
				BR	PI0918672	A2	2016年 8月 23日
				IN	201105636	P4	2012年 11月 30日
CN	104200134	A	2014年 12月 10日	无			
WO	2001073428	A1	2001年 10月 4日	无			