(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 111640468 B
(45)授权公告日 2021.08.24

(21)申请号 202010418499.3
(22)申请日 2020.05.18
(65)同一申请的已公布的文献号
申请公布号 CN 111640468 A
(43)申请公布日 2020.09.08
(73)专利权人 天士力国际基因网络药物创新中心有限公司
地址 300451 天津市滨海新区自贸试验区(东疆保税港区)亚洲路6865号金融贸易中心北区1-1-2108-3
(72)发明人 李旭 任静 王学敏 张文闫凯境 王文佳
(74)专利代理机构 北京君尚知识产权代理有限公司 11200
代理人 司立彬
(51)Int.Cl.
G16B 40/20(2019.01)
G16B 25/10(2019.01)
G06N 20/00(2019.01)
(56)对比文件
CN 110400600 A,2019.11.01
CN 109411033 A,2019.03.01
Peng Yang 等.Ensemble Positive Unlabeled Learning for Disease Gene Identification.《PLOS ONE》.2014,第9卷(第5期),
Aditya Grover 等.node2vec: Scalable Feature Learning for Networks.《arXiv》.2016,
Qi Zhao等.SSCMDA: spy and super cluster strategy for MiRNA-disease association prediction.《Oncotarget》.2017,第9卷(第2期),
审查员 姜蝶

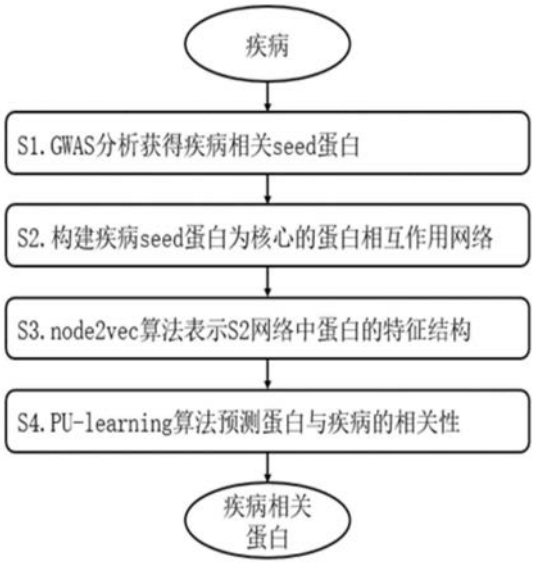
权利要求书1页 说明书6页 附图1页

(54)发明名称

一种基于复杂网络筛选疾病相关蛋白的方法

(57)摘要

本发明公开了一种基于复杂网络筛选疾病相关蛋白的方法,本方法为:1)获取目标疾病相关的种子基因;2)基于蛋白-蛋白相互作用数据库,构建以该种子基因为核心的蛋白相互作用网络;3)提取该蛋白相互作用网络中蛋白的特征数据;4)将蛋白的所述特征数据作为训练数据,采用机器学习算法训练得到PU分类器;5)根据所述PU分类器预测该蛋白相互作用网络中与该目标疾病有关的蛋白。本发明的方法能够快速高效鉴定出与疾病相关的蛋白,有助于生物医学专家进行实验验证或相关研究人员开展工作。



CN 111640468 B

1.一种基于复杂网络筛选疾病相关蛋白的方法,其步骤包括:

1)获取目标疾病相关的种子基因;

2)基于蛋白-蛋白相互作用数据库,构建以该种子基因为核心的蛋白相互作用网络;

3)利用node2vec算法提取该蛋白相互作用网络中蛋白的特征数据;

4)将蛋白的所述特征数据作为训练数据,采用半监督算法PU-learning训练得到PU分类器;

5)根据所述PU分类器预测该蛋白相互作用网络中与该目标疾病有关的蛋白。

2.如权利要求1所述的方法,其特征在于,利用node2vec算法提取该蛋白相互作用网络中蛋白的特征数据,其方法为:

31)基于蛋白相互作用网络数据构建无向图G,获得节点集和边集;其中节点集中的每一节点对应一蛋白,边集中的边代表蛋白与蛋白之间相互作用关系;

32)利用node2vec算法对该无向图G进行图嵌入,得到蛋白相互作用网络中的蛋白特征。

3.如权利要求2所述的方法,其特征在于,所述node2vec算法通过两个超参数p和q控制概率反复经过节点;其中超参数p的取值范围为[2,5],超参数q的取值范围为[0.1,3]。

4.如权利要求3所述的方法,其特征在于,将所述蛋白特征表示为d维向量,维数d的取值范围为[128,256]。

5.如权利要求1所述的方法,其特征在于,采用半监督算法PU-learning训练得到PU分类器的方法为:

41)从正例标注集P中随机选取部分“间谍”样本得到一集合S,并将集合S放置于无标签的数据集U中;将P-S作为正例集合PS,U+S作为反例集合US,训练得到NB分类器对反例集合US中的每个蛋白样本进行分类;其中,正例标注集P为与目标疾病相关的种子蛋白集合,数据集U中的蛋白为与目标疾病无法确定相关性的蛋白;

42)重复步骤41),将满足设定要求的负样本作为高频稳定的负样本,构成一负样本集合RN;

43)对正例样本P和负样本集合RN迭代运行学习算法SVM,直到SVM收敛或达到设定停止条件,得到PU分类器。

6.如权利要求5所述的方法,其特征在于,将同一样本被分类为负样本的次数占重复总次数的比例超过设定阈值的样本作为高频稳定的负样本。

7.如权利要求1所述的方法,其特征在于,对目标疾病组人群和对照组人群进行全基因组扫描,通过全基因组关联分析获得该目标疾病相关的种子基因。

8.一种存储介质,所述存储介质中存储有计算机程序,其中,所述计算机程序被设置为运行时执行权利要求1-7中任一所述方法中各步骤的指令。

9.一种服务器,其特征在于,包括存储器和处理器,所述存储器存储计算机程序,所述计算机程序被配置为由所述处理器执行,所述计算机程序包括用于执行权利要求1至7任一所述方法中各步骤的指令。

一种基于复杂网络筛选疾病相关蛋白的方法

技术领域

[0001] 本发明涉及蛋白筛选技术领域，具体涉及一种基于复杂网络筛选疾病相关蛋白的方法。

背景技术

[0002] 疾病相关蛋白的识别在疾病的分子分型、诊断、治疗等方面发挥重要的作用。准确且高效的识别疾病相关蛋白有助于发现致病基因、鉴定药物的靶标，在疾病诊治和药物设计方面意义深远。GWAS作为探讨疾病易感基因的重要研究工具，能够快速发现较为显著的疾病易感位点。但GWAS对数据利用度不高，掩盖了大量可能具有显著性的疾病相关蛋白。同时，传统GWAS的单位点关联分析将机体内各个基因独立对待，忽略了生物体内基因间的相互作用，难以发现真正与疾病相关的蛋白。

[0003] 蛋白-蛋白相互作用(PPI)网络分析弥补了上述不足。近年来，随着PPI数据的日益完善，使用计算机网络和图论的理论及方法，从系统的角度研究蛋白相互作用网络，成为热门领域。不少学者逐渐转向基于计算的蛋白识别研究，提出了许多经典的算法，如度中心性(Degree Centrality，DC)、介数中心性(Betweenness Centrality，BC)、接近度中心性(Closeness Centrality，CC)等等，然而这些算法的识别准确率普遍不高。因此，如何基于蛋白互作网络获得蛋白的特征表示，并用于寻找与已知疾病相关蛋白功能相似的蛋白，是蛋白网络分析的难点。

[0004] 疾病相关蛋白的识别的另一难点在于，在以疾病种子基因为核心的蛋白相互作用网络中，有标签的疾病蛋白非常稀少，仅有种子基因为正样本，大量的无标签的蛋白与疾病的关系未知，即这些无标签的蛋白可能为与疾病相关的蛋白，也可能与疾病无关。通常，这些无标签的蛋白与疾病的相关性依赖于生物学实验或人工文献检索矫正，价格昂贵，耗时长。大数据和人工智能技术的飞速发展为疾病蛋白筛选提供了一条低成本、高效率的途径。通过机器学习技术预测出与疾病最相关的蛋白，从而对开发新药，临床治疗起到巨大的推动作用。然而，大量的无标签数据在机器学习过程中容易造成模型欠拟合或过拟合问题，导致模型不足以学习到整个样本空间中的信息或者模型范化能力不足。如何合理使用无标注数据构建模型，大大减轻对标注数据的需求，是亟待解决的技术难题。

发明内容

[0005] 针对现有技术中的缺陷与不足，本发明的目的在于提供一种基于复杂网络筛选疾病相关蛋白的方法。本发明的方法能够快速高效鉴定出与疾病相关的蛋白，有助于生物医学专家进行实验验证或相关研究人员开展工作。

[0006] 本发明提供了一种基于复杂网络筛选疾病相关蛋白的方法。通过全基因组关联分析(GWAS)发现与疾病相关的种子基因，并基于蛋白相互作用数据库(如Biogrid、String、Intact、HPRD数据库等)获得以种子基因为核心的蛋白互作网络，使用node2vec算法提取蛋白互作网络中蛋白的特征数据，利用半监督学习PU-learning算法预测与蛋白互作网络中

与疾病有关的蛋白。具体步骤如下：

[0007] S1:采用病例-对照研究的方法,对目标疾病组人群和对照组人群进行全基因组扫描,全基因组关联分析(GWAS)获得该目标疾病相关的种子基因；

[0008] S2:基于蛋白-蛋白相互作用数据库,构建以该种子基因为核心的蛋白相互作用网络；

[0009] S3:基于node2vec算法提取该蛋白相互作用网络中蛋白的特征数据；

[0010] 所述步骤S3具体包括：

[0011] S31:将S2获得的蛋白相互作用网络数据,构建无向图G,获得节点集(蛋白集)和边集(蛋白-蛋白相互作用关系集)；

[0012] S32:利用node2vec方法将S2蛋白相互作用网络中的蛋白特征表示为d维向量,以表示蛋白之间的关系,进而研究网络中蛋白的内容相似性和结构相似性。

[0013] 在机器学习的具体实践任务中,特征选择的功能是减少特征数量、降维,提高模型性能,使模型泛化能力更强,减少过拟合。选择一组具有代表性的特征用于构建模型是非常重要的问题。针对现有的特征学习方法不能够捕捉到复杂网络(蛋白相互作用网络)连接的多样性模式,本发明利用node2vec算法通过二阶马尔科夫链模拟偏向随机游走过程代替深度搜索(DFS)和广度搜索(BFS),搜索节点的一阶邻居和多阶同构邻居实现节点邻居采样的多样化。其采样策略原理如下：

[0014] 给定当前顶点v,访问下一个顶点x的概率为：

[0015] $$P(c_i = x | c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & if\ (v,x) \in E \\ 0 & otherwise \end{cases}$$

[0016] 其中π_{vx}是顶点v和顶点x的转移概率,Z为归一化常数。Node2vec引入两个超参数p和q来控制随机游走策略,假设当前随机游走经过边(t,v)到达顶点v,令$\pi_{vx}=\alpha_{pq}(t,x)\cdot w_{vx}$,$w_{vx}$是顶点v和x之间的权重。则

[0017] $$\alpha_{pq}(t,x) = \begin{cases} 1/p & if\ d_{tx} = 0 \\ 1 & if\ d_{tx} = 1 \\ 1/q & if\ d_{tx} = 2 \end{cases},$$

[0018] 其中d_{tx}为顶点t和顶点x之间的最短路径距离。

[0019] S33:在S32中描述的二阶随机游走函数中,参数p控制重复访问刚刚访问过的顶点的概率,p越高,访问刚刚访问过的顶点的概率越低。参数q控制游走是向内还是向外,如果q>1,随机游走倾向访问和顶点t接近的顶点(偏向BFS)；如果q<1,倾向访问远离顶点t的顶点(偏向DFS)。本发明使用node2vec方法将S2蛋白相互作用网络中的蛋白特征表示为d维向量过程中,p的较优范围为[2,5],q的较优范围为[0.1,3],维数d的较优范围为[128,256]。

[0020] S4:将S3获得的蛋白结构化特征数据作为输入,利用半监督算法PU-learning算法进行分析,实现蛋白相互作用网络中与目标疾病有关的蛋白的预测。

[0021] PU-learning是一种用于解决二分类问题的半监督学习方法,与传统的方法不同的是PU-learning方法可以处理正例样本数目较少且负样本缺失的情况,从而提高疾病相关蛋白的预测性能。PU-learning的训练集是由正例和无标记样本构成,具体的,在S3获得

的蛋白相互作用网络中蛋白结构化特征数据中，仅有少数蛋白(种子基因)是已知与疾病相关的基因，为正例样本；与这些种子基因有相互作用关系的蛋白为无标记样本(可能与疾病相关，也可能与疾病无关)。

[0022] 所述步骤S4具体包括：

[0023] S41:使用两步法建立PU分类器：第一步，从未标注样本中找到可靠负样本；第二步，用确定的正样本和可靠的负样本训练分类器。

[0024] S411:具体的，S41中第一步寻找可靠负样本使用Spy technique(间谍技术)，从正例标注集P(与目标疾病相关的种子蛋白)中随机选取部分“间谍”样本(部分种子蛋白)S，并将其放置于无标签的数据集U(即与目标疾病关系无法确定的蛋白，为无标签的蛋白)中。将P-S看作正例集合(PS)，U+S看作反例集合(US)，运用朴素贝叶斯算法，训练得到NB分类器。通过NB分类器，对反例集合US中的每个蛋白k进行分类，即为每个蛋白k赋予一个概率列表标识Pr(1|k)，其中1标识正例类别。间谍集合S的概率标识决定无标签的蛋白集合U中哪些最有可能与疾病无关。具体的，通过间谍集合S的概率标识，给定一个阈值H，将蛋白的概率标识Pr(1|k)<H的蛋白视为负样(与疾病无关的蛋白)。

[0025] S412:将步骤S411重复100-1000次，得到高频稳定的负样本(与疾病无关蛋白)集合RN。本发明通过对算法进行改进，多次实施步骤S411，每次都会得到负样，然后选取高频次出现的蛋白为负样蛋白，而不是一次试验的结果。比如某一蛋白在100次试验中，有100次均被选出为负样蛋白，认为该蛋白属于稳定负样集合RN。

[0026] S413:具体的，S41中第二步使用SVM算法。对正例样本P(已知与疾病有关蛋白)和S43得到的负样(与疾病无关蛋白)集合RN迭代运行学习算法SVM直到它收敛，或达到设定停止条件，得到PU分类器。

[0027] S42:使用上述S41获得的PU分类器，预测蛋白互作网络中每个蛋白是与疾病有关蛋白的概率，实现与疾病有关的蛋白的预测。

[0028] S5:本发明涉及对预测得到的蛋白的评价指标为：准确率、精准率、召回率、F1值。理想情况下，精准率和召回率两个指标都高最好，但一般情况下，精准率高，召回率就低；召回率高，精准率就低。选择保证精准率的条件下，提升召回率。

[0029] S51:准确率大体上评价了模型的准确情况，公式为，$准确率=\frac{真阳性+真阴性}{样本数}$。

[0030] S52:精准率表明了分类器检测某蛋白为疾病相关蛋白的准确性，公式为，$精准率（\mathrm{Precision}）=\frac{真阳性}{真阳性+假阳性}$。

[0031] S53:召回率表明了分类器能否把所有的目标都检测出来，公式为，$召回率(\mathrm{Recall})=\frac{真阳性}{真阳性+假阴性}$。

[0032] S54:F1值是精确率和召回率的调和均值，即$\mathrm{F1}\ 值=\frac{2\times 精准率\times 召回率}{精准率+召回率}$，相当于精确率和召回率的综合评价指标。

[0033] 本发明的技术效果是：

[0034] (1)本发明使用node2vec算法提取蛋白互作网络中蛋白的特征结构，与传统拓扑性质相比，可以提高蛋白互作网络中蛋白识别的准确率。

[0035] (2)本发明组合使用了node2vec算法和PU-learning算法,降低生物学实验或人工识别疾病相关蛋白的工作量,同时提高精度,尽可能达到自动分类的效果。

附图说明

[0036] 图1为本发明实施例提供的一种基于复杂网络筛选疾病相关蛋白的方法流程示意图。

具体实施方式

[0037] 为了使本领域的技术人员更好地理解本发明的技术方案,下面结合附图和具体实施例对本发明作进一步的详细说明。

[0038] 实施例一:冠心病

[0039] 该实施例展示发明技术效果1,使用node2vec算法提取蛋白互作网络中蛋白的特征结构,与传统拓扑性质相比,可以提高蛋白互作网络中蛋白识别的准确率。在本例S5部分,应用“node2vec算法和PU-learning算法”组合算法,获得958个冠心病有关蛋白,准确率(92.35%)、召回率(56.89%);而使用“拓扑性质算法和PU-learning算法”算法组合,获得8348个与冠心病相关的蛋白,准确率(32.93%)、召回率(10.66%)。

[0040] 如图1所示,本发明实施例提供一种基于复杂网络筛选疾病相关蛋白的方法,该方法包括:

[0041] S1:采用病例-对照研究的方法,对某种疾病组(冠心病)人群和对照组人群进行全基因组扫描,全基因组关联分析(GWAS)获得疾病(冠心病)相关种子基因409个(集合P);

[0042] S2:基于蛋白-蛋白相互作用数据库(Biogrid,版本为3.5.169,网址为https://thebiogrid.org/),构建疾病(冠心病)种子基因为核心的蛋白相互作用网络,该网络由11303个蛋白,44194个蛋白-蛋白相互作用关系组成;

[0043] S3:基于node2vec算法提取S2蛋白相互作用网络中11303个蛋白的特征数据;实施过程中,设置随机游走参数walk_length=80,num_walks=50,window=10,min_count=1,batch_words=4,超参数p=3,q=3,所提取的蛋白特征数据维度为128。

[0044] S4:将S3获得的蛋白结构化特征数据作为输入,利用半监督算法PU-learning算法进行分析,实现蛋白互作网络中与疾病有关的蛋白的预测。具体的,使用Spy技术随机选取15%已知与冠心病有关的蛋白作为“间谍”样本集合S,生成正例集合(PS=P-S)和反例集合(US=U+S),运用朴素贝叶斯算法,得到NB分类器并对反例集合US中的每个蛋白k进行分类,即为每个蛋白赋予一个概率列表标识Pr(1|k),间谍集合S的概率标识决定无标签的蛋白U种哪些最有可能与疾病无关。具体的,通过间谍集合S的概率标识,给定一个阈值H(间谍集合S概率的10分位值),将蛋白的概率标识Pr(1|k)<H的蛋白视为负样(与疾病无关的蛋白)。重复上述步骤100次得到高频负样集合RN(含有蛋白9455个)。在正例样本(409个已知与冠心病有关蛋白)和得到的负样RN(与疾病无关蛋白)集合运行SVM算法,并预测蛋白互作网络中每个蛋白是与疾病有关蛋白的概率,共获得958个可能与冠心病有关的蛋白。

[0045] S5:评价部分:选用文献检索并人工矫正获得与冠心病有关的基因(共997个)为金标准,评价S4获得的958个冠心病有关蛋白的准确率(92.35%)、精准率(54.66%)、召回率(56.89%)、F1值(55.75%)。使用6个拓扑性质(绝对点中心度degree、接近中心度

closenness、点的特征向量中心度evcent、点的中心度betweenness、平均最短路径average distance、Pagerank得分)作为蛋白的特征值,使用PU-learning算法预测得到8348个与冠心病相关的蛋白,其准确率(32.93%)、精准率(89.47%)、召回率(10.66%)、F1值(19.05%)。结果表明,本发明使用node2vec算法提取蛋白互作网络中蛋白的特征结构,与传统拓扑性质相比,可以提高蛋白互作网络中蛋白识别的准确度。

[0046] 实施例二:缺血性心肌病

[0047] 该实施例为了保护方法node2vec的参数范围,维数d,选择维度128-256维,能够得到更加稳定的蛋白互作网络解析结果。而选择维度为64时,得到的稳定负样集合RN结果较少,选择512维时,数据模型在验证集上灵敏度大幅度降低。

[0048] S1:基于GWAS catalog、IPA、DisGeNET等数据库获得缺血性心肌病相关种子基因270个(集合P);

[0049] S2:基于蛋白-蛋白相互作用数据库(BIOGRID、HPRD、INTACT、STRING蛋白互作数据库),构建缺血性心肌病种子基因为核心的蛋白相互作用网络,该网络由9329个蛋白,30274个蛋白-蛋白相互作用关系组成;S1中的270个种子基因在蛋白互作数据库中识别263个,7个蛋白未识别。

[0050] S3:基于node2vec算法提取S2蛋白相互作用网络中9329个蛋白的特征数据;实施过程中,设置随机游走参数walk_length=80,num_walks=50,window=10,min_count=1,batch_words=4,所提取的蛋白特征数据维度d分别为64维、128维、256维、512维。

[0051] S4:将S3获得的蛋白结构化特征数据作为输入,利用半监督算法PU-learning算法进行分析,实现蛋白互作网络中与疾病有关的蛋白的预测。具体的,使用Spy技术随机选取15%已知与缺血性心肌病有关的蛋白作为“间谍”样本集合S,生成正例集合(PS=P-S)和反例集合(US=U+S),运用朴素贝叶斯算法,得到NB分类器并对反例集合US中的每个蛋白k进行分类,即为每个蛋白赋予一个概率列表标识Pr(1|k),间谍集合S的概率标识决定无标签的蛋白U种哪些最有可能与疾病无关。具体的,通过间谍集合S的概率标识,给定一个阈值H(间谍集合S概率的10分位值),将蛋白的概率标识Pr(1|k)<H的蛋白视为负样(与疾病无关的蛋白)。重复上述步骤1000次得到高频负样集合RN。对正例样本和得到的负样RN集合运行SVM算法,并预测蛋白互作网络中每个蛋白是与疾病有关蛋白的概率。结果如下:

[0052]

维度	高频负样集合 RN	正样 P	训练集准确率	验证集准确率	训练集灵敏度	验证集灵敏度
64	**941**	263	0.9979	0.9959	1	1
128	1710	263	0.9981	0.9975	1	1
256	2641	263	1	0.9897	1	0.95455
512	2890	263	1	0.9429	1	**0.6735**

[0053] 选择维度128-256维,能够得到更加稳定的蛋白互作网络解析结果。而选择维度为64时,得到的稳定负样集合RN结果较少,选择512维时,数据模型在验证集上灵敏度降低至0.6735。

[0054] 实施例三:房颤

[0055] 该实施例为了保护方法node2vec的参数范围,p的较优范围为[2,5],q的较优范围为[0.1,3]。

[0056] S1:基于GWAS catalog、Malacards、DisGeNET等数据库获得房颤相关种子基因141个(集合P);

[0057] S2:基于蛋白-蛋白相互作用数据库(BIOGRID、HPRD、INTACT、STRING蛋白互作数据库),构建房颤种子基因为核心的蛋白相互作用网络,该网络由5745个蛋白,13606个蛋白-蛋白相互作用关系组成;S1中的141个种子基因在蛋白互作数据库中识别131个,10个蛋白未识别。

[0058] S3:基于node2vec算法提取S2蛋白相互作用网络中9329个蛋白的特征数据;实施过程中,设置随机游走参数walk_length=80,num_walks=50,window=10,min_count=1,batch_words=4,选取p值分别为0.1、0.2、0.5、1、2、3、4、5,选取q值分别为0.1、0.2、0.5、1、2、3、4、5,提取的蛋白特征数据维度分别为128维。

[0059] S4:将S3获得的蛋白结构化特征数据作为输入,利用半监督算法PU-learning算法进行分析,实现蛋白互作网络中与疾病有关的蛋白的预测。具体的,使用Spy技术随机选取15%已知与房颤有关的蛋白作为"间谍"样本集合S,生成正例集合(PS=P-S)和反例集合(US=U+S),运用朴素贝叶斯算法,得到NB分类器并对反例集合US中的每个蛋白k进行分类,即为每个蛋白赋予一个概率列表标识Pr(1|k),间谍集合S的概率标识决定无标签的蛋白U种哪些最有可能与疾病无关。具体的,通过间谍集合S的概率标识,给定一个阈值H(间谍集合S概率的10分位值),将蛋白的概率标识Pr(1|k)<H的蛋白视为负样(与疾病无关的蛋白)。重复上述步骤100次得到高频负样集合RN。对正例样本和得到的负样RN集合运行SVM算法,并预测蛋白互作网络中每个蛋白是与疾病有关蛋白的概率。结果如下:

[0060]

准确率 q/p	0.1	0.2	0.5	1	2	3	4	5
0.1	0.99855	0.9937	0.99115	0.9939	0.99265	0.99475	0.99375	0.99495
0.2	0.99805	0.9883	0.99805	0.99135	0.9982	0.99205	0.99115	0.9917
0.5	0.9988	0.99875	0.993	0.9928	0.9917	0.9967	0.9967	0.9938
1	0.99765	0.99735	0.9966	0.9942	0.9952	0.9968	0.99675	0.99175
2	0.994	0.99605	0.99275	**0.9994**	0.99725	0.99445	0.9963	0.99455
3	0.9934	0.9961	0.9979	0.9966	0.9961	0.9962	0.9952	0.99335
4	0.99785	0.99925	0.99445	0.99775	0.9934	0.9955	0.9947	0.9968
5	0.9979	0.99745	0.99635	0.9981	0.99695	0.99675	0.99405	0.99405

[0061] 在选择超参数p=2,q=1时,模型最优,准确率最高为0.9994。

[0062] 上述对本发明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,都属于本发明保护范围。

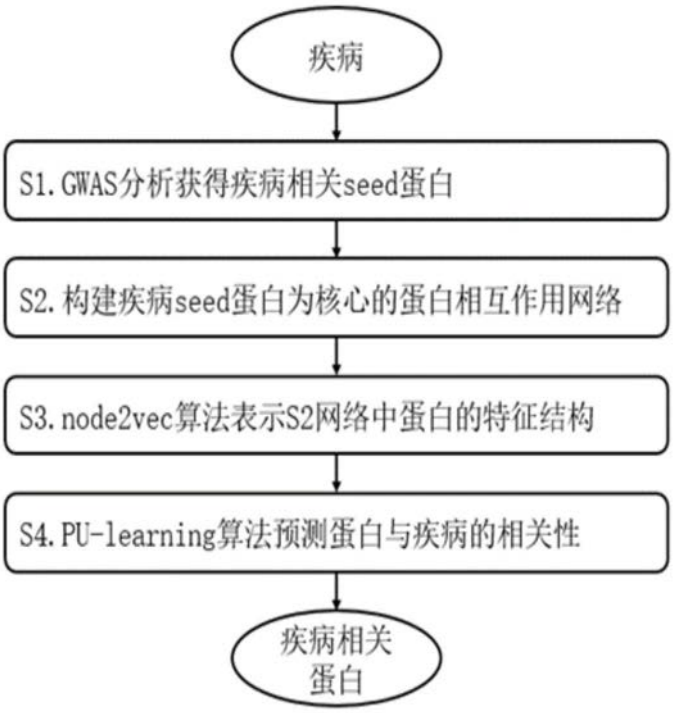

疾病
S1.GWAS分析获得疾病相关seed蛋白
S2.构建疾病seed蛋白为核心的蛋白相互作用网络
S3.node2vec算法表示S2网络中蛋白的特征结构
S4.PU-learning算法预测蛋白与疾病的相关性
疾病相关蛋白

图1