计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500082

基于相对危险度的儿童先心病风险因素分析算法

徐慧慧　晏　华
电子科技大学计算机科学与工程学院　成都 611731
(494655043@qq.com)

摘　要　对疾病相关风险项的分析是数据挖掘理论在医疗领域应用的一个重要内容,可以帮助医生分析疾病成因,从而有效地开展防治工作。医学领域的疾病数据有其自身的特征,例如其高度不平衡性的特点往往使得大量珍贵的信息蕴藏于支持度小的属性项中,直接采用经典的基于支持度的关联规则挖掘算法易造成重要信息的丢失。因此,文中结合医疗领域的知识,基于医学领域常用的统计标准——相对危险度,提出了一种挖掘疾病高风险项集的算法(Mining Algorithm for high Relative Risk Itemsets,MARRI),以及与之相匹配的两种规则剪枝方法,即作用叠加剪枝和样本数剪枝,并在儿童先心病数据集上对算法进行验证。实验结果表明,该算法具有挖掘低支持度项集信息的能力,挖掘出的疾病关联因素更有价值。

关键词:关联规则;相对危险度;数据挖掘;疾病分析
中图法分类号　TP181

Relative Risk Degree Based Risk Factor Analysis Algorithm for Congenital Heart Disease in Children

XU Hui-hui and YAN Hua
School of Computer Science & Engineering,University of Electronic Science and Technology of China,Chengdu 611731,China

Abstract　The analysis of disease-related risk factors is an important part of application of data mining theory in the medical field,which is helpful for doctors to analyze causes of disease and carry out effective work of disease prevention and control. But disease data in the medical field have their own characteristics,such as high imbalance,which means that most valuable information is contained in the attribute items with a small support. It is easy to lose important information when applying the classical association rule algorithm based on the support directly. Therefore,based on the knowledge of medical field and the common statistical standard of medical field——Relative Risk,this paper proposes a mining algorithm for high relative risk itemsets(MARRI) and two corresponding pruning methods,which are interaction pruning and sample number pruning,and verifies the algorithm on the dataset of children's congenital heart disease. Experimental results show that the algorithm is effective to mine the information in low support items and disease-related factors mined out are more valuable.

Keywords　Association rules,Relative risk,Data mining,Disease analysis

1　引言

随着数据挖掘技术的发展和成熟,人们不断地探索其在各个领域的应用。由于应用领域数据的实际情况存在差异,经典算法在应用中往往需要根据数据的特定分布进行改进。本文研究关联规则挖掘[1]在医疗疾病方面的应用,针对疾病数据的不平衡性问题,基于医学统计指标相对危险度,提出一种挖掘疾病高风险项集的算法(MARRI)。

在医疗领域中,关联规则可以用于总结用药规律,帮助指导治疗,例如,文献[2]研究了肺系疾病症状与对应治疗药物之间的关联规则,文献[3]使用关联规则算法对陈可冀院士多年治疗心血管疾病的医案进行了用药规律总结。首先,关联规则可用于发现病因,进行健康预警,如文献[4]以胃癌和心肌炎数据为例,利用关联规则发现易患病人群的特征。其次,关联规则也可用于发现疾病之间的关联,如文献[5]从病患记录中挖掘出了疾病之间的伴生关系。除直接利用经典算法以外,对关联规则算法进行改进使其适用于更多样的任务也是一大研究重点,如文献[6]融合了关联规则和 logistic 回归模型建立了疾病预警系统,文献[7]将 11 种度量(Jaccard,cosine等)划分为 4 类来代替支持度并利用层次聚类法来选择所采用的度量,文献[8]利用遗传算法和数据采样来优化关联规则算法。近年来,随着大数据技术和人工智能的快速发展,关联

到稿日期:2020-05-19　返修日期:2020-08-08　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61976046);四川省重点研发计划项目(2018SZ0065)
This work was supported by the National Natural Science Foundation of China(61976046) and Key Research and Development Projects of Sichuan Province(2018SZ0065).
通信作者:晏华(huayan@uestc.edu.cn)

规则与这些新技术的结合也越发紧密，文献[9-13]研究了关联规则在分布式领域和大数据领域中的挖掘，文献[14]将关联规则与深度神经网络相结合来预测抗癌药物的效用。

对疾病相关风险项的分析是医疗领域数据挖掘的一个重点内容，其有助于分析疾病的成因和有效开展疾病的预防工作，但疾病相关数据具有极度不平衡性[15-16]的特点，具体表现为异常项的比重远小于正常记录的比重。此处的异常项不仅包括是否患病，还包括吸烟、喝酒等生活习惯异常，射线、化学药剂等接触环境异常，以及肥胖、消瘦等身体指标不在正常区间的异常。当研究疾病相关风险项时，数据集中的异常项往往更值得被关注，其所蕴含的信息密度也更大[17]。但传统的关联规则算法挖掘频繁项集的首要依据就是支持度，即项集在数据集中出现的频率[1]。在面对疾病数据这类不平衡性数据时，单一支持度的设定将直接过滤掉更具挖掘价值的异常项。若想通过降低支持度阈值来挖掘异常项，则又会造成结果过度冗余的问题[18]。针对这一问题，本文利用医疗领域常用的统计标准——相对危险度，提出了高风险项集的概念和挖掘算法，以进行疾病风险因素的分析。

2 相对危险度介绍

相对危险度(Relative Risk，RR)是医疗领域中评估某风险对发病率影响的一个重要衡量标准。相对危险度的计算方法是暴露组的发病率与对照组的发病率之比，其中，暴露组是指暴露在某风险因素下的样本人群，而对照组是指与该风险因素隔离的样本人群。例如，想要统计孕期感冒是否存在诱发儿童先天性心脏病的风险，则需要统计的数据如表1所列。暴露组为孕期患过感冒的孕产妇群体，其新生儿患先心病的人数为 A，未患病人数为 B；对照组为孕期未曾感冒的孕产妇群体，其新生儿患先心病的人数为 C，未患病人数为 D。计算孕期感冒对儿童先心病的相对危险度的方法如式(1)所示。

$$RR=\frac{A/(A+B)}{C/(C+D)} \tag{1}$$

表1 孕期感冒与儿童先心病关系的统计

Table 1 Statistics of relationship between cold during pregnancy and congenital heart disease in children

	患病	未患病
暴露组(孕期感冒)	A	B
对照组(孕期未感冒)	C	D

若相对危险度大于1，则表明该暴露因素是风险因素，会提升相应疾病的患病概率，应尽量避免。该数值越大，表明危险性越高。若相对危险度小于1，则表明该暴露因素是保护因素，会降低相应疾病的患病概率，且数值越小，保护性越强。数值区间与疾病的关联性强弱的对照如表2所列[19]。

表2 暴露因素相关性的对比

Table 2 Comparison of exposure factors' relativity

RR数值区间	相关性
0.9～1.1	不相关
0.7～0.8或1.2～1.4	弱关联
0.4～0.6或1.5～2.9	中关联
0.1～0.3或3.0～9.9	强关联
<0.1或>10	很强关联

3 基于相对危险度的风险因素分析算法的设计

给定一个疾病数据集 $D=\{t_1,t_2,\cdots,t_n\}$，数据集中的每条记录可表示为一个二元组 $t_i=(x_i,y_i)$，二元组中的第二项 y_i 表示是否患病(用0/1表示，0表示未患病，1表示患病)，第一项 x_i 则是该条记录其他属性项的集合，$x_i=\{i_1,i_2,\cdots,i_m\}$。属性项的集合称为项集，若项集中项的个数为 k，则称之为 k 项集。

对于一个项集 I，该项集的相对危险度计算如式(2)－式(6)所示。

$$RR=\frac{A/(A+B)}{C/(C+D)} \tag{2}$$

$$A=|\{t_i\mid I\subseteq x_i,y_i=1,t_i\in D\}| \tag{3}$$

$$B=|\{t_i\mid I\subseteq x_i,y_i=0,t_i\in D\}| \tag{4}$$

$$C=|\{t_i\mid I\not\subset x_i,y_i=1,t_i\in D\}| \tag{5}$$

$$D=|\{t_i\mid I\not\subset x_i,y_i=0,t_i\in D\}| \tag{6}$$

MARRI算法需要设置一个最小的相对危险度作为阈值，该值必须大于1.1。相对危险度大于阈值的项集称为高风险项集。MARRI算法的目标是挖掘出数据集中所有的高风险项集。

高风险项集的挖掘过程可分为6步，其伪代码表示如算法1所示。

步骤1 输入数据集 D 和指定的最小相对危险度($min\_rr>1$)。

步骤2 统计出数据集中的患病例数、健康例数以及总例数，以备后续计算使用。

步骤3 第一次扫描数据集，找出高风险1项集，即相对危险度大于阈值的属性项。

步骤4 根据高风险 $k-1$ 项集的交叉组合，得到候选的高风险 k 项集。

步骤5 扫描数据集，计算各个候选高风险 k 项集的相对危险度；筛选出相对危险度大于阈值的 k 项集，即为高风险 k 项集。

步骤6 返回步骤4循环操作，直至无法产生新的候选项集。

算法1 MARRI

```
Input：数据集 D，最小相对危险度 min_rr
Ouput：高风险项集 L
1. health_num，ill_num，total_num＝scan_count(D)
2. L1＝ find_risk_1 _itemsets(D)
3. for k＝2；Lk−1 !＝Φ；k＋＋ do
4.     Ck＝candidate_itemsets_gen(Lk−1)
5.     for each transation t in dataSet do
6.         Ct＝subset(Ck，t)
7.         for each candidate in Ct do
8.             if it is health then
9.                 c. health＋＋
10.            else
11.                c. ill＋＋
12.            end if
13.        end for
```

```
    end for
    for c in C_k do
        total_c=c.ill+c.health
        c.rr=(c.ill/total_c)/((ill_num-c.ill)/(total_num-total_c))
    end for
    L_k={c∈C| c.rr>min_rr }
end for
return L∪_k L_k
```

图 1 给出了一个高风险项集挖掘过程的简单示例。数据集中有 7 条记录，含 3 条患病记录和 4 条未患病记录，设置最小相对危险度为 1.3，最终挖掘到 4 个风险项集{A}，{B}，{C}，{B,C}，对应的相对危险度分别为 1.5，2.67，1.5，2.67。

数据集

TID	X(项集)	Y(是否患病)
1	A,D	0
2	C	0
3	A,B,C,D	0
4	D	0
5	A	1
6	A,B,C	1
7	B,C,D	1

高风险1项集

项集	相对危险度
A	1.50
B	2.67
C	1.50

候选高风险2项集

项集	相对危险度
A,B	1.25
A,C	1.25
B,C	2.67

高风险2项集

项集	相对危险度
B,C	2.67

图 1 挖掘过程的示例图

Fig. 1 Exemplary diagram of mining process

该算法使用了医疗领域常用的统计标准——相对危险度，来代替传统 Apriori 算法中的支持度。相对危险度反映的是暴露组和对照组中病例比重的差异，而非某一因素在数据集中的出现频率。因此，面对不平衡的数据集，比例极小的数据项也不会被过滤掉。

4 剪枝方法

高风险项集挖掘算法通过高风险 $k-1$ 项集的交叉组合得到下一层的候选高风险 k 项集。若没有阈值的强烈约束，随着项数 k 的不断增大，高风险项集数目可能呈现爆炸式增长，这一方面会造成计算资源的浪费，另一方面会造成大量冗余的高风险项集，对后续的结果分析工作造成压力，甚至有可能导致因无法分析而使结果作废的情况。因此，适当的剪枝操作是必要的。我们采用的剪枝方法从作用叠加和样本数两方面进行限制。

4.1 作用叠加剪枝

高风险项集的挖掘由 $k-1$ 项集产生 k 项集，我们预期的是 k 项集的相对危险度高于 $k-1$ 项集的相对危险度，这说明暴露因素的组合产生的影响是正向叠加的，而非相互干扰。因此，我们在对候选高风险项集进行筛选时，除了与最小相对危险度进行比较以外，还加入了与高风险 $k-1$ 项集相对危险度的比较。伪代码表示如算法 2 所示。

算法 2 PruningByInteraction( )

```
Input: 候选高风险 k 项集 C_k，项集相对危险度记录字典 Dic，最小相对危险度 min_rr
Output: 高风险 k 项集 L_k
L_k=[]
for c in C_k do
    c.rr=calculate_rr(c)
    c_subsets=get_subset(c)
    max_sub_rr=get_max_value(Dic,c_subsets)
    if c.rr>min_rr and c.rr>max_sub_rr then
        L_k.append(c)
    end if
end for
return L_k
```

4.2 样本数剪枝

因为相对危险度使用样本比值进行计算，对样本数没有约束，所以会存在相对危险度突出但实际上样本数并不具有统计意义的情况，这种情况可被视为噪声干扰。例如，存在一个项集{A,B,C}，数据集中包含该项集的记录数仅为 3，即使这 3 条记录均为患病数据，且计算得到该项集的相对危险度大于阈值，该结果的可疑性仍然很大，不具有统计意义，应当排除。因此我们提出基于样本数的剪枝，指定结果具有统计意义的项集最小样本数，大于该样本数阈值的项集才纳入高风险项集。伪代码表示如算法 3 所示。

算法 3 PruningByNum( )

```
Input: 候选高风险 k 项集 C_k，项集相对危险度记录字典 Dic，最小相对危险度 min_rr，最小样本数 N
Output: 高风险 k 项集 L_k
L_k=[]
for c in C_k do
    c.rr=calculate_rr(c)
    c.num=count(c)
    if c.rr>rr_risk and c.num≧N then
        L_k.append(c)
    end if
end for
return L_k
```

5 实验及结果分析

本文在儿童先心病数据集上进行实验以验证算法效果。该数据集共包含先心病儿童记录 16 329 例，健康儿童记录 7 714 例，总共 24 043 条记录。数据集共有 30 个属性，包含了如民族、性别、胎龄等儿童基本资料内容，如发烧、抗生素、镇静药等环境因素内容，以及如近亲婚配、遗传缺陷等过往家族史内容。经缺失值填充和数值属性离散化操作后，共划分为 70 个属性项，如表 3 所列。在所有属性项中，支持度小于 0.2 的属性项共 49 个，占比 70%；支持度小于 0.1 的属性项共 35 个，占比 50%，即数据集中存在大量低支持度项集。

表 3 数据集简介

Table 3 Introduction of dataset

心脏病例数	健康例数	总例数	属性项总数
16329	7714	24043	70

我们使用基于相对危险度的高风险项集挖掘算法

(MARRI)和基于支持度的频繁项集挖掘算法(Apriori[1])分别对数据集进行挖掘，并对比了挖掘结果。将 MARRI 算法的最小相对危险度设置为 1.2，并进行样本数剪枝来消除噪声干扰，最小样本数设置为 100，Apriori 算法的最小支持度设置为 0.2。因为挖掘结果的项集较多，所以无法全部展示，下面仅展示了部分项集，如表 4 所列，表内项集均为随机选取，彼此独立，无顺序关联。与 Apriori 算法相比，MARRI 算法的优点表现在以下 3 个方面。

表 4 基于相对危险度和基于支持度的部分关联项集的对比

Table 4 Comparison of partial relative itemsets based on relative risk and support

基于支持度		基于相对危险度		
项集	支持度/%	项集	相对危险度	支持度/%
产次=1	55.33	三胎及以上	1.47	0.14
孕次=1	33.84	产次>3	1.21	0.58
城镇户籍	68.19	文盲	1.20	0.41
性别女	43.06	同卵	1.42	3.32
大专及以上学历	44.75	产次=3	1.25	3.49
汉族	93.35	双胎	1.41	5.54
单胎	94.31	异卵	1.37	2.36
年收入 8000+	63.25	年龄<20	1.38	15.1
性别男	54.22	体重 2.5～3kg	1.42	11.27
孕次=2	28.14	产次=2	1.27	23.54
农村户籍	31.81	胎龄 38～39 周	1.57	18.86
性别女；汉族	40.14	胎龄 36～37 周	1.45	10.24
大专及以上学历；汉族	42.64	胎龄 42 周以上	1.52	11.11
汉族；年收入 8000+	60.51	体重 2～2.5kg；双胎	1.45	1.06
单胎；产次=1；性别女	23.12	胎龄 36～37 周；双胎	1.46	2.07
单胎；城镇户籍；汉族	60.74	体重 2～2.5kg；胎龄 36～37 周；双胎	1.46	1.01

(1)可挖掘到低支持度数据。Apriori 算法挖掘的是频繁项集，即数据集中出现频率较多的项集。在疾病数据中，频繁项集往往意味着正常。由表 4 列出的频繁项集可知，这些项集描述了实例通常具有的特征，但这些特征在先心病儿童和未患病儿童中普遍存在，从而导致的结果是，众多频繁项集所能推导出的形如“项集 A=>患病”的关联规则数为 0。也就是说，挖掘到的频繁项集对了解疾病几乎没有帮助。而 MARRI 算法根据相对危险度挖掘高风险项集，表 4 列出的高风险项集如文盲、多胎、多产次等都是低支持度项集，即 MARRI 算法可以挖掘出这些低支持度的属性项与儿童先心病之间存在的关联。

(2)支持属性的细粒度划分。MARRI 算法不对支持度做限制，因此可以支持属性的细粒度划分。如本例中的胎龄以两周为一个间隔进行划分，在 Apriori 算法中，因为任一分区都不满足支持度大于 0.2 的要求，所以该属性被废弃了；但在 MARRI 算法中，29 周以下和 36 周以上的胎龄的属性项都被挖掘出来。

(3)调节阈值即可增强限制。若 Apriori 算法想挖掘低支持度的属性项，则必须降低支持度阈值。但由表 5 可知，降低支持度阈值将使挖掘到的频繁项集数目激增，一方面会造成大量冗余结果，为后续分析和处理造成困扰；另一方面也消耗了更多的计算资源。但 MARRI 算法通过简单地调整相对危险度阈值，即可限制项集的关联性强度，阈值越高，风险性越大，且阈值限制越强，计算消耗越小。

表 5 频繁 k 项集与高风险 k 项集数量的对比

Table 5 Itemset quantity comparison of frequent k-itemsets and high relative risk k-itemsets

不同的 k 值	阈值条件					
	Apriori			MARRI		
	sup>0.3	sup>0.2	sup>0.1	RR>1.4	RR>1.2	RR>1.1
k=1	15	21	35	10	20	24
k=2	40	81	213	20	106	170
k=3	40	124	473	8	139	337
k=4	16	80	514	1	36	169
k≥5	0	16	380	0	0	10
小计	111	322	1615	39	301	710

为验证第 4 节所提两种剪枝方法的效果，我们分别在样本数剪枝和作用叠加剪枝的情况下进行了数据挖掘实验。

(1)样本数剪枝效果的对比。在最小相对危险度为 1.2 的前提下，本文分别统计了最小样本数限制为 0，10，50，100 这 4 种情况下所挖掘到的高风险项集数。在数据统计方面，样本数越多意味着统计分析结果的可信度更高。因此，不断提高最小样本数限制意味着挖掘到的项集是风险因素的可信度更高。例如，项集{“三胎及以上”}的相对危险度为 1.47，属于高风险项集，但包含该项集的样本数仅有 23 条，这导致该条结果的可信度不高，因为该结果可能是样本采集过程中的偶然性因素使其存在较大误差而产生的；而高风险项集{“年龄<20”}的相对危险度为 1.38，低于前者的 RR，但包含该项集的样本数多达 2 465 条，这使得命题“低龄产妇的新生儿患先心病的风险更大”具有极高的可信度，该条结果也更加有意义。如图 2 所示，随着最小样本数限制的不断增强，挖掘到的高风险项集个数不断减少，但其结果的可信度将不断提高。同时，由图 2 可知，即使最小样本数仅设置为 10，也可以大幅度地减少高风险项集的个数。这表明在不对样本数进行限制的情况下，算法会产生大量不具有统计意义的冗余结果，即可信度极低的结果。因此，在实际应用中，样本数剪枝是必不可少的一个环节。

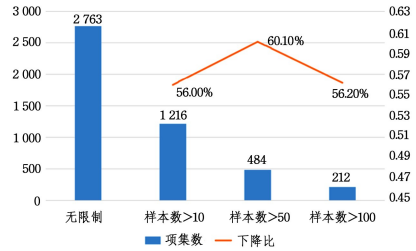


图 2 样本数剪枝效果对比图

Fig. 2 Comparison of sample number pruning effect

(2)作用叠加剪枝效果的对比。在最小相对危险度设置为 1.2 的前提下，本文分别统计了无作用叠加限制和有作用叠加限制这两种情况下所挖掘到的高风险项集数目，结果如图 3 所示。无剪枝的高风险项集总数为 2 763，而作用叠加剪枝后的高风险项集总数为 132，下降了 95.2%。因此，作用叠加限制较样本数限制更加强烈。推测原因有两点：1)属性项

之间本身存在复杂的关联关系，即使同为高风险因素，因素之间也有相互排斥的情况；2)高风险项集的支持度本身就低，在样本数较少的情况下，数量的微小波动或偏差就会对相对危险度计算结果造成较大影响，因此可能存在两个关联项之间的作用叠加，但因为样本数过少，所以统计偏差造成叠加的高风险项集被过滤掉。

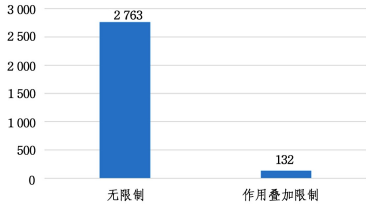


图 3 作用叠加剪枝效果对比图

Fig. 3 Comparison of interaction pruning effect

结束语 本文针对医疗疾病数据的不平衡性问题，基于医学统计指标——相对危险度，提出了一种挖掘疾病高风险项集的算法 MARRI 和相应的剪枝方法(即作用叠加剪枝和样本数剪枝)。在儿童先心病数据集上进行实验，验证了 MARRI 算法具有聚焦低支持度数据的能力，以及两种剪枝方法均可有效地减少冗余的高风险项集。后续将在本文的基础上继续研究关联规则在疾病分析上的应用，探讨属性之间的关系(如因果关系、主从关系、互斥关系等)对关联规则结果的影响，尝试使关联规则结果更加清楚并准确地传达信息。

参 考 文 献

[1] AGRAWAL R, IMIELIŃSKI T, SWAMI A. Mining association rules between sets of items in large database[J]. ACM SIGMOD Record, 1993, 22(2): 207-216.

[2] GAO L, WANG J, LI F G, et al. Symptoms-herbs relationship in lung diseases based on association rules[J]. Journal of Traditional Chinese Medicine, 2013, 54(8): 697-700.

[3] JIANG Y R, XIE Y H, ZHANG J C, et al. Data mining of the medication rule of Chen Keji in the treatment of blood stasis syndrome of cardiovascular disease[J]. Journal of Traditional Chinese Medicine, 2015, 56(5): 376-380.

[4] LI Q, CHEN D T, LUO X L. Implementation of the association rule algorithm in medical big data[J]. Software Engineering, 219, 22(1): 12-15.

[5] LEE W H, WANG E T, CHEN A L P. Mining accompanying relationships between diseases from patient records[C]// IEEE International Conference on Big Data. IEEE, 2018: 3861-3868.

[6] WANG M X. The prediction model for disease based on logistic regression and association rules[D]. Jinan: Shandong University, 2016.

[7] GAO S Y, CHENG S Z. Application of clustering-based entropy weighted association analysis[J/OL]. [2019-01]. http://dpi-proceedings.com/index.php/dtcse/article/view/27565.

[8] OJHA D, PANDEY P. Optimizing Association Rule using Genetic Algorithm and Data Sampling Approach[J]. International Journal of Computer Applications, 2018, 179(11): 15-19.

[9] DING Y, ZHU C S, WU Y Y. Association Rule Mining Algorithm Based on Hadoop[J]. Computer Science, 2018, 45(11A): 409-411, 416.

[10] IBRAHIM A, SHEHADA D. Study of Association Rule Mining for Discovery of Frequent Item Sets on Big Data Sets[J]. International Journal of Materials Science, 2018, 13(4): 345-358.

[11] LIU Z, HU L, WU C, et al. A novel process-based association rule approach through maximal frequent itemsets for big data processing[J]. Future Generation Computer Systems, 2018, 81: 414-424.

[12] WANG Q S, JIANG F S, LI F. Multi-label Learning Algorithm Based on Association Rules in Big Data Environment[J]. Computer Science, 2020, 47(5): 90-95.

[13] RATHEE S, KASHYAP A. Adaptive-Miner: an efficient distributed association rule mining algorithm on Spark[J]. Journal of Big Data, 2018, 5(1): 6.

[14] VOUGAS K, KROCHMAL M, JACKSON T, et al. Deep learning and association rule mining for predicting drug response in cancer[J/OL]. https://www.biorxiv.org/content/10.1101/070490v3.full.

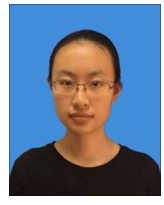
[15] KHAN A, USMAN M. Early diagnosis of Alzheimer's disease using machine learning techniques: a review paper[C]// 2015 7th International Joint Conference on Knowledge Discovery. IEEE, 2016: 380-387.

[16] ZHOU W, NIELSEN J B, FRITSCHE L G, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies[J]. Nature Genetic, 2018, 50(9): 1335-1341.

[17] WANG W P. A dissertation for the master degree of engineering[D]. Zhangzhou: Minnan Normal University, 2016.

[18] CUI X J. The study of association rule based classification for imbalanced data[D]. Dalian: Dalian University of Technology, 2015.

[19] LI M L. Epidemiology[M]. Beijing: People's Medical Publishing House, 2008: 71.



XU Hui-hui, born in 1995, postgraduate. Her main research interests include data mining and so on.

YAN Hua, born in 1970, Ph.D, associate professor. Her main research interests include computational intelligence and data mining.