面向健康体检数据的多目标Top-k频繁模式挖掘方法

邱剑锋[1,2], 武梦雨[2], 储建军[3], 张兴义[1,2], 苏延森[1,2†]

(1. 安徽大学 计算智能与信号处理教育部重点实验室，合肥 230039；2. 安徽大学 计算机科学与技术学院，合肥 230039；3. 合肥市第二人民医院，合肥 230012)

摘 要: 为解决现有的模式挖掘方法没有充分利用体检数据中检查项的异常程度与特定疾病之间相关性的问题, 提出一种面向健康体检数据的多目标Top-k频繁模式挖掘方法. 首先, 针对体检数据的特点, 提出异常度和覆盖率两个指标, 在此基础上, 将Top-k频繁模式挖掘建模为一个多目标优化问题; 其次, 针对该问题, 提出一种基于偏好的种群初始化策略和一个面向模式和项的双层更新策略, 并基于此设计一种高效的进化多目标优化算法进行求解. 实验结果表明, 所提出方法所获得的Top-k个模式不仅能够有效地反映其与特定疾病之间的关联性, 而且能够提供多样化的模式, 为健康管理提供重要的参考依据.

关键词: Top-k频繁模式；健康体检数据；多目标进化优化；异常度；覆盖率；基于偏好的初始化策略；双层更新策略

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0657 **开放科学(资源服务)标识码(OSID):**

引用格式: 邱剑锋, 武梦雨, 储建军, 等. 面向健康体检数据的多目标Top-k频繁模式挖掘方法 [J]. 控制与决策, 2023, 38(1): 190-200.

A multi-objective Top-k frequent pattern mining approach oriented for health examination data

QIU Jian-feng[1,2], *WU Meng-yu*[2], *CHU Jian-jun*[3], *ZHANG Xing-yi*[1,2], *SU Yan-sen*[1,2†]

(1. Key Lab of Intelligent Computing and Signal Processing of Ministry of Education，Anhui University, Hefei 230039，China；2. School of Computer Science and Technology，Anhui University，Hefei 230039，China；3. The Second People's Hospital of Hefei，Hefei 230012，China)

Abstract: In order to solve the problem that the existing pattern mining methods do not make full use of the correlation between the abnormality of check items in the health examination data and specific diseases, this paper proposes a multi-objective Top-k frequent pattern mining approach oriented for health examination data. First, according to the characteristics of health examination data, two indicators of abnormality and coverage are proposed, and with these metrics, the Top-k frequent pattern mining is modeled as a multi-objective optimization problem. Then, an efficient evolutionary multi-objective optimization algorithm is designed to solve the problem, in which a preference-based population initialization strategy and a two-layer update strategy oriented to patterns and items are respectively proposed. The experimental results show that the achieved Top-k frequent patterns not only effectively reflect the correlation with the specific diseases, but also provide a variety of patterns, which gives an important reference for health management.

Keywords: Top-k frequent pattern；health examination data；multi-objective evolutionary optimization；abnormality；coverage；preference based initialization strategy；two-level updating strategy

0 引 言

频繁模式挖掘通过识别出现频率较高的项集来揭示隐藏在数据背后的信息, 成为当前医疗数据挖掘领域研究的一个热点问题[1]. 在频繁模式挖掘中, 支持度是衡量其是否为频繁模式的常用指标之一. 当模式的支持度大于给定的最小支持度阈值时, 该项集为频繁模式, 代表性的算法有Apriori[2]、FP-Growth[3]等. 陈治等[4]利用Apriori算法挖掘与宫颈癌发病相

收稿日期: 2021-04-16；**录用日期:** 2021-09-10.

基金项目: 科技部2030新一代人工智能重大项目(2018AAA0100105)；国家自然科学优秀青年基金项目(61822301)；国家自然科学基金项目(61822301, 62076001, U1804262)；安徽省自然科学基金项目(1908085MF218, 2008085QF294)；安徽省重点研发项目(202004j07020005)；安徽高校自然科学研究项目(KJ2019A0029, KJ2021A0048, KJ2021A0634).

†通讯作者. E-mail: suyansen@ahu.edu.cn.

关的因素,通过分析和试验,确定最小支持度阈值,从而发现符合条件的模式集合. Jung等[5]提出了利用频繁模式树挖掘导致慢性病的频繁模式,并利用挖掘结果对慢性病进行预测. Noma等[6]提出了一个五步知识发现模型,利用FP-Growth算法构建FP-tree,以实现对频繁项集的挖掘. 在这些算法中,最小支持度阈值的设置对挖掘结果有着重要的影响;如果设置太小,则会产生大量冗余的候选频繁模式;如果设置太大,尽管有效地压缩了频繁模式的数量,但容易遗漏重要的项集组合[7].

秦琦冰等[8]根据中医数据的特点提出了基于带权无向图的Top-k频繁模式挖掘算法以快速回溯到该频繁模式所对应的方剂名. 胡法奎等[9]提出了一种面向医疗数据的模糊频繁模式挖掘算法,利用模式之间模糊权重因素解决了中医方剂频繁模式挖掘中低频率有效模式的挖掘问题. 此外,Le等[10]针对不确定环境下的频繁模式挖掘问题提出了一个阈值提升策略,以减少不确定环境下候选模式的数量. Zhao等[11]在其已有工作的基础上,通过定义新的指标进一步挖掘出潜在有价值的Top-k频繁模式,实验结果验证了其在肺炎类型分类问题中的有效性. 这些工作针对医疗数据的特点,设计了不同的策略来改进Top-k频繁模式挖掘算法的性能,并且取得了较好的实验结果.

在上述方法中,通过设计不同策略提升了算法性能,但主要还是以支持度作为评价模式质量的指标. 由于问题的复杂性,只考虑单一指标往往使得模式结构相似,缺乏多样性[12]. 对此,一些学者提出通过定义多个度量指标,将频繁模式挖掘建模为多目标优化问题以获得更加全面的挖掘结果. 例如,Zhang等[13]针对频繁模式挖掘提出了效用性这一指标,将模式挖掘问题建模为一个多目标优化问题,并设计了多目标优化算法进行求解. Zhang等[12]引入了覆盖率指标,将Top-k频繁模式挖掘建模为一个多目标优化问题,提出了一个多目标进化算法(ISR-MOEA)对问题进行求解. 这些研究表明,同时考虑模式挖掘问题中的多个因素,利用多目标优化理论对频繁模式挖掘问题建模,并在多目标进化算法框架下对该模型进行求解是一种行之有效的方法.

在体检数据的频繁模式挖掘中,体检项的异常程度往往与疾病之间存在一定的相关性. 直观上看,异常程度较为明显的体检项可能是导致出现病症的关键,而异常程度较弱的往往与病症的关联性偏弱. 从临床的角度看,希望挖掘出的模式不仅能够很好地反映其与病症之间的关联性,而且不同模式之间具有很好的多样性,从而为医生进行健康管理提供不同的参考意见. 为此,本文针对医疗体检数据中的Top-k频繁模式挖掘问题,设计新的度量指标来刻画模式中体检项的异常程度,将Top-k频繁模式挖掘建模为一个多目标优化问题. 在此基础上,提出一个多目标进化算法对该问题进行求解,以发现多样化、关联性强的频繁模式或模式组合. 本文的主要贡献如下:

1) 针对体检数据的特点,本文提出一个新的异常度指标,用于描述检查项的异常值与标准值之间的偏离程度,其大小反映了该项与病症之间关联性的强弱. 进而,考虑模式之间的多样性,结合传统的覆盖率指标,将体检数据中的Top-k频繁模式挖掘问题建模为一个多目标优化问题.

2) 针对上述多目标优化问题,本文提出一个多目标进化优化算法(MOEA-FIMED)进行求解. 在MOEA-FIMED中,分别提出了基于偏好的种群初始化策略以及面向模式和项的双层更新策略以提高初始种群的质量和算法的性能,实现对Top-k频繁模式挖掘多目标优化问题的有效求解.

3) 在体检数据集上的实验结果表明,利用本文所提出的方法挖掘出的Top-k频繁模式能够同时兼顾异常度和覆盖率两个目标,并且通过与已有医学文献的对比,进一步验证了本文所提出方法的有效性.

1　Top-k频繁模式挖掘多目标优化方法

1.1　多目标Top-k频繁模式挖掘模型

在面向体检数据的频繁模式挖掘中,检查项的异常程度往往有助于挖掘出具有指导意义的频繁模式. 为此,本文提出一个新的评价指标,利用项的异常程度来反映模式的异常程度.

定义1 (项的异常度)　给定一个医疗数据库(medical database, MDB)和事务t,假设事务t包括m个不同的项,即$t=\{i_1,i_2,\ldots,i_m\}$. 定义第$j(1\leqslant j\leqslant m)$项的异常度为该项与标准值之间的偏离程度,记为$\mathrm{Diff}_j=\mathrm{norm}(|S_j-i_j|)$. 其中:$S_j$和$i_j$分别表示第$j$项指标的标准值和第$j$项指标的检查结果;$\mathrm{Diff}_j$为两者差的绝对值,并做归一化处理[14]. Diff_j越大,反映了该项对诊断结果(病症)的影响越大.

定义2 (模式异常度)　在定义1的基础上,模式的异常度定义为其中每个项的异常度之和,即

$$\mathrm{Abnormal\_}I=\sum_{j\in I\subseteq t}\mathrm{Diff}_j. \tag{1}$$

该结果反映了模式中项的偏离程度对模式的影响,$\mathrm{Abnormal_}I$越大,说明该模式与诊断的结果之间的关联度越大.

定义3 (k个模式的异常度) 假设包含k个模式的模式集$P=\{I_1,I_2,\dots,I_k\}$. 其中: I_i表示第i个模式, $1\leqslant i\leqslant k$. 其异常度定义为

$$\text{Abnormal}(P)=\frac{1}{k}\sum_{I_i\in P}\text{Abnormal\_}I_i. \tag{2}$$

式(2)描述了所选择的k个模式与诊断结果之间的关联程度, 对于任意两组包含k个模式的集合P_1和P_2, 如果$\text{Abnormal}(P_1)<\text{Abormal}(P_2)$, 则说明模式集$P_2$所选择的$k$个模式与诊断结果之间的关联性更强.

为了使模式集合中所选择的模式之间具有多样性, 本文引入覆盖率指标[12]用于描述k个模式之间的多样性, 定义如下:

$$\text{Coverage}(P)=\frac{|I_1\bigcup\dots\bigcup I_k|}{\sum\limits_{j=1}^{k}|I_j|}. \tag{3}$$

其中: $|I_1\bigcup\dots\bigcup I_k|$表示模式集中所包含的不同项的数目, $\sum\limits_{j=1}^{k}|I_j|$表示$P$中所选择的$k$个模式的长度之和. $\text{Coverage}(P)$值越大, 说明P中的k个模式之间的重复项越少, 该模式具有较好的多样性, 从而为医生提供更加多样化的频繁模式组合.

模式集合的异常度和覆盖率是两个具有相互冲突特征的指标. 具体而言, 在体检数据中往往会有一些与疾病高相关性的项, 因此, 在异常度较高的模式中这些项也会频繁出现. 对于模式集合而言, 提高其异常度即提高集合中各模式的异常度, 而异常度高的模式之间又因上述原因存在重复的项, 从而导致整个模式集合的覆盖率下降. 类似地, 如果要提高模式集合的覆盖率, 则可以删去重复的异常度高的项或者增加未选中的项, 但会降低模式集合的异常度. 如果采用单目标优化问题模型(如仅考虑异常度这一目标), 尽管更易于搜出最优解, 但最终只能发现一个最优的Top-k频繁模式, 并且这k个模式之间相似度较大, 难以发现潜在的多样化的模式组合.

基于上述分析, 本文将Top-k频繁模式挖掘问题建模为一个多目标优化问题, 即发现k个模式的集合, 使其在异常度与覆盖率上达到较好的平衡. 该问题定义如下.

定义4 (Top-k频繁模式挖掘多目标优化问题) $F(P)=(f_1(P),f_2(P))$定义为

$$\begin{cases}\text{minimize}:\ f_1(P)=1-\text{Abnormal}(P),\\ \text{maximize}:\ f_2(P)=\text{Coverage}(P).\end{cases} \tag{4}$$

其中: P为包含k个模式的模式集合, 最小化$f_1(P)$是为了发现异常度最大的k个模式集合, 最大化$f_2(P)$是为了使发现的k个模式集合同时具有尽可能高的覆盖率.

1.2 多目标Top-k频繁模式挖掘算法

1.2.1 算法框架

多目标进化算法能够在相互冲突的目标之间寻找最佳的权衡, 并在一次运行中发现一组非支配解[15-16]. 为此, 本文采用多目标进化算法对多目标Top-k频繁模式挖掘问题进行求解.

由于模式长度不一, 本文采用一种基于变长索引的编码方式对k个模式的集合进行编码[12], 种群中的每个个体代表一种可能的k个模式, 可表示为$P=\langle I_1,\Delta,I_2,\dots,\Delta,I_k\rangle$, I_i表示P中的第i个模式, Δ表示模式之间的分割符. 图1给出了一个包含5个模式的个体表示形式$(I_1,I_2,\dots,I_5)$, 模式之间以-1作为分隔符.

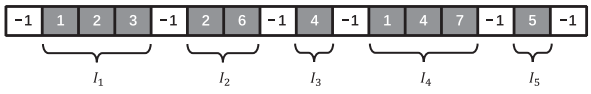


图1 基于变长索引的个体编码

基于所采用的变长索引的编码, 本文以基于分解的多目标进化算法MOEA/D (multi-objective evolutionary algorithm based on decomposition)[17]作为框架, 提出一个多目标进化算法MOEA-FIMED. 下面的算法1给出了整个算法的框架, 主要包括3个阶段: 种群初始化、种群更新和环境选择. 在种群初始化阶段, 提出一个基于偏好的初始化策略(算法1的第1行), 以获得较好的初始种群. 其次, 为了提高算法的收敛性和分布性, 提出一个基于模式和项的双层更新策略用于个体之间的交叉变异操作(算法1的第8行). 下面将对上述两个策略进行详细介绍.

算法1 MOEA-FIMED算法框架.

输入: 医疗数据库MDB, 最大迭代次数maxGen, 种群大小NP, 均匀分布的NP个权重向量$\{\lambda_1,\dots,\lambda_{\text{NP}}\}$, 邻域大小$T$, 变异概率$p_m$, 个体包含的模式数$k$;

输出: 非支配解集EP.

step 1: 基于偏好的种群初始化策略.

1) $\text{Pop}\leftarrow\text{RefInitialPop}(\text{MDB},\text{NP},k)$; // 基于偏好的种群初始化策略(见1.2.2节)

2) $z^*\leftarrow$根据Pop中解的目标值初始化参考点;

3) $N=\{N_1,N_2,\dots,N_{\text{NP}}\}$; // 计算任意两个权重向量之间的欧氏距离, 选取距离每个权向量最近的T个向量所对应的解作为该权向量的邻域, N_i表述第i个个体的邻域解集

4) $\text{iter}=1$.

step 2: 种群更新阶段.

5) while iter $<$ maxGen do

6)　for $i = 1$ to NP do

7)　$p_{i1}, p_{i2} \leftarrow$从当前个体$p_i$的邻域$N_i$中随机选择两个个体作为父代;

8)　$p'_{i1}, p'_{i2} \leftarrow$ BilevelUpdate$(k, p_{i1}, p_{i2}, p_m,$ NP); // 基于模式和项的双层更新策略(见1.2.3节)

9)　$[N_i, z^*] \leftarrow$更新邻域和参考点;

10)　end for

11) end while

step 3: 环境选择阶段.

12) EP $\leftarrow$选择Pop中的非支配解集作为算法的输出.

1.2.2　基于偏好的种群初始化策略

随机初始化往往难以发现潜在的优势区域,降低了算法的性能[18]. 为此,本文提出一种基于偏好的种群初始化策略以产生高质量和多样性的初始种群,主要包括以下3个步骤.

1) 对数据库中的每个项,计算其异常度值,即评价每个长度为1的模式的异常度.

2) 利用轮盘赌机制[19],随机选择k个模式,模式的异常度越大,被选择的概率越大. 因为选择的模式彼此不相同,所以此时覆盖率最高,即偏好覆盖率目标最大.

3) 模式的扩充阶段. 首先产生一个$0 \sim k$之间的随机数,记为num,表示选择前num个模式,对其进行扩充;其次,对于第w个$(w \in [1, \text{num}])$模式,从剩余的候选项集中随机选择一项,记为i_j,加入第w个模式中,计算第w个模式的异常度. 如果异常度增加或者不发生变化,则将第i_j项加入第w个模式,并继续上述过程;否则,将该项i_j从候选项集中删除. 这一操作的目的在于保证覆盖率的情况下,采用贪婪的机制增加异常度,即偏好异常度目标. 算法2给出了基于偏好的种群初始化策略的算法框架.

算法2　RefInitialPop.

输入: 医疗数据库MDB,种群大小NP,个体包含的模式数k;

输出: 初始种群Pop.

1) $(\text{row}, \text{col}) \leftarrow$ sizeof(MDB); // 计算医疗数据库的行数和列数

2) $\{\text{abn}_1, \text{abn}_2, \ldots, \text{abn}_{\text{col}}\} \leftarrow$根据定义1计算数据库中每一项的异常度;

3) for $i = 1$ to NP do

4)　$p_i \leftarrow (I_1, I_2, \ldots, I_k)$; // 根据项的异常度利用轮盘赌机制选择$k$个项生成一个个体,其中包含$k$个模式,每个模式只含有1个项

5)　num $\leftarrow \lceil \text{rand} * k \rceil$; // 产生一个$0 \sim k$之间的随机数num

6)　for $w = 1$ to num do

7)　Selectpool $\leftarrow$将除I_w项之外的$(\text{col} - 1)$项随机放入候选池Selectpool中;

8)　for $j = 1$ to $(\text{col} - 1)$ do

9)　if $\text{Abnormal}_{\{I_w, \text{Selectpool}_j\}} \geqslant \text{Abnormal}_{I_w}$ then

10)　$I_w \leftarrow$将Selectpool_j加入到I_w中;

11)　end if

12)　end for

13)　end for

14) end for

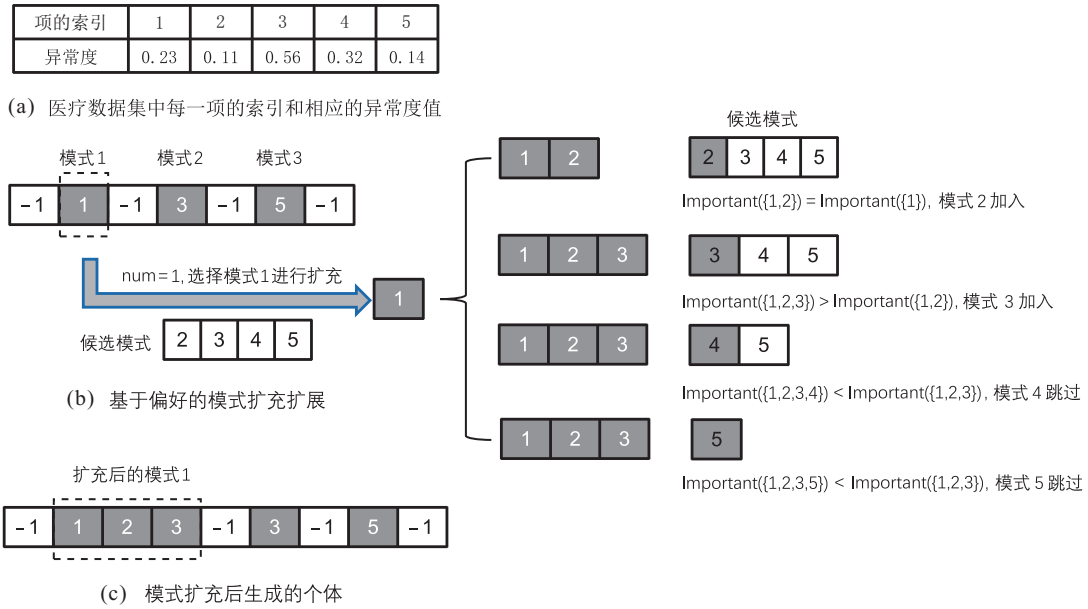


图 2　基于偏好的种群初始化过程

图2给出了一个例子来说明基于偏好的初始化过程. 图2(a)给出了一个包含5个项的数据集和这些项的索引号,同时给出了每个项的异常度值. 在图2(b)中,首先随机选择k个模式(这里k为3,每个模式只包含一个项,其索引号分别为$(1,3,5)$,-1为分隔符). 其次,随机选择前num个模式分别进行扩充,这里$\mathrm{num}=1$,即选择第1个模式(图2(b)中的模式1)进行扩充操作. 对于模式1,在当前项的基础上,依次从候选集2、3、4、5中选择一个新的项加入,由于在添加项4、5时,模式1的异常度下降,这两个项将不被添加进模式1中. 最终形成的个体如图2(c)所示.

1.2.3 基于模式和项的双层更新策略

在进化算法中,传统的交叉变异操作难以解决本文所提出的变长编码方式. 为此,本文提出一种新的基于模式和项的双层更新策略,以更好地在全局探索和局部搜索之间进行平衡,提高算法的性能. 具体而言,第1层是交叉算子,通过引入学习机制发现更好的、异常度更高的模式组合. 第2层是变异算子,在交叉操作的基础上,对每一个个体进行变异操作,其目的在于调整模式中的体检项的出现频率,以获得覆盖率较高的模式组合.

1) 面向模式的基于学习机制的交叉算子.

在MOEA-FIMED中,种群中的个体由k个模式组成,对于任意两个个体,假设为p_i和p_j,存在两种关系:一种是p_i支配p_j (p_j支配p_i是类似的),一种p_i和p_j为非支配的. 前者表明相对于p_j, p_i中包含了更好的模式,可以成为p_j的学习对象. 对于后者,一种随机的方法在两个非支配个体中交换相应的模式,目的在于探索更广泛的区域,发现更有潜力的个体. 主要步骤如下:

① 对于父代种群中的任一个个体p_i,在其大小为T的邻域内选择两个个体,分别记为p_{i1}和p_{i2},利用非支配排序判断两者的支配关系.

② 如果p_{i1}支配p_{i2}(反之类似),则说明p_{i1}中存在更优的模式可以被p_{i2}学习,从而提高p_{i2}的质量,探索新的搜索区域. 具体而言,首先计算p_{i1}和p_{i2}中每个模式的异常度;其次,在p_{i1}和p_{i2}中分别随机选择randIndi个模式作为候选待交叉的模式. 如果p_{i1}中的randIndi个模式,存在其异常度大于p_{i2}中所对应模式的情况,则p_{i2}中的候选模式将被替换,产生新的子代. 通过这种学习机制,利用优势个体中的模式淘汰较劣等个体中的模式可以提高劣等个体的性能.

③ 如果p_{i1}和p_{i2}是非支配关系,则分别从p_{i1}和p_{i2}中随机选择randIndi个模式进行交叉,产生2个新的子代. 这种情况发生在非支配关系的个体中,其目的在于探索新的搜索区域,发现高质量的解.

2) 面向项的基于历史信息的变异算子.

为了进一步提高个体的局部搜索的能力,本文提出一种面向项的基于历史信息的变异算子对个体进行变异操作.主要步骤如下.

① 种群历史信息的采集. 对于当前种群中的所有个体,计算每项在种群中出现的频率,这种频率反映了项对模式的重要程度. 对于某一项而言,如果在不同个体、不同模式中频繁出现,则说明这一项对问题的关联度较强,这样的项将更有可能被选择用于替换变异个体中的项.

② 个体变异位置的选择. 对于给定的个体,计算其中的每一项在个体中出现的频率,频率高者,说明个体中的模式大多选择了该项. 为了提高个体覆盖率,在计算项出现的频率的基础上,采用轮盘赌机制来确定变异的位置,如果这一项出现的次数越高,则越有可能作为变异的位置.

③ 个体中项的改变. 在确定变异位置后,根据种群历史信息,利用轮盘赌机制选择关联度较强的项替换当前的项,从而产生新的变异个体.

基于模式和项的双层更新策略产生了新的个体. 其中:面向模式的基于学习机制的交叉算子是通过模式之间的相互学习,引导种群在更大的范围内去探索未知区域;对于面向项的基于历史信息的变异算子,利用进化过程中的历史信息对模式中的项进行变异,以产生新的模式,发现更优的个体. 算法3给出了基于模式和项的双层更新过程.

算法3 BilvelUpdate.

输入: 个体包含的模式数k,个体p_{i1},个体p_{i2},变异概率p_m,种群大小NP;

输出: 子代个体p'_{i1}, p'_{i2}.

// 面向模式的基于学习机制的交叉算子

1) randIndi $\leftarrow \lceil \mathrm{rand} * k \rceil$; // 产生一个$0 \sim k$之间的随机数randIndi

2) 分别从p_{i1}和p_{i2}中随机选取randIndi个模式,计算每个模式的异常度,记为$\mathrm{Abnormal}_{p_{i1}}$和$\mathrm{Abnormal}_{p_{i2}}$;

3) if is_nonDominant(p_{i1}, p_{i2}) then

4) for $u=1$ to randIndi do

5) if $\mathrm{Abnormal}_{p_{i1}}(u) > \mathrm{Abnormal}_{p_{i1}}(u)$ then

6) 将p_{i1}中的第u个模式与p_{i2}中所对应的模式交换,得到子代p'_{i1}, p'_{i2};

7) end if

8)　end for

9) else

10)　交换从p_{i1}和p_{i2}中选取的randIndi个模式, 得到子代p'_{i1}, p'_{i2};

11) end if

// 面向项的基于历史信息的变异算子

12) for $k = 1$ to 2 do

13)　if rand $\leqslant p_m$ then

14)　计算p'_{ik}中不同项出现的频率, 利用轮盘赌机制选择被替换项;

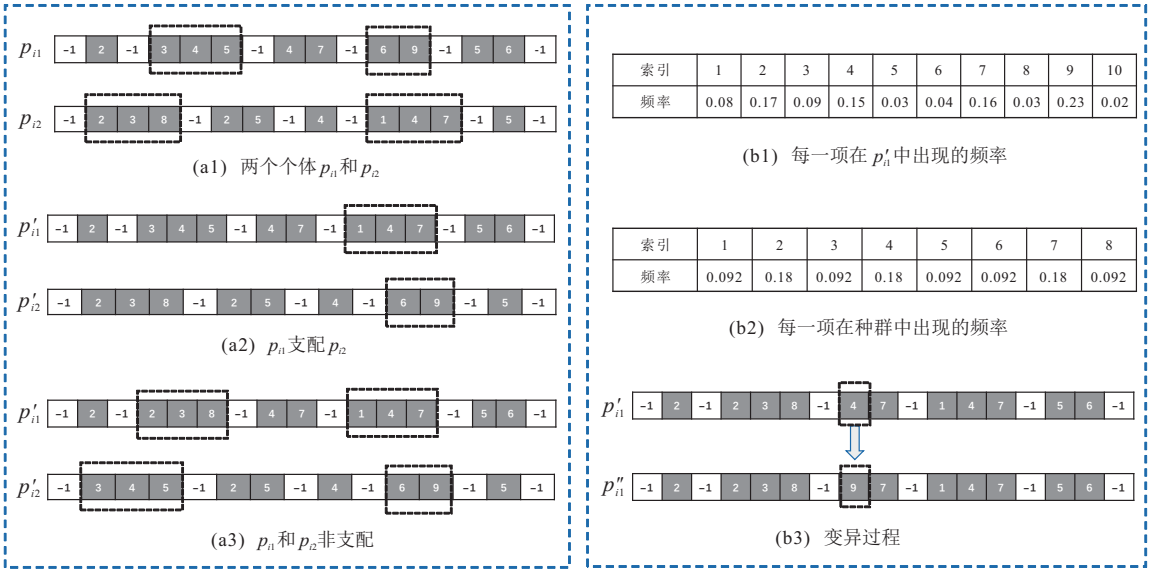
15)　计算种群中不同项出现的频率, 利用轮盘赌机制选择替换项;

16)　使用替换项代替被替换项, 得到更新后的子代p'_{ik};

17　end if

18) end for

图 3 给出一个例子来解释本文所提出的双层更新策略. 图 3(a) 展示的是利用面向模式的基于学习机制的交叉算子产生子代的过程. p_{i1} 和 p_{i2} 是从邻域 N_i 中随机选取的两个个体, $\{3,4,5\}$、$\{6,9\}$ 和 $\{2,3,8\}$、$\{1,4,7\}$ 分别是在 p_{i1} 和 p_{i2} 中选择的 2 个模式 (图 3(a1) 所示). 如果 p_{i1} 支配 p_{i2}, 则 p_{i2} 向 p_{i1} 学习. 由于模式 $\{6,9\}$ 的异常度大于模式 $\{1,4,7\}$ 的异常度, p_{i1} 中的 $\{6,9\}$ 与 p_{i2} 中的 $\{1,4,7\}$ 发生了交换 (图 3(a2) 所示). 如果 p_{i1} 和 p_{i2} 是非支配关系, 则上述两个个体中随机选择的模式将发生交换, 以发现更好的解 (如图 3(a3) 所示). 图 3(b) 给出了面向项的基于历史信息的变异过程. 图 3(b1) 和图 3(b2) 分别给出了在待突变个体 p'_{i1} 中出现的项的频率和整个种群中项出现的频率, 即种群的历史信息. 根据上述信息, 利用轮盘赌机制选择第 3 个模式中的第 1 位进行突变, 如图 3(b3) 所示.



(a) 面向模式的基于学习机制的交叉算子　　(b) 面向项的基于历史信息的变异算子

图 3　基于模式和项的双层更新策略

2　实验与分析

本节在一个公开的心脏病数据集和从某医疗卫生机构采集的近 5 年的健康体检数据集上验证本文所提出算法的性能, 并将挖掘出的 Top-k 频繁模式与现有的医学文献结论进行对比, 验证其临床上的指导意义.

2.1　数据集及数据预处理方法

心脏病数据集可从公开的数据库 UCI machine learning repository[20] 中下载, 脂肪肝数据集来自某社区医院获得的近 5 年的医疗体检数据, 并提取标记有脂肪肝的样本. 在这些样本中, 每一个检查项均为数值型, 并且给出了明确的上下界取值. 表 1 给出了每一个数据集的名称、受检者数目及所包含的项的数目. 本文采用下面的预处理方法对数据进行处理.

表 1　心脏病和脂肪肝数据集

序号	数据集	受检者数目	项的数目
1	心脏病	120	23
2	脂肪肝	1 828	68

首先将正常指标值设为 0, 其次计算异常值与给定的标准值的差值以表示异常的程度. 这个差值越大, 代表该异常对患者的影响越大, 越有可能与相应的疾病产生关联 (差值为 0 意味着该受检者在对应的指标上的检查结果是正常的). 不同的指标, 其在数值的量纲上可能是不同的, 为此, 本文采用 L_2 范数[14] 对

差值进行归一化操作.

2.2 实验设置

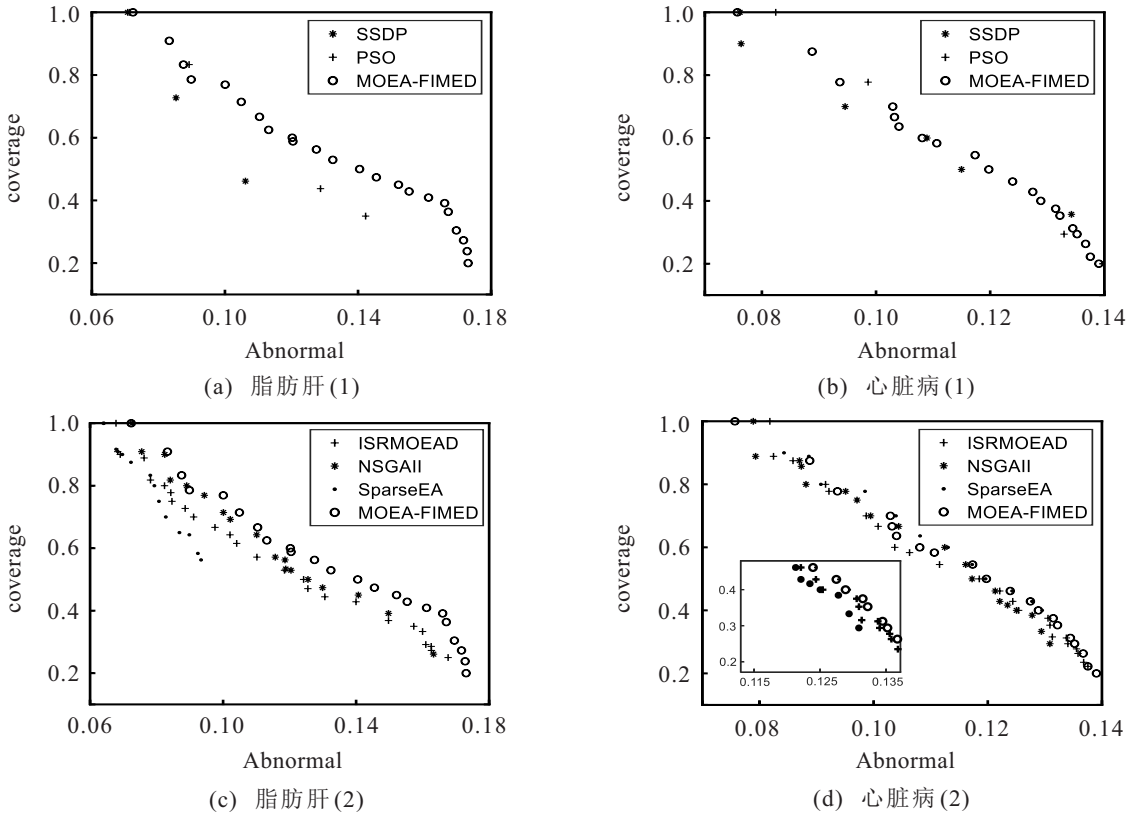
为了验证MOEA-FIMED的有效性,本文选取5种对比算法,即PSO[21]、SSDP[22]、ISR-MOEA[12]、SparseEA[23]、NSGA-II[24]. PSO和SSDP是单目标进化算法,用于挖掘Top-k频繁模式. 其中:PSO是一种高效的粒子群优化算法,在很多实际优化问题中得到了广泛应用[25-26],其基本思想是通过种群中个体之间的协作和信息共享来寻找最优解;SSDP是一种单目标优化算法,通过利用归属特征来挖掘Top-k频繁模式. ISR-MOEA、SparseEA和NSGA-II是3个多目标进化算法用于求解Top-k频繁模式挖掘问题,其中,ISR-MOEA是一种基于索引集的多目标进化方法,通过引入覆盖率度量模式集合的多样性,并给出了一个Top-k频繁模式挖掘多目标优化算法;针对模式挖掘中最优解往往具有稀疏性的特点,SparseEA设计了一种通用的求解大规模稀疏多目标优化问题的进化算法,用于解决模式挖掘问题;NSGA-II是一种广泛使用的基于支配关系的多目标进化算法.

因为单目标进化算法每次运行只能得到一个解,所以,本文采用文献[12]的处理方法,即对单目标进化算法进行r次实验,其中r为利用MOEA-FIMED所得到的前沿面中解的个数. 同时,在每次运行时,对两个目标分配不同的权重α和β,并进行加权求和. 其中:$\beta = 1-\alpha, \alpha$在$[0,1]$上均匀采样. 对于多目标进化算法,本文采用其各自文献中推荐的参数. 对于MOEA-FIMED,本文设邻域大小为10,变异概率为0.25. ISR-MOEA也是在MOEA/D框架下的进化多目标的Top-k频繁模式挖掘算法,使用与MOEA-FIMED相同的邻域大小和子问题数目. 在所有的实验中,种群大小为100,最大迭代次数为30,每个个体包含5个模式,即Top-k中的k取5. 所有算法在每个数据集上独立运行30次,记录下每次的Pareto前沿面中的非支配解. 所有实验在Intel i5-7500 @3.4 GHz四核处理器、16 GB主存的机器上采用Matlab R2017a编程语言实现.

2.3 实验结果

2.3.1 MOEA-FIMED与PSO、SSDP、ISR-MOEA、SparseEA、NSGA-II算法之间的比较

图4给出了MOEA-FIMED和5种对比算法在目标空间的帕累托前沿面的分布情况. 其中,图4(a)和(b)给出了MOEA-FIMED与两种单目标算法对比的结果,可以看出,MOEA-FIMED在两个数据集上展示了较好的分布性,在脂肪肝数据集上,其收敛性的优势非常明显并具有良好的分布性. 图4(c)和(d)给出了MOEA-FIMED与3种多目标算法对比的结果,可以看到,MOEA-FIMED获得了更好的收敛性和多样性. SparseEA是一种新的用于解决频繁模式挖掘



(a) 脂肪肝(1) (b) 心脏病(1)

(c) 脂肪肝(2) (d) 心脏病(2)

图4 MOEA-FIMED和5种对比算法在两个数据集上的帕累托前沿面分布

问题的稀疏多目标优化算法,从脂肪肝数据集上的实验结果来看,MOEA-FIMED所获得的解在目标空间的收敛性和分布性均明显优于SparseEA,意味着MOEA-FIMED能够提供更加多样性的频繁模式组合. 类似地,在心脏病数据集上,MOEA-FIMED也展示了较好的实验结果.

表2给出了6种算法所获得的非支配解在目标空间的平均HV(hypervolume)值[27],HV值越大,表明所获得的非支配解具有越好的收敛性和多样性. 实验采用0.05显著性水平下的Wilcoxon秩和检验[28]来评价对比算法与MOEA-FIMED的差异性. 实验结果还表明,MOEA-FIMED在统计上是显著优于其他5种基于进化算法的Top-k频繁模式挖掘算法.

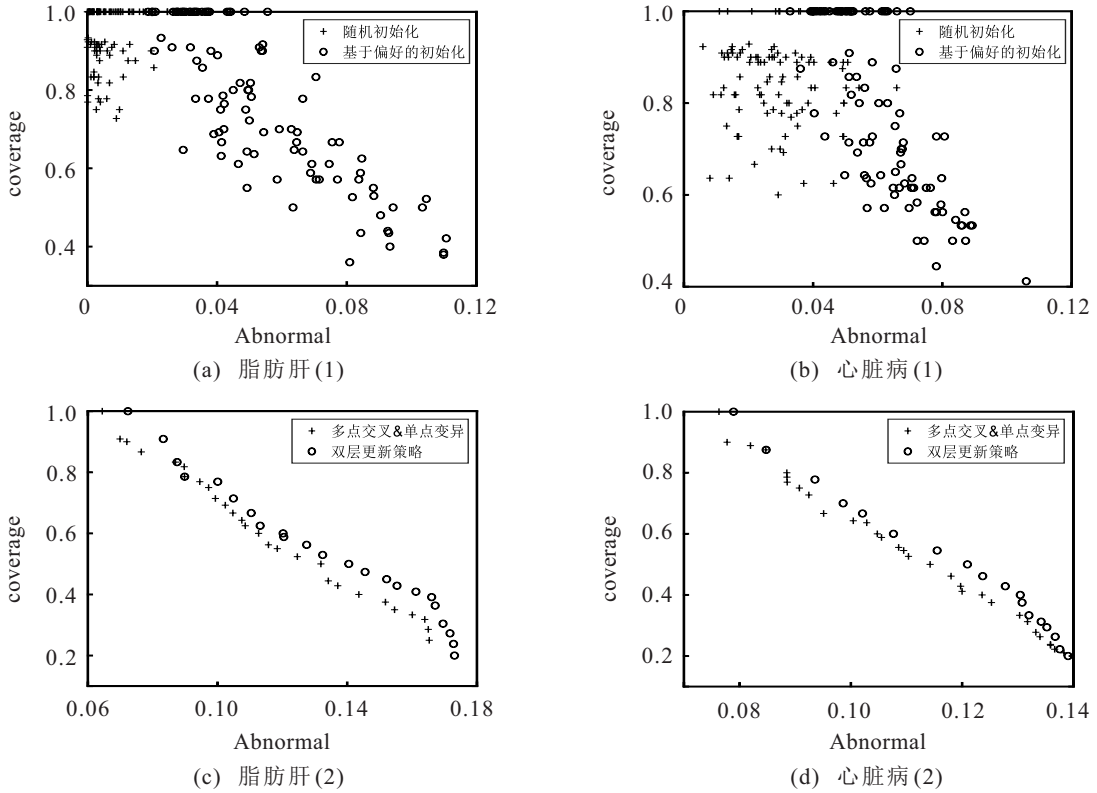
表 2　6种进化算法在两个数据集上的平均HV值

数据集	SSDP	PSO	ISR-MOEA	SparseEA	NSGA-II	MOEA-FIMED
脂肪肝	0.099 0±0.004 3-	0.104 4±0.003 6-	0.123 1±0.001 3-	0.097 0±0.007 0-	0.120 6±0.002 5-	0.130 5±0.001 0
心脏病	0.108 0±0.002 7-	0.104 0±0.002 0-	0.112 6±0.000 6-	0.111 5±0.001 0-	0.1114±0.001 0-	0.113 2±0.000 7

2.3.2　策略有效性和参数敏感性分析

本文在MOEA/D框架下,提出一个基于偏好的初始化策略和面向模式和项的双层更新策略以提高算法的性能. 本节首先验证所提出策略的有效性;其次,对变异概率这一重要参数的选择进行敏感性分析以确定最优的参数值.

图5给出了本文所提出的两个策略的有效性验证. 从图5(a)和(b)中可以看出,采用随机初始化策略得到的初始个体往往聚集在目标空间的某一区域,分布性较差. 利用本文所提出的初始化策略考虑了对不同目标的偏好,使得最终的个体在目标空间上有较好地分布. 图5(c)和(d)给出了仅利用传统的更新策略,即多点交叉[29]和单点变异算子[30],和利用本文所提出的更新策略的对比结果. 可以看出,通过引入学习机制和历史信息的双层更新策略在目标空间获得了更好的非支配前沿.



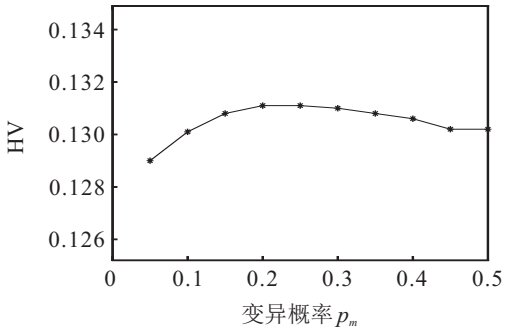
(a) 脂肪肝(1)　(b) 心脏病(1)

(c) 脂肪肝(2)　(d) 心脏病(2)

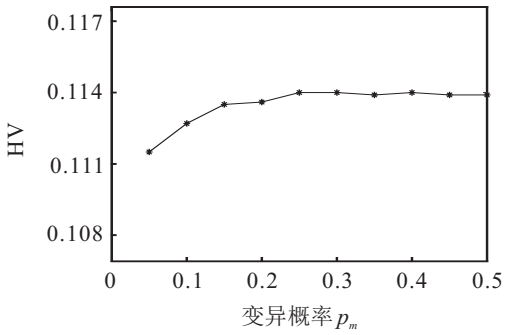
图 5　基于偏好的初始化策略有效性验证

在MOEA-FIMED中,变异概率p_m是一个重要的参数. 为了选择合适的p_m,本文以0.05为间隔,在0～0.5之间测试p_m在不同取值下算法的性能. 图6给出了在不同p_m下算法得到的HV值. 可以看出:对于脂肪肝数据集,当p_m取0.25时,算法获得了最好的HV值;而对于心脏病数据集,当p_m在0.25时,算法获

得了较好的HV值. 综合考虑,在本文中,参数变异概率p_m取0.25.



(a) 脂肪肝



(b) 心脏病

图 6 变异概率p_m的参数敏感性分析

2.3.3 异常度指标的有效性分析

本文提出一个新的异常度指标用于描述体检项与病症之间相关性的强弱. 为了验证所提出的异常度指标的有效性,将MOEA-FIMED算法所得到的前沿面中的Pareto解抽取出来,与Aprior算法的结果进行对比,以验证在不需要人为设置支持度的情况下,同样可以获得支持度较高的频繁模式,并且在这些模式之间表现出较好的多样性.

以脂肪肝数据集为例,从MOEA-FIMED算法获得的前沿面中提取48个模式,其中模式的最大长度为6. 因此,将算法Apriori的模式最大长度设为6,最小支持度设为0,总共863 834个模式. 表3给出了对前沿面中所有模式按异常度排序后取前5位的模式,并计算了模式的支持度. 从表3中可以看出,引入异常度指标后,在不需要人工设置支持度阈值的前提下,所挖掘出的模式中异常度较高的,支持度也比较高,同时也表现出较好的多样性.

2.3.4 与现有医学文献结果的对比分析

仍以脂肪肝数据为例,通过查阅相关的文献来验证MOEA-FIMED所挖掘出的模式中蕴含的项与实际临床诊断结果的相关性. 本文统计了利用MOEA-FIMED算法获得的前沿面中Pareto解,及其中每一项出现的频次,按降序排列取前5项如下: 男性、红细胞压积异常、血红蛋白异常、高血压、红细胞计数异常. 表4给出的这些项以及相关文献给出的佐证表明,利用MOEA-FIMED算法挖掘出的这些项与脂肪肝高度相关.

表 3 MOEA-FIMED挖掘出的按异常度排序前5位的模式及其支持度

序号	模式	支持度
1	男, 红细胞压积异常, 血红蛋白异常, 高血压	0.173 077
2	男, 红细胞计数异常, 红细胞压积异常, 血红蛋白异常, 高血压	0.171 606
3	男, 红细胞压积异常, 血红蛋白异常, 体重指数超重, 高血压	0.167 634
4	男, 红细胞压积异常, 胃幽门螺杆菌抗体异常, 血红蛋白异常, 高血压	0.162 341
5	男, 红细胞计数异常, 红细胞压积异常, 血红蛋白异常	0.158 341

表 4 与现有医学文献依据的对比分析结果

序号	与脂肪肝异常相关的项集	出现的文献
1	男	[31-32]
2	红细胞压积异常	[33-34]
3	血红蛋白异常	[35-36]
4	高血压	[37, 32]
5	红细胞计数异常	[31, 35]

3 结 论

针对健康体检数据中的Top-k频繁模式挖掘问题,本文提出了一种基于多目标进化优化的Top-k频繁模式挖掘方法. 首先针对健康体检数据的特点,提出了一个模式异常度的指标,并结合传统的覆盖率指标,将Top-k频繁模式挖掘问题建模为一个多目标优化问题. 为了提高算法的性能和所挖掘出的模式的质量,提出一个基于偏好的种群初始化策略和一个面向模式和项的双层更新策略,对该多目标优化问题进行求解. 实验结果表明,本文所提出的算法能够发现一组在异常度和多样性方面具有较好表现的Top-k频繁模式,此外,已有的医学文献也验证了所发现的k个频繁模式的有效性. 本文所提出的方法可以帮助医生发现更多潜在的、可能会导致该疾病的体检指标(组合),可根据这些指标有针对性地安排体检事宜. 对体检者而言,可以更加关注这些与疾病相关的体检指标,一旦这些指标出现异常情况,可提前对相

应的疾病进行干预和治疗.

本文所提出的算法表明,通过设计新的度量指标,利用多目标优化理论进行医疗健康数据挖掘是一种行之有效的思路. 然而,在健康体检数据中仍然存在着大量非数值型的体检数据,如何针对这些非数值型体检数据特点,设计更加有效的度量指标,构建复杂的高维多目标优化模型,并设计相关的求解算法将是一个非常有意义的工作.

参考文献(References)

[1] Benatia M A, Baudry D, Louis A. Detecting counterfeit products by means of frequent pattern mining[J]. Journal of Ambient Intelligence and Humanized Computing, DOI:10.1007/s12652-020-02237-y.

[2] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]. Proceedings of the 20th International Conference on Very Large Data Bases. Santiago, 1994: 487-499.

[3] Han J W, Pei J, Yin Y W. Mining frequent patterns without candidate generation[J]. ACM SIGMOD Record, 2000, 29(2): 1-12.

[4] 陈治, 吴娟娟. 基于关联规则的医疗数据挖掘研究[J]. 统计与决策, 2020, 36(6): 174-177.
(Chen Z, Wu J J. Research on medical data mining based on association rules[J]. Statistics & Decision, 2020, 36(6): 174-177.)

[5] Jung H, Chung K Y, Lee Y H. Decision supporting method for chronic disease patients based on mining frequent pattern tree[J]. Multimedia Tools and Applications, 2015, 74(20): 8979-8991.

[6] Noma N G, Abd Ghani M K. Discovering pattern in medical audiology data with FP-growth algorithm[C]. IEEE-EMBS Conference on Biomedical Engineering and Sciences. Langkawi, 2012: 17-22.

[7] 武优西, 王振坤, 史巧硕, 等. 无重叠条件下的Top-k序列挖掘[J]. 小型微型计算机系统, 2019, 40(10): 2170-2174.
(Wu Y X, Wang Z K, Shi Q S, et al. Top-k sequence mining with nonoverlapping condition[J]. Journal of Chinese Computer Systems, 2019, 40(10): 2170-2174.)

[8] 秦琦冰, 谭龙. 基于中医方剂数据库的Top-Rank-k频繁模式挖掘算法[J]. 计算机应用, 2017, 37(2): 329-334.
(Qin Q B, Tan L. Top-Rank-k frequent patterns mining algorithm based on TCM prescription database[J]. Journal of Computer Applications, 2017, 37(2): 329-334.)

[9] 胡法奎, 陈高云, 龚程, 等. 面向大规模医疗数据的模糊频繁模式挖掘研究[J]. 信息通信, 2017, 30(3): 14-16.
(Hu F K, Chen G Y, Gong C, et al. Research on fuzzy frequent pattern mining for large-scale medical data[J]. Information & Communications, 2017, 30(3): 14-16.)

[10] Le T, Vo B, Huynh V N, et al. Mining top-k frequent patterns from uncertain databases[J]. Applied Intelligence, 2020, 50(5): 1487-1497.

[11] Zhao Y H, Yin Y, Wang G R. Identifying top-k vital patterns from multi-class medical data[C]. International Conference on Future BioMedical Information Engineering (FBIE). Sanya, 2009: 536-539.

[12] Zhang L, Yang S S, Wu X P, et al. An indexed set representation based multi-objective evolutionary approach for mining diversified top-k high utility patterns[J]. Engineering Applications of Artificial Intelligence, 2019, 77: 9-20.

[13] Zhang L, Fu G L, Cheng F, et al. A multi-objective evolutionary approach for mining frequent and high utility itemsets[J]. Applied Soft Computing, 2018, 62: 974-986.

[14] Dai Z, Chen W N, Huang X H, et al. CNN descriptor improvement based on L_2-normalization and feature pooling for patch classification[C]. IEEE International Conference on Robotics and Biomimetics. Kuala Lumpur, 2018: 144-149.

[15] 耿焕同, 周山胜, 陈哲, 等. 基于分解的预测型动态多目标粒子群优化算法[J]. 控制与决策, 2019, 34(6): 1307-1318.
(Geng H T, Zhou S S, Chen Z, et al. Decomposition-based predictive dynamic multi-objective particle swarm optimization algorithm[J]. Control and Decision, 2019, 34(6): 1307-1318.)

[16] 么双双, 董志明, 王显鹏. 基于分解的多目标多因子进化算法[J]. 控制与决策, 2021, 36(3): 637-644.
(Yao S S, Dong Z M, Wang X P. A multiobjective multifactorial evolutionary algorithm based on decomposition[J]. Control and Decision, 2021, 36(3): 637-644.)

[17] Zhang Q F, Li H. MOEA/D: A multiobjective evolutionary algorithm based on decomposition[J]. IEEE Transactions on Evolutionary Computation, 2007, 11(6): 712-731.

[18] Rocco C M, Ramirez-Marquez J E, Salazar D E, et al. Assessing the vulnerability of a power system through a multiple objective contingency screening approach[J]. IEEE Transactions on Reliability, 2011, 60(2): 394-403.

[19] Lipowski A, Lipowska D. Roulette-wheel selection via stochastic acceptance[J]. Physica A: Statistical Mechanics and Its Applications, 2012, 391(6): 2193-2196.

[20] Dua D, Graff C. UCI machine learning repository[EB/OL]. [2021-04-01]. http://archive.ics.uci.edu/ml.

[21] Liu B, Wang L, Jin Y H. An effective PSO-based memetic algorithm for flow shop scheduling[J]. IEEE Transactions on Systems, Man, and Cybernetics: Part B, 2007, 37(1): 18-27.

[22] Lucas T, Silva T C P B, Vimieiro R, et al. A new evolutionary algorithm for mining top-k discriminative patterns in high dimensional data[J]. Applied Soft Computing, 2017, 59: 487-499.

[23] Tian Y, Zhang X Y, Wang C, et al. An evolutionary algorithm for large-scale sparse multiobjective optimization problems[J]. IEEE Transactions on Evolutionary Computation, 2020, 24(2): 380-393.

[24] Deb K, Pratap A, Agarwal S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182-197.

[25] 胡广浩, 毛志忠, 何大阔. 基于两阶段领导的多目标粒子群优化算法[J]. 控制与决策, 2010, 25(3): 404-410.
(Hu G H, Mao Z Z, He D K. Multi-objective PSO optimization algorithm based on two stages-guided[J]. Control and Decision, 2010, 25(3): 404-410.)

[26] Bonavolontà F, Noia L P D, Liccardo A, et al. A PSO-MMA method for the parameters estimation of interarea oscillations in electrical grids[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(11): 8853-8865.

[27] Zitzler E, Thiele L. Multiobjective optimization using evolutionary algorithms — A comparative case study[C]. International Conference on Parallel Problem Solving from Nature. Berlin: Springer, 1998: 292-301.

[28] Wilcoxon F, Katti S K, Wilcox R A. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test[J]. Selected Tables in Mathematical Statistics, 1970, 1: 171-259.

[29] Spears W M, de Jong K A. An analysis of multi-point crossover[C]. Foundations of Genetic Algorithms. Amsterdam: Elsevier, 1991: 301-315.

[30] Ji M J, Tang H W, Guo J. A single-point mutation evolutionary programming[J]. Information Processing Letters, 2004, 90(6): 293-299.

[31] Wang H L, Zhang H, Wu S L, et al. Red blood cell count has an independent contribution to the prediction of ultrasonography-diagnosed fatty liver disease[J]. PLoS One, 2017, 12(2): e0172027.

[32] Firneisz G. Non-alcoholic fatty liver disease and type 2 diabetes mellitus: The liver disease of our age? [J]. World Journal of Gastroenterology, 2014, 20(27): 9072-9089.

[33] Li G L, Hu H, Shi W, et al. Elevated hematocrit in nonalcoholic fatty liver disease: A potential cause for the increased risk of cardiovascular disease? [J]. Clinical Hemorheology and Microcirculation, 2012, 51(1): 59-68.

[34] Li Y, Liu L, Wang B, et al. Hematocrit is associated with fibrosis in patients with nonalcoholic steatohepatitis[J]. European Journal of Gastroenterology & Hepatology, 2014, 26(3): 332-338.

[35] Giorgio V, Mosca A, Alterio A, et al. Elevated hemoglobin level is associated with advanced fibrosis in pediatric nonalcoholic fatty liver disease[J]. Journal of Pediatric Gastroenterology and Nutrition, 2017, 65(2): 150-155.

[36] Li H B, Guo M H, An Z, et al. Prevalence and risk factors of metabolic associated fatty liver disease in Xinxiang, China[J]. International Journal of Environmental Research and Public Health, 2020, 17(6): 1818.

[37] Liu P Y, Tang Y H, Guo X P, et al. Bidirectional association between nonalcoholic fatty liver disease and hypertension from the Dongfeng-Tongji cohort study[J]. Journal of the American Society of Hypertension: JASH, 2018, 12(9): 660-670.

作者简介

邱剑锋 (1979–), 男, 副教授, 博士, 从事机器学习及多目标进化优化的研究, E-mail: qiujianf@ahu.edu.cn;

武梦雨 (1996–), 女, 硕士生, 从事频繁模式挖掘及多目标优化的研究, E-mail: 412265778@qq.com;

储建军 (1976–), 男, 教授, 硕士, 从事医疗信息数据挖掘等研究, E-mail: chujianjun@163.com;

张兴义 (1982–), 男, 教授, 博士生导师, 从事多目标进化优化理论等研究, E-mail: xyzhanghust@gmail.com;

苏延森 (1985–), 女, 副教授, 博士生导师, 从事医疗数据挖掘、进化优化等研究, E-mail: suyansen@ahu.edu.cn.

(责任编辑：李君玲)