一种基于模糊选项关系的关键属性提取方法

熊 熙[1)] 乔少杰[1)] 韩 楠[2)] 元昌安[3)] 张海清[4)] 李斌勇[1)]

[1)](成都信息工程大学网络空间安全学院 成都 610225)

[2)](成都信息工程大学管理学院 成都 610103)

[3)](广西师范学院计算机与信息工程学院 南宁 541004)

[4)](成都信息工程大学软件工程学院 成都 610225)

摘 要 模糊分析方法已广泛应用于医学实践包括对心理疾病的辅助诊断.属性约简方法在过滤冗余信息并提取关键信息时起到了重要作用,使整个临床决策过程更加准确和高效.这些方法抽取的有价值信息可以从新的视角揭示深层次医学知识.很多未经培训的参与者很难识别心理量表中选项间模糊的界线,即很难区分拥有相同意义但程度不同的选项.临床心理学自身的模糊性和心理测量数据的模糊性都将带来噪声.如果将心理测量数据中的属性看作信息系统的条件属性,利用降维算法可提取关键属性,从而简化对疑似患者的临床筛查过程.实际使用时,可对提取的关键属性或者拥有高权重的属性进行重点关注,从而迅速定位拥有异常关键属性的患者,对其优先处理.由此该文提出一种称为 FOAD(Fuzzy-Option based Attribute Discriminant method)的基于模糊选项关系的关键属性提取方法,包括三个主要步骤:数据获取、模糊选项的选择与约简以及关键属性的排序与提取.每个参与者样本包含若干身体症状属性,为每个属性都选择一个程度选项.选择模糊选项时须同时考虑选择该选项的样本数量和选项的程度含义.而模糊选项约简算法作为整个方法的核心,可以将模糊选项合并到其他选项,以降低心理测量数据中选项的模糊度.实验中采用两个真实临床数据集验证 FOAD 算法的性能.首先使用各种属性提取算法对测试数据集进行处理,获取关键属性,然后将输出的关键属性作为条件属性,以诊断结论作为分类标签,利用逻辑回归方法对样本数据进行分类.实验结果表明:FOAD 算法在不增加时间复杂度的前提下能将分类准确率普遍提高3.3%~14.1%.虽然选项约简操作造成部分信息的损失,但是合并模糊选项使选项分布更加清晰.FOAD 作用下的 LDA(Linear Discrimination Analysis)对各种参数敏感,尤其是对保留属性的个数.LDA 的预测准确率从保留最少属性时提高 6.7%,上升到保留最多属性时提高 14.1%. PCA(Principal Component Analysis)算法选择的投影方向会使数据方差最大,保留的信息量最多,但分类效果差.因此 FOAD 算法很难应用于提高 PCA 的预测准确率,甚至在个别情况下,出现了 FOAD 引起 PCA 分类准确率降低的情况.此外,实验发现基于 FOAD 的 LDA 算法比其他属性模糊提取算法具有更高预测准确率.心理诊断数据具有明显的模糊性,一般的统计分析方法往往不能得到需要的结果.而利用最新的模糊集和粗糙集等特殊的数据预处理方法可以消除这种数据噪声,提高临床诊断效果.

关键词 选项约简;模糊集;医学数据挖掘;临床决策;属性提取

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2019.00190

A Fuzzy-option Based Attribute Discriminant Method

XIONG Xi[1)] QIAO Shao-Jie[1)] HAN Nan[2)] YUAN Chang-An[3)] ZHANG Hai-Qing[4)] LI Bin-Yong[1)]

[1)](*School of Cybersecurity, Chengdu University of Information Technology, Chengdu* 610225)

[2)](*School of Management, Chengdu University of Information Technology, Chengdu* 610103)

[3)](*School of Computer and Information Engineering, Guangxi Teachers Education University, Nanning* 541004)

[4)](*School of Software Engineering, Chengdu University of Information Technology, Chengdu* 610225)

Abstract Fuzzy analysis method has been widely used in medical domains including auxiliary diagnosis of mental diseases. Attribute reduction methods play an important role in filtering

收稿日期:2017-09-27;在线出版日期:2018-05-14.本课题得到国家自然科学基金(61772091,61802035)、教育部人文社会科学研究青年基金(17YJCZH202)、四川省科技计划项目(2018GZ0253,2018JY0448)、成都信息工程大学科研基金(KYTZ201637,KYTZ201715,KYTZ201750)、成都市软科学研究项目(2017-RK00-00125-ZF,2017-RK00-00053-ZF)、成都信息工程大学中青年学术带头人科研基金(J201701)、四川高校科研创新团队建设计划(18TD0027)、广西自然科学基金项目(2018GXNSFDA138005)、广东省重点实验室项目(2017B030314073)资助.**熊 熙**,男,1983 年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为数据挖掘、机器学习和社会计算.E-mail:xiongxi@cuit.edu.cn.**乔少杰**(通信作者),男,1981 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为大数据、移动对象数据库和复杂网络.E-mail:sjqiao@cuit.edu.cn.**韩 楠**,女,1984 年生,博士,讲师,主要研究方向为数据挖掘.**元昌安**,男,1964 年生,博士,教授,主要研究领域为数据挖掘.**张海清**,女,1986 年生,博士,副研究员,主要研究领域为数据挖掘与决策系统.**李斌勇**,男,1982 年生,博士,讲师,主要研究方向为大数据和云计算.

redundant information and extracting essential information, and facilitating the whole decision-making process. Valuable information extracted by these methods can reveal underlying medical knowledge through a novel perspective of clinical medicine. It is difficult for many untrained participants to identify the fuzzy boundaries between the options in psychometric scales, i. e. , it is difficult to distinguish options with the same meaning and different degrees. The noise data are generated due to the intrinsic fuzziness of clinical psychology and the psychometric data. If the attributes of psychological data are viewed as the condition attributes of an information system, the key attribute can be obtained by attribute discriminant methods, which will simplify the clinical screening process for suspected patients. This study focuses on the extracted key attributes or the attributes with high weight values, in order to quickly discover the patients with abnormal key attributes and give them prior treatment. A Fuzzy-Option based Attribute Discriminant method is proposed, called FOAD, which contains three main phases: data collection, fuzzy option selection and reduction as well as sort and extraction of key attributes. In regard to psychometric data, each sample contains several physical symptoms, which can be viewed as attributes, then it selects an option for each attribute. It is necessary to take the number of samples and the meanings of options into consideration simultaneously when selecting fuzzy options which will be removed. As the key part of the whole approach, the fuzzy option reduction algorithm can merge fuzzy options into other reserved options in order to reduce the fuzziness of psychometric data. Two real clinical datasets are used to verify the performance of FOAD algorithm. The key attributes are obtained from datasets by multiple categories of attribute discriminant algorithms. Then, it classifies samples by logistic regression based on the key attributes and diagnosis results, which are viewed as conditional attributes and classification labels, respectively. The experimental results on the real datasets demonstrate that the prediction accuracy can be improved by 3.3% — 14.1% without increasing the computational complexity. Although the operation of option reduction loses some information in datasets, the option distribution becomes clearer by the merging operation. Linear Discrimination Analysis (LDA) under FOAD is sensitive to various parameters, especially to the number of reserved attributes. The prediction accuracy of LDA is increased from 6.7% when reserving the least attributes to 14.1% when reserving the most attributes. Principal Component Analysis (PCA) algorithm chooses the projection direction with the maximal variance of data and retains the maximal information. Due to the poor classification performance, PCA can hardly be improved through FOAD. The prediction accuracy of PCA degrades even under some specific conditions. Moreover, LDA based on FOAD demonstrates better prediction accuracy than other fuzzy attribute discriminant methods. It is concluded that it is difficult to process the fuzzy clinical psychometric data by conventional statistical analysis methods. The special preprocessing methods, such as the state-of-the-art fuzzy set and rough set techniques, can eliminate the noise of data and improve the clinical diagnosis effect.

Keywords　option reduction; fuzzy sets; medical data mining; clinical decision-making; attribute discrimination

1　引　言

近些年来，人工智能在许多复杂场景中得到广泛应用[1]. 以模糊集理论和粗糙集理论为代表的模糊分析方法使决策过程更加准确和高效，并已用于医学实践，其中就包括对心理疾病的诊断[2]. 在模糊分析过程中，属性约简算法和决策方法在过滤冗余

信息并提取关键信息时起到了重要作用.从临床医学的角度来看,这些方法抽取的有价值信息可以对疾病进行辅助诊断,并能从新的医学视角揭示深层次医学知识[3].

心理疾病近年来在世界范围内迅速蔓延,成为一种发病率较高的疾病,世界卫生组织 2009 年[4]预测世界上四分之一的人在一生中将受到心理疾病或神经疾病的困扰.心理测试量表是评估心理疾病的重要手段,可辅助心理医生进行相关疾病的诊断.典型的心理量表是由若干问题组成的问卷,参与者需要根据自身心理和精神状况给出每个问题的答案.每个问题实际上都反映用户的一个属性,而每个属性只有几个选项,分别表示症状的不同程度,例如严重的、一般的、轻微的.参与者只能从这几个选项中进行选择.

PHQ-13 是一个典型心理测试量表,包含 13 个问题,每个问题有 5 个选项,分别表示不同症状程度.表 1 是 100 个样本患者分别完成 PHQ-13 问卷后汇总所得到的数据,每个症状可以选择一个选项,即范围在 1 到 5 之间的一个整数.表 2 表示每个选项值所表示的症状程度.

表 1 PHQ-13 量表[5]包括 13 个问题和 5 个选项

样本	1.胃痛	2.背痛	3.四肢痛	…	13.睡眠问题
样本 1	2	1	1		2
样本 2	3	3	2	…	1
…				…	
样本 100	5	1	5	…	2

表 2 PHQ-13 量表中每个选项值的含义

无	轻微	中度	较严重	严重
1	2	3	4	5

从这两个例子可以看到心理量表中包含可用于诊断心理疾病的重要信息,因此,从心理量表数据中提取关键属性主要基于如下几点考虑:

(1)很多未经培训的参与者很难识别心理量表中选项间模糊的界线,即很难区分拥有相同意义但不同值的选项.如果他们不能准确理解量表中问题的意思,就会选择一些近似的选项,从而给医生确诊病情带来障碍.这些情况都会产生冗余数据.临床心理学自身的模糊性和心理测量数据的模糊性都将带来噪声.因此在从模糊心理测量数据中挖掘有价值信息和提取关键属性之前,有必要降低数据的模糊性.

(2)虽然目前实验标准数据的维度大多只有 10～200 维,但是对于心理医生临床诊断来说,其中很多维度信息具有模糊性,可以将模糊选项合并到其他选项,以降低心理测量数据中选项和属性的模糊度;此外,属性维度较多会极大地降低心理医生诊断病情的效率.如果将心理测量数据中的属性看作降维算法中的特征,利用降维算法可提取关键属性,从而简化对疑似患者的临床筛查过程.实际使用时,可对提取的关键属性或者拥有高权重的属性进行重点关注,从而迅速定位拥有异常关键属性的患者,并对其优先处理.因此,基于模糊选项关系的关键属性提取方法对挖掘心理测量数据中的有用信息具有重要的研究意义和应用价值.

(3)诊断是一个复杂的过程,经常出现错误.构建基于大量的心理疾病诊断数据的专家系统,对辅助心理医生或相关专家做出相对准确的诊断决策具有极大的帮助.

通过仔细分析模糊的心理测量数据,本文提出一种关键属性提取方法,以辅助心理医生和相关专家做出相对准确的诊断,主要贡献包括:

(1)提出了选项熵和选项影响度的概念.选项熵用于描述每个选项中包含的决策信息量,选项影响度则表示不同选项间的关联程度.

(2)提出了一种基于模糊选项关系的关键属性提取方法 FOAD(Fuzzy-Option based Attribute Discriminant method),该方法可以从心理量表的模糊数据中提取关键属性.该方法分为 3 个部分:①从数据库提取数据并进行数据预处理;②选择并约简模糊选项;③对属性排序并提取关键属性.

(3)基于临床数据集对 FOAD 算法进行了对比实验,结果表明选项的约简及合并降低了选项间的关联度,能有效改进已有属性提取算法.在不增加时间复杂度的基础上能将分类准确率提高 3.3%～14.1%.

2 相关工作

机器学习方法可以通过模仿人类推理过程来解决诊断中出现的困难并辅助决策,并且类似的方法也用于处理不确定和不完整的信息.Masri 等人[6]提出了一种用于诊断心理健康状况的专家系统.Rahman 等人[7]将贝叶斯网络、多层感知器、决策树、模糊干扰系统等多种分类方法应用于糖尿病的诊断,发现在使用不同测量手段时,这些分类方法的准确率存在差异.Seixas 等人[8]提出了一种用于诊

断痴呆和轻微认知障碍的贝叶斯网络决策模型，该模型在面对不确定性数据时的表现优于其他分类器. Khemphila 等人[9]发现带反向传播学习的多层感知器在属性约简前提下可有效诊断帕金森症. Dabek 等人[10]提出的神经网络模型在预测焦虑、行为紊乱、抑郁和灾后创伤等心理问题时的准确率达到了 82.35%. 虽然利用这些机器学习方法可以完成对包括心理疾病在内的不同疾病的诊断，但是它们却忽略了带有模糊性的心理测量数据中包含的一些重要特征.

属性约简[11]，是一种处理数据的常用方法，在许多研究领域具有重要的作用，其中包括模式识别、机器学习和数据挖掘等. 在诸多的属性提取方法中，Pawlak 的粗糙集模型[12]提供了一种典型理论研究框架. 可以通过粗糙集理论来为模式识别[13]选择最适合的属性子集. 粗糙属性选择又被称为属性约简[14]，可以在保持区分度的情况下选出最重要的特征[13]. 近年来又出现了一些新的约简算法[15]. Zhao 等人[16]在属性约简时考虑了属性的冗余，采用基于属性相似性的组合优化方法，将其扩展到多输出回归方法，以达到提高约简效果的目的. Liu 等人[17]提出了一种利用属性的区分度（也就是属性之间的差异）进行约简的方法. 现有约简算法都是从属性的维度进行约简，即利用属性值产生决策值过程中所出现的冗余数据查找多余属性，未考虑选项维度的模糊性，并且无法指定需保留关键属性的数目，在样本量较大时会出现属性约简效果较差的情况. 因此有必要构造一种新方法对这种模糊选项进行约简.

典型的降维算法可用于提取数据中的关键属性，在处理模糊性的心理测量数据时，采用基于模糊集的降维算法可以获得更好的效果. Wu 等人[18]提出一种基于 LDA 算法的模糊降维算法，每个数据具有不同的权重，该权重值由其重叠程度决定，从而使该算法在许多场景下的表现优于 LDA. Xu 等人[19]提出的算法中重新定义了模糊局部均值，使数据的模糊分布差异最大化，从而找到最优子空间，克服了其他模糊降维算法在处理局部模糊数据时的局限性. Zhao 等人[20]在模糊降维算法中引入了马氏随机游走和模糊隶属度的概念，可以最大程度地保留局部与全局的统计特征，并且可以有效地查找异常数据. 现有模糊降维算法都是针对属性维度的模糊性进行处理，却无法降低选项（即属性值）维度的模糊性. 例如表 2 中 5 个选项值具有相同的含义，而相邻选项值所表示的程度存在模糊的差异. 因此有必要构造一种数据预处理方法以减小模糊选项对属性提取算法性能的影响.

通过以上分析，可以发现上述算法主要关注具有明显边界的临床数据，很少考虑带有模糊性的心理测量数据. 这类数据不仅包含具有关联性的属性，而且拥有含义类似但程度不同的选项集. 典型的属性约简算法和模糊降维算法都很难应用于心理测量数据中关键属性的约简与提取.

3　基础理论

本节对算法中使用的主要概念进行形式化定义. 为描述方便，首先给出文中用到的一些符号，如表 3 所示.

表 3　文中用到的符号

符号	含义
U	论域
A	模糊选项集
O	选项集合
O'	保留的选项集合
μ_A	A 的隶属函数
R_σ	基于选项 σ 的等价关系
$H(\sigma)$	选项 σ 的熵
$Q(\sigma)$	选项 σ 的质量
$f_u^k(x)$	样本 k 中选项值 x 在选项 u 的隶属值
$r(\sigma_i,\sigma_j)$	选项 σ_i 与 σ_j 之间的影响度
$\boldsymbol{W}$	影响度关系矩阵

3.1　模糊选项集

不同于经典集合，模糊集用于表示一个元素属于一个集合的程度. 模糊集的特征函数被称为隶属函数，其取值范围在 0 到 1 之间. 类似地，可以给出模糊选项集的概念.

定义 1（模糊选项集）.　U 是对象的集合，其中每个对象用 x 来表示，则 U 中的模糊选项集 A 可定义为

$$A=\{(x,\mu_A(x))\mid x\in U)\} \tag{1}$$

其中 $\mu_A(x)$ 是模糊选项集 A 的隶属函数. 隶属函数将 U 中每个对象映射到位于区间 $[0,1]$ 的一个隶属值.

$$\mu_A: x\rightarrow\mu_A(x)\in[0,1] \tag{2}$$

通常选项集 O 由多个离散值组成. 模糊选项集的概念来源于模糊集，但是二者在含义上却有差异. 模糊选项集是一种特殊的模糊集，表示一系列具有相似属性意义而程度不同的选项集合，是模糊集在处理心理测量数据时的一个具体应用. 模糊选项集

可以由包含等间隔选项值的初始选项集合(如表 2)生成,而普通模糊集的数据却没有该特征.

此处选项值与属性值具有相同的含义,任一属性值都可以取选项集合中的一个选项值.

例 1. 如图 1 所示,一个对象由若干属性组成并且每个属性的原始值都是 1 到 5 之间的某整数值.$O=\{1,3,5\}$是选项集合,即每个原始属性值都可以用集合 O 及其中各元素的隶属度表示.那么,模糊选项集 $A=$"属性值 4 的归属选项"可以表示为

$$A=\{(1,0.1),(3,0.7),(5,0.5)\} \tag{3}$$

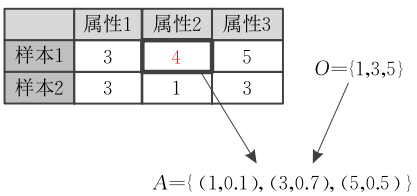


图 1 模糊选项集与隶属度

A 中每个元素都是无序离散对象.从上面的例子可以看出,模糊选项集包含两个要素:选项集合和隶属函数.一般隶属函数的选择较为主观,通常来源于对历史数据的经验分析.

定义 2(归一化隶属函数). 假设隶属函数服从正态分布,其最大值等于中心选项被某样本选中的次数.$\varphi(a,b)$表示变量 a 和 b 是否相等:

$$\varphi(a,b)=\begin{cases}1, & a=b\\ 0, & a\neq b\end{cases} \tag{4}$$

则 $\varphi(\boldsymbol{v},a)$表示向量 $\boldsymbol{v}$ 中值为 a 的元素个数:

$$\varphi(\boldsymbol{v},a)=\sum_i \varphi(v_i,a) \tag{5}$$

其中,v_i表示向量 $\boldsymbol{v}$ 的第 i 个元素.如果 $g_u^k(x)$表示样本 k 的中心为 u 的分布函数,则中心点 u 的值为

$$g_u^k(u)=\varphi(\boldsymbol{V}_k,u) \tag{6}$$

其中,$\boldsymbol{V}_k$表示向量集合 $\boldsymbol{V}$ 中第 k 个向量.将 $g_u^k(u)$归一化得

$$norm(g_u^k(u))=\frac{g_u^k(u)}{\max_u(g_u^k(u))}\in[0,1] \tag{7}$$

以 u 为中心的 x 归一化值为

$$f_u^k(x)=norm(g_u^k(x))=\frac{g_u^k(x)}{\max_u(g_u^k(u))} \tag{8}$$

例 2. 以选项 7 和选项 10 为中心的两个隶属函数,它们分别在选项 8 的位置产生不同的隶属值,$f_7^k(8)=0.75$,$f_{10}^k(8)=0.5$,这表示选项 8 更接近于选项 7 而不是选项 10.

3.2 选项数量差异

定义 3(选项数量差异). 选项数量差异表示选项之间的整体影响程度,可以定义为

$$r(\sigma_i,\sigma_j)=\frac{|n(\sigma_i),n(\sigma_j)|}{\max(n(\sigma_i),n(\sigma_j))}\in[0,1] \tag{9}$$

其中 $n(\sigma)$表示选项 σ 的个数.如果选项 σ_i远远多于选项 σ_j,则 $r(\sigma_i,\sigma_j)$的值接近于 1;否则,$r(\sigma_i,\sigma_j)$的值接近于 0.当两个选项接近并且很难被区分,样本会频繁选择其中一个选项而忽略另外一个.这种情况下,两个选项就处于彼此严重干扰的状态.

定义 4(影响度关系). $\boldsymbol{W}$ 是 $U\times\boldsymbol{V}$ 上的一个模糊选项集,其中 U 和 $\boldsymbol{V}$ 分别是不同的选项集,$U\times\boldsymbol{V}\triangleq\{u\in U,\boldsymbol{v}\in\boldsymbol{V}\}$是笛卡尔乘积.$\boldsymbol{W}$ 的隶属函数是:

$$\boldsymbol{W}:U\times\boldsymbol{V}\rightarrow[0,1] \tag{10}$$

$$(u,\boldsymbol{v})\rightarrow\boldsymbol{W}(u,\boldsymbol{v}) \tag{11}$$

该隶属函数表明 U 中的元素 u 和 $\boldsymbol{V}$ 中的元素 $\boldsymbol{v}$ 之间的相互影响,可以表示为 $U\xrightarrow{W}\boldsymbol{V}$,这种关系实际上是一种二元关系.特别地,$U$ 和 U 之间的关系可以表示为 $\boldsymbol{W}(U\times U)$.

定义 5(影响度关系矩阵). 影响度关系矩阵 $\boldsymbol{W}$ 表示选项集中任意两个选项之间的影响度.

$$W^k=[w_{ij}^k]_{n\times n},\ w_{ij}^k\in[0,1] \tag{12}$$

定义 6(选项影响度). 选项之间的影响度大小可以用影响度关系矩阵中的元素来表示.w_{ij}^k是矩阵 $\boldsymbol{W}^k$ 中的元素,表示选项 σ_i 对选项 σ_j 的影响度,其值为

$$w_{ij}^k=f_{\sigma_i}^k(\sigma_j)\times r(\sigma_i,\sigma_j) \tag{13}$$

例 3. 某测试数据共包含 10 个样本,5 个选项,样本 k 的所有属性中,选项 1~5 的数量如表 4 所示.

表 4 样本 k 的选项个数

选项	样本 k 的选项数量	总的选项数量
1	6	80
2	7	37
3	2	43
4	0	10
5	1	22

求样本 k 的影响度关系值 $w_{1,2}^k$ 和 $w_{2,1}^k$.

解. 假设隶属函数服从正态分布,归一化中心值 1 和 2 分别等于:

$$norm(g_1^k(1))=6/7=0.86,$$

$$norm(g_2^k(2))=1.$$

如果 $f_2^k(1)=0.75$,

则 $f_1^k(2)=0.75\times0.86=0.64$.

根据式(9),选项影响度可以计算得到

$r(1,2)=r(2,1)=(80-37)/80=0.54$，

进而计算得到

$w_{1,2}^k=f_1^k(2)\times r(1,2)=0.64\times 0.54=0.344$，

$w_{2,1}^k=f_2^k(1)\times r(2,1)=0.75\times 0.54=0.405$.

$w_{1,2}^k<w_{2,1}^k$，表明样本 k 中，选项 2 对选项 1 的影响度大于选项 1 对选项 2 的影响度.

4　基于模糊选项关系的关键属性提取方法

4.1　整体框架

在提出模糊选项约简方法之前，首先需要介绍心理量表数据的表示方法. 心理量表可以用 $n\times m$ 的矩阵 $\boldsymbol{T}$ 来表示，例如表 1 所示量表数据可以用下面的矩阵来表示：

$$\boldsymbol{T}_{100\times 13}=\begin{bmatrix}2 & 1 & 1 & \cdots & 2\\ 3 & 3 & 2 & \cdots & 1\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 5 & 1 & 5 & \cdots & 2\end{bmatrix}.$$

同时，$\boldsymbol{T}[i]$表示样本 i 的属性组成的向量，如 $\boldsymbol{T}[1]=\{2,1,1,\cdots,2\}$.

心理疾病临床诊断过程中，医生会使用各种心理测试量表来辅助诊断. 本文提出一种名为 FOAD 的方法，该方法可以利用大量心理测试数据对属性进行排序，提取关键属性，进而用于疾病的分类预测. 如图 2 所示，FOAD 方法可以分为 3 个主要步骤：

(1) 从数据库中提取心理测试数据集并对数据进行清理. 数据集中的每个参与者样本包含若干身体症状属性，他为每个属性都选择一个选项，以表示症状的程度.

(2) 约简冗余选项，保留主要选项，并将删除的选项数据合并到其他选项. 如图 2 所示，如果选项 2 和选项 4 是冗余选项，需要合并到其他选项，则根据选项之间的关系将所有值为 2 和 4 的属性修改为其他值.

(3) 利用典型的属性提取算法来提取关键属性. 如果同时需要对属性重要度进行排序，则可以使用线性回归方法.

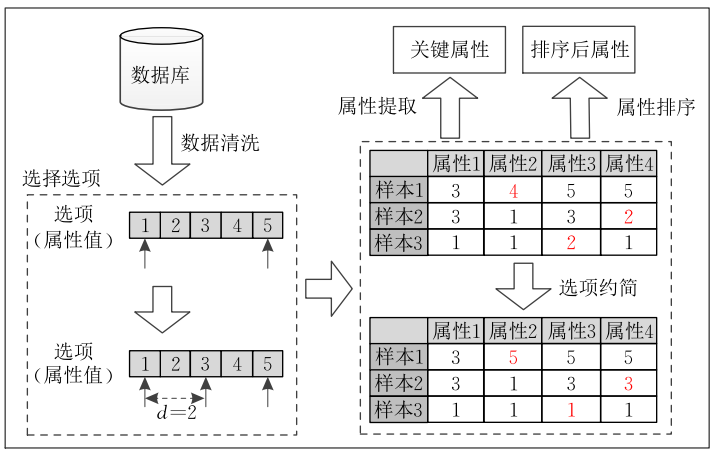


图 2　基于 *FOAD* 算法的关键属性提取方法的工作原理

下面将详细叙述这些步骤.

4.2　选择选项

量表中某属性的程度由量表选项来指定. 根据 5.1 节提供的 PTSD 数据集，可以统计出 5 个选项分别占 75.2%、19.5%、3.8%、1.1%和 0.4%. 这是因为患病的参与者毕竟是少数，因此程度越小的选项占比越大. 但是相近选项在程度上也是近似的，其模糊性使参与者在选择时存在误差. 因此保留哪些选项，不能完全根据选择该选项的样本数量来确定，而应该同时考虑其程度含义. 因此应该分别至少保留一个程度较大的选项和一个程度较小的选项，其他保留的选项根据选项间的距离来确定. 选项之间的距离定义为两个选项所表示的程度值之差的绝对值.

算法 1 给出了选择选项的具体过程，如下所示.

算法 1.　选择选项.

输入：$m\times s$：量表矩阵 $\boldsymbol{T}$，初始选项个数 s，保留选项个数 s'

输出：O'：保留选项的集合

1. IF $s'<2$ or $s'>\lfloor(s+1)/2\rfloor$ THEN
2. 　return;
3. END IF
4. max_L, $max_R\leftarrow GetMax(T)$;
5. add max_L, max_R to set O';
6. FOR EACH $i\in[1,s'-2]$ DO

7. FOR EACH $j\in[1,s]$ and $j\notin O'$ DO
8. $d[j]$=SearchMinDistance(O',j);
9. END FOR
10. $next\_max\leftarrow MaxOf(d)$;
11. add $next\_max$ to *set* O';
12. ENDFOR

算法 1 的工作原理如下：

(1)为保证选项约简的效率，保留选项的个数必须在 2 到$\lfloor(s+1)/2\rfloor$之间(第 1～3 行).保留选项的个数必须保证基本的程度含义，即至少有“无”和“严重”两个选项.此外，约简选项以后仍然需要保证选项含义的合理性，使删除的选项尽量等间距分布.例如 5 个选项可以保留选项组合{1,3,5}，但是不能保留选项组合{1,2,5}或者{1,2,3,5}.因此保留选项个数最大为$\lfloor(s+1)/2\rfloor$.

(2)将所有选项分为两部分，即$[1,\lfloor s/2\rfloor]$和$[\lfloor s/2\rfloor+1,s]$，并且分别查找两部分中拥有最大数量的选项 max_L和max_R(第 4 行)，从而保证约简后的选项集合涵盖了两种基本的程度含义.

(3)将 max_L，max_R加入到O'中(第 5 行).

(4)在剩余选项中查找其他 $s'-2$ 个需要保留的选项(第 6 行).

(5)对不在O'中的选项进行处理(第 7 行).

(6)计算该选项与O'中所有选项的最小距离为$d[j]$(第 8 行).

(7)获得 d 中最大元素并加入到O'中(第 10～12 行)，使得O'中元素之间的距离尽量大.

上述选择选项的方法同时考虑选项的含义和选项在整个数据集中的真实分布，因此保留选项集同时具有主观性和客观性，而不仅仅是基于经验进行选择.图 3 是选择选项的示意图，其中 $O=\{1,2,3,4,5,6,7\}$，需保留 4 个选项.

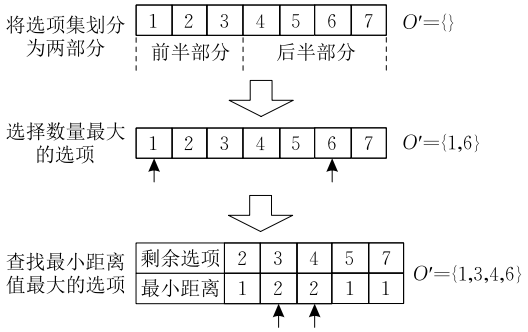


图 3 选择选项流程示意图

4.3 模糊选项约简

通常任意两个选项间存在关联，基于这一考虑，本文提出的模糊选项约简方法通过合并模糊选项以降低心理量表数据的选项模糊度.该方法可以降低参与者在面对模糊选项时随机选择所带来的误差，更能反映样本个体的真实情况.

当确定需要删除的选项之后，这部分选项可以合并到其他选项.O 和O'分别表示初始选项集和保留选项集，则$O''=O-O'$表示删除选项集.

对一个需删除的选项 $\sigma\in O''$，保留选项集中每一个选项 σ_j 对它的影响度可以表示为 $w(\sigma_j,\sigma)$，那么 σ 修改为 σ_j 的比例为 $\mu=w(\sigma_j,\sigma)\Big/\sum_{k=1}^{n_{O'}}w(\sigma_k,\sigma)$.参与者 i 选择选项 σ 的个数为 $n_A=\varphi(\boldsymbol{T}[i],\sigma)$，其中，$\boldsymbol{T}[i]$是参与者 i 的属性向量.修改为 σ_j 的选项 σ 的个数为 $n_j=\lfloor n_A\times\mu\rfloor$，$j\in O'-\sigma_{\max}$，其中，$\sigma_{\max}$表示 O'中影响度最大的选项.向下取整保证 σ 尽量修改为影响度最大的的选项 $\sigma_{\max}$.因此，选项 σ 修改为 $\sigma_{\max}$的个数为

$$n_{\max}=n_A-\sum_{j\in O'-\sigma_{\max}}w(\sigma_j,\sigma) \tag{14}$$

算法 2 给出了模糊选项的约简过程，如下所示.

算法 2. 模糊选项约简.

输入：O：初始选项集，O'：保留选项集，$\boldsymbol{T}$：量表矩阵($m\times s$)

输出：$\boldsymbol{T}'$：合并选项后的量表矩阵

1. $O''\leftarrow O-O'$
2. FOR EACH $i\in[1,m]$ DO
3. FOR EACH $\sigma\in O''$ DO
4. $n_A\leftarrow\varphi(\boldsymbol{T}[i],\sigma)$;
5. FOR EACH $\sigma_j\in O'$ DO
6. $w(\sigma_j,\sigma)\leftarrow CalcInfluence(\sigma_j,\sigma)$;
7. END FOR
8. $\sigma_{\max}\leftarrow$使 $w(\sigma_j,\sigma)$最大的选项;
9. $sum_\sigma\leftarrow\sum_{k=1}^{|O'|}w(\sigma_k,\sigma)$;
10. FOR EACH $\sigma_j\in O'-\sigma_{\max}$ DO
11. $w(\sigma_j,\sigma)\leftarrow CalcInfluence(\sigma_j,\sigma)$;
12. $\mu\leftarrow w(\sigma_j,\sigma)/sum_\sigma$;
13. $n_j\leftarrow\lfloor n_A\times\mu\rfloor$;
14. $Label\_j(\boldsymbol{T}[i],\sigma,\sigma_j,n_j)$;
15. END FOR
16. $n_{\sigma_{\max}}\leftarrow n_A-\sum_{j\in O'-\sigma_{\max}}w(\sigma_j,\sigma)$;
17. END FOR
18. END FOR
19. FOR EACH $i\in[1,m]$ DO
20. FOR EACH $j\in[1,s]$ DO
21. IF $\boldsymbol{T}[i,j]$ labeled as $Change\_k$ THEN
22. $\boldsymbol{T}[i,j]\leftarrow\sigma_k$;
23. END IF
24. END FOR
25. END FOR

算法 2 中，输入参数 $\boldsymbol{T}$ 是 $m\times s$ 的矩阵，表示包含 m 个样本和 s 个属性的数据集.

算法 2 的基本思想为

(1) O''是O'的补集，保存了需要删除的选项(第 1 行).

(2) 对某样本中需要删除的选项，逐一标记该选项对应的属性值需要修改为哪个保留选项(第 2～18 行).

(3) 根据式(13)计算每个保留的选项对该删除选项的影响度(第 5～7 行).

(4) 将影响度最大的选项记为 $\sigma_{\max}$(第 8 行).

(5) 计算所有保留选项的影响度之和(第 9 行).

(6) 标记该删除选项对应的属性需修改为除 $\sigma_{\max}$以外的哪个保留选项(第 10～15 行).

(7) 第 13 行是向下取整，为保证所有属性值都可以得到处理，此处将该删除选项对应的剩余未标记属性标记为 $\sigma_{\max}$(第 16 行).

(8) 修改所有待删除选项为标记的保留选项(第 19～25 行).

考虑到所有选项都会对删除选项产生影响，尤其是样本的属性值较为分散，即各选项的数量较为接近时，这种影响更为明显. 因此第 9 行需要计算所有保留选项对选项 σ 的影响度之和，而不仅仅是相邻保留选项对 σ 的影响度之和.

5　实验与结果

5.1　数据集介绍

为验证 FOAD 算法的性能，实验中采用四川大学华西医院提供的两个数据集，分别是：

(1) PTSD 数据集. 通过使用 PHQ-13 量表，可以评估 2013 年四川宝兴地震后六个月时幸存者的 PTSD(Post-Traumatic Stress Disorder，创伤后应激障碍)状况和躯体症状[5,21]. PHQ-13 量表是一个自评量表，其中包括 13 个问题，分别用于评估 13 种躯体症状的程度. 针对每一个问题，每个参与者需要从五个选项中选出一个合适的选项以表示他在该属性上的严重程度. 这五个选项分别是：① 没有困扰，② 有点困扰，③ 有中度困扰，④ 有较大困扰，⑤ 有很大困扰. PHQ-13 量表在中国已经被证明具有令人满意的可信度和一致度[22]. 删除冗余数据后，该数据集包括 3099 个有效样本，其中 51.9% 为女性，87.5% 是汉族，拥有从 14 岁到 91 岁的年龄跨度. 睡眠问题、感觉疲乏、恶心与消化道不适这三种症状因其与 PTSD 有较大的关联值[5]，被认为对 PTSD 的发病有最直接的影响. 表 5 是 PTSD 数据集的结构，与表 1 不同的地方在于每个样本都包含针对各种症状得到诊断结论，其中诊断结论的 0 表示未患病，1 表示患病.

(2) 精神障碍数据集. 该数据集来源于四川省青少年的流行病调查项目[23]，该项目的目的是了解精神疾病的各种症状行为的发生率并查找高风险因素. 20752 位 6 至 16 岁的青少年参与了本次测试，他们首先要回答 CBCL 问卷(Achenbench 儿童行为问卷)，最终在与心理医生的面谈中得到确诊. 中文 CBCL 问卷包括 113 个问题，每个问题包括 3 个选项，分别是：① 无，② 一般，③ 经常. 每个参与者需要依据近两个月的表现来为每个问题选择一个选项. CBCL 中包括了对多种精神疾病的分析，为缩小研究范围，只提取了与攻击行为相关的 23 个问题与数据. 与 PTSD 数据集类似，该数据集也包含基于症状的诊断结论.

表 5　PTSD 数据集的结构

样本	1. 胃痛	2. 背痛	3. 四肢痛	…	13. 睡眠问题	诊断结论
样本 1	2	1	1		2	0
样本 2	3	3	2	…	1	0
…				…		
样本 100	5	1	5	…	2	1

基于这两个心理量表数据集的诊断过程可以表示为信息系统 $S=\{U,C\cup D,V,f\}$. 其中 D 是决策属性集，即诊断结果，可以取 0 和 1，分别表示诊断为无病与有病；C 是条件属性集，PTSD 数据集中包含 13 个条件属性($|C|=13$)，而攻击行为数据集包含 23 个条件属性($|C|=23$). $\bigcup_{a\in C\cup D}V_a$ 由 V_a 组成，其中 V_a 表示属性 a 的值域，在 PTSD 数据集中取 1 到 5(条件属性)或者 0 和 1(决策属性)，在攻击行为数据集中取 1 到 3(条件属性)或者 0 和 1(决策属性). f 表示决策函数. 表 6 给出了两个数据集的概要信息.

表 6　两个数据集的概述信息

数据集	PTSD	攻击行为
采用的量表	PHQ-13	CBCL
样本数	3099	20752
属性数	13	23
选项数	5	3

5.2　对比算法

本文将 FOAD 方法进行数据预处理后的数据作为输入，对比验证该方法对 PCA(Principal Com-

ponent Analysis,主成份分析法)[24]和 LDA(Linear Discrimination Analysis,线性降维分析法)[25]这两种算法降维后的分类性能的影响,从而验证 FOAD 方法的有效性.为了对属性进行排序,对比的算法中还包括线性回归法.同时,作为一种数据预处理方法,选项约简与属性约简分别针对不同的数据维度进行处理,因此采用以下两种基于粗糙集的属性约简算法进行对比:

(1) SPSF-LAR(Similarity Preserving Feature Selection-Least Angle Regression,基于最小角度回归的相似性保留特征提取法)[16].在属性约简时考虑了属性的冗余,采用基于属性相似性的组合优化方法,并将其扩展到多输出回归方法,以达到提高约简效果的目的.

(2) QIFS[17](Quality of Information based Feature Selection,基于信息质量的特征提取法).利用属性的区分度(也就是属性之间的差异)进行约简.

为进一步验证基于 FOAD 的属性提取算法的分类性能,将采用以下属性模糊提取算法进行对比:

(1) FDA(Fuzzy Discriminant Analysis,模糊降维分析法)[18].与传统的 LDA 不同,FDA 中每个数据具有不同的权重,该权重值由其重叠程度决定,这使得 FDA 在许多场景下的表现优于 LDA.

(2) FLMDA(Fuzzy Local Mean Discriminant Analysis,模糊局部均值降维分析法)[19].FLMDA 算法重新定义了模糊局部均值,使数据的模糊分布差异最大化,从而找到最优子空间,克服了 FDA 等传统模糊降维算法在处理局部模糊数据时的局限性.

(3) MF-LDA(Markov-random-walks Fuzzy Linear Discriminant Analysis,马氏随机游走的模糊线性降维分析法)[20].该算法同时引入了马氏随机游走和模糊隶属度的概念,可以最大程度地保留局部与全局的统计特征,并且可以有效地查找异常数据.

表 7 是上述所有算法的主要特征的联系与区别.

表 7 对比算法的主要特征

对比算法	是否基于粗糙集	是否基于模糊集	特点
FOAD+PCA	否	是	尽可能保留原特征
FOAD+线性回归	否	是	属性排序
FOAD+LDA	否	是	类间差异最大
SPSF-LAR	是	否	多输出回归
QIFS	是	否	引入属性区分度
FDA	否	是	引入数据权重
FLMDA	否	是	数据的模糊分布差异最大化
MF-LDA	否	是	引入马氏随机游走

5.3 测试方法

本文提出的方法可以降低心理测量数据中选项的模糊度,获得在诊断过程中起决定作用的关键属性.将采用以下步骤比较上述各种算法的属性提取效果:

(1) 使用表 7 的各种属性提取算法对测试数据集进行处理,获取关键属性.

(2) 以步骤(1)输出的关键属性作为条件属性,以诊断结论作为分类标签,利用逻辑回归方法对样本数据进行分类.

(3) 对比不同的属性提取算法的预测(分类)准确率,其中预测准确率定义为

$$Accuracy=N_{\text{right}}/N,$$

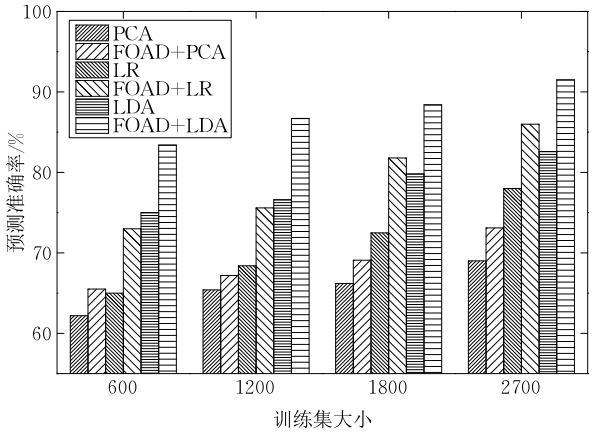
其中 N_{right} 分别表示诊断正确的样本,N 表示总样本.

测试过程中采用十折交叉验证法去除不相干因素对结果的影响.

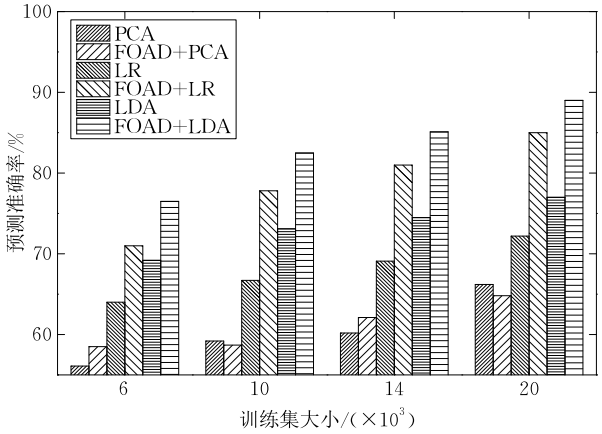
5.4 FOAD 预处理效果对比

在使用 PTSD 数据集和攻击行为数据集时分别保留三个选项和两个选项.图 4 和图 5 分别展示了不同算法在两个数据集中对比的结果.

如图 4 和图 5 所示,FOAD 算法使属性提取算法提高了至少 7.1%的预测准确率,其中唯一的例外是 PCA 算法.PCA 算法是一种无监督算法,选择的投影方向会使数据方差最大,保留的信息量最多,但分类效果差,因此 FOAD 算法很难应用于提高 PCA 的预测准确率.在图 5(b)的个别情况下,出现了 FOAD 引起 PCA 分类准确率降低的情况.总体上看,虽然选项约简操作造成部分信息的损失,但是合并模糊选项使选项分布更加清晰.此外,LDA 对各种参数敏感,尤其是对保留属性的个数.LDA 的预测准确率从保留最少属性时的提高 6.7%上升到保留最多属性时的提高 14.1%.由于线性回归不是一种典型属性提取算法,其性能起伏很大,其准确率提高值在 7.1%到 9.4%的范围内上下波动.

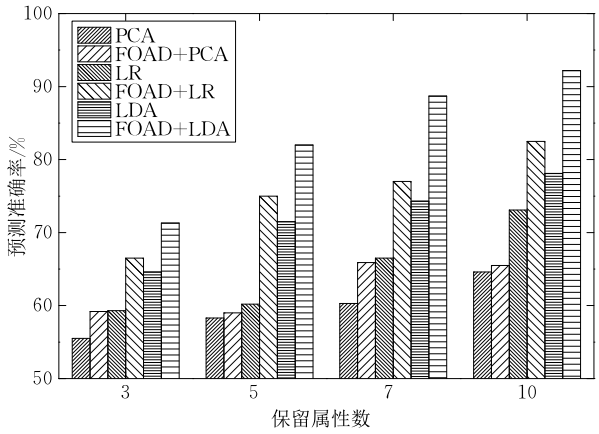


(a) PTSD数据集

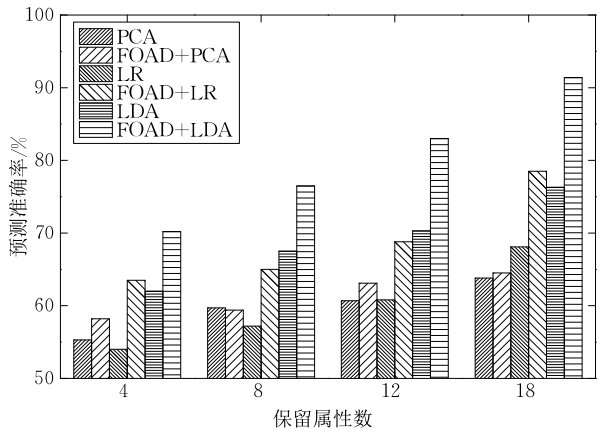


(b) 攻击行为数据集

图 4　三种属性提取算法在不同规模的训练集中的分类准确率



(a) PTSD数据集



(b) 攻击行为数据集

图 5　三种属性提取算法在保留不同数量属性时的分类准确率

5.5　属性约简算法对比

图 6 是基于 FOAD 方法与两种基于粗糙集的属性约减方法的性能对比，可以发现 FOAD 方法的预测准确率提高了 3.3%—4.7%. 心理测量数据中的模糊选项对预测准确率具有较大影响，因此基于模糊选项关系的 FOAD 方法比基于粗糙集的属性约减方法更能消除数据中的冗余关系.

5.6　属性模糊提取算法对比

图 7 展示了在保留 5 个属性的条件下，5.2 节中不同模糊降维方法的预测准确率的对比结果. 基于 FOAD 的 LDA 算法比其他属性模糊提取算法具有更好预测准确率. PCA 算法降维时保留了最多

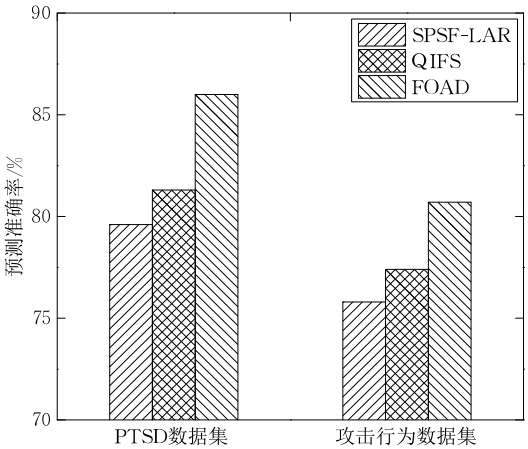


图 6　FOAD 方法与基于粗糙集的属性约简方法的对比

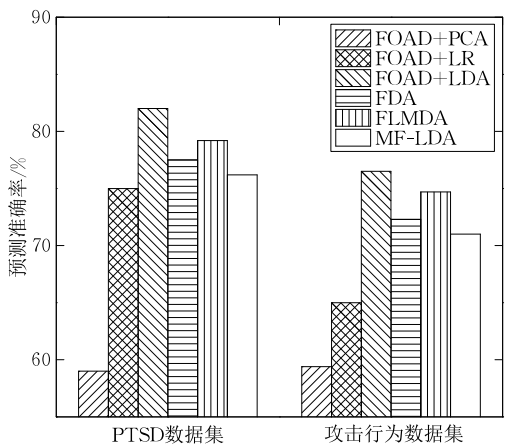


图 7　属性模糊提取算法的分类准确率比较

信息量,却使分类准确率降低,因而基于 FOAD 的 PCA 算法表现也最差. LDA(FOAD)算法与 PCA(FOAD)算法之间存在较大差距,分别为 23.7%(PTSD 数据集)和 17.5%(攻击行为数据集). 虽然线性回归(FOAD)比 LDA(FOAD)的准确率低 6.8%(PTSD 数据集)和 11.3%(攻击行为数据集),但是它可以对所有属性进行排序,而不仅局限于输出关键属性. 其他模糊降维算法虽然引入了模糊处理方法,但是仍然未考虑选项间的模糊关系,因此分类效果与 LDA(FOAD)算法相比较差.

属性个数和选项个数往往是有限的,比如 PHQ-13 量表中有 13 个属性和 5 个选项,因此在忽略这两个参数的情况下可以得到 FOAD 的时间复杂度为 $O(n^2)$. 进而可以得到这几种属性模糊提取算法的时间复杂度都是 $O(n^2)$. 图 8 分别比较了不同属性模糊提取算法在两个数据集中的时间消耗,

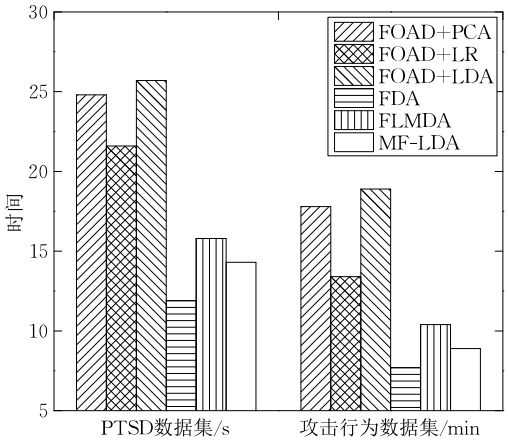
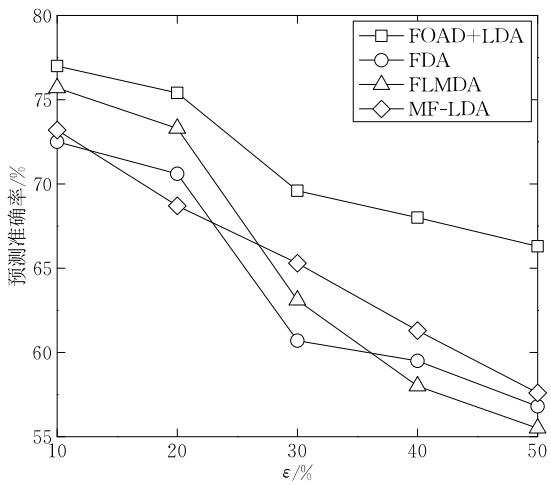


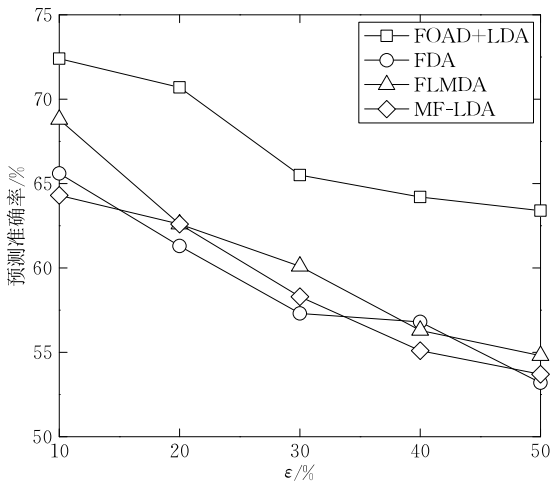
图 8 属性模糊提取算法的时间比较

其中 PTSD 数据集以秒作为时间单位,而攻击行为数据集的时间以分计算. 可以发现 FOAD 算法需要在其他属性提取算法的基础上消耗额外的时间,比如选项合并,因而其效率略低于其他算法. 线性回归(FOAD)的预测准确率和时间消耗都优于 PCA(FOAD),原因是 PCA 降维时保留了最多的信息量而使分类效果最差,同样 FLMDA 的预测准确率和时间消耗均优于 MF-LDA. 鉴于选项约简操作比较耗时,三种基于 FOAD 的模糊算法相比其他三种算法消耗时间更多,在两个数据集上需要多耗时至少 5.8 s 和 3.0 min.

为对比算法的稳定性,可以在测试数据中人为加入噪声数据. 首先生成 $m'\times s$ 的随机噪声数据,其中 m' 和 s 分别表示噪声样本数量和属性数量. $m'=m\varepsilon$,其中 ε 是噪声数据占原始数据的比例. 这些噪声数据都具有错误的决策属性,即错误的诊断结论. 图 9 是不同的属性模糊提取算法在两个数据集中的稳定性比较,可以看到经过 FOAD 预处理后的 LDA 算法的稳定性明显优于其他几种算法. LDA(FOAD)算法在 ε 较小时下降缓慢,随后在 20%～30%的区间内下降迅速,最后又趋于平稳. 在整个 10%～50%的区间内仅下降了约 5%的预测准确率. 即使存在大量的噪声($\varepsilon=50\%$),其预测准确率仍然较高. 这是因为 FOAD 方法通过降低选项的模糊性减少了错误决策,使 FOAD 具有一定的容错率,在包含一定比例的噪声数据时仍然具有良好的预测准确率. 随着 ε 的增大,其他几种算法的预测准确率持续快速下降,在噪声数据比例达到 50%时已无法正常工作.



(a) PTSD数据集



(b) 攻击行为数据集

图 9 属性模糊提取算法的稳定性比较

6 总　结

本文引入了选项熵和选项影响度的概念，用于描述选项间的关系与分布. 为去除心理测量数据中的冗余数据，本文提出了一种称为 FOAD 的基于模糊选项关系的关键属性提取方法. 对比实验的结果显示：该方法与其他典型算法具有相似的时间复杂度，但却能有效提高分类准确率.

心理疾病的诊断数据具有明显的模糊性，一般的统计分析方法往往不能得到需要的结果，利用模糊集和粗糙集方法对数据进行预处理可以消除这种数据噪声. 未来可以进一步将知识表示与模糊推理应用于心理疾病诊断，以获得更加精确的诊断结论.

参　考　文　献

[1] Sumathi M R, Poorna B. Prediction of mental health problems among children using machine learning techniques. International Journal of Advanced Computer Science & Applications, 2016, 7(1): 552-557

[2] Chen H L, Huang C C, Yu X G, et al. An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert Systems with Applications, 2013, 40(1): 263-271

[3] Son C S, Kim Y N, Kim H S, et al. Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches. Journal of Biomedical Informatics, 2012, 45(5): 999-1008

[4] Kessler R C, Aguilar-Gaxiola S, Alonso J, et al. The global burden of mental disorders: an update from the WHO World Mental Health (WMH) surveys. Epidemiologia e Psichiatria Sociale, 2009, 18(1): 23-33

[5] Zhang J, Zhu S, Du C, Zhang Y. Posttraumatic stress disorder and somatic symptoms among child and adolescent survivors following the Lushan earthquake in China: A six-month longitudinal study. Journal of Psychosomatic Research, 2015, 79(2): 100-106

[6] Masri R Y, Jani H M. Employing artificial intelligence techniques in Mental Health Diagnostic Expert System// Proceedings of the International Conference on Computer & Information Science. Kuala Lumpur, Malaysia, 2012: 495-499

[7] Rahman R M, Afroz F. Comparison of various classification techniques using different data mining tools for diabetes diagnosis. Journal of Software Engineering & Applications, 2013, 6(3): 85-97

[8] Seixas F L, Zadrozny B, Laks J, et al. A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. Computers in Biology & Medicine, 2014, 51C(7): 140-158

[9] Khemphila A, Boonjing V. Parkinsons disease classification using neural network and feature selection. World Academy of Science, Engineering and Technology, 2012, 64: 15-18

[10] Dabek F, Caban J J. A neural network based model for predicting psychological conditions. 2015, 9250: 252-261

[11] Lôpez J, Maldonado S. Group-penalized feature selection and robust twin SVM classification via second-order cone programming. Neurocomputing, 2017, 235: 112-121

[12] Pawlak Z, Skowron A. Rudiments of rough sets. Information Sciences, 2007, 177(1): 3-27

[13] Wang C Z, Qi Y, Shao M, et al. A fitting model for feature selection with fuzzy rough sets. IEEE Transactions on Fuzzy Systems, 2016, 25(4): 741-753

[14] Wang F, Liang J, Dang C. Attribute reduction for dynamic data sets. Applied Soft Computing, 2013, 13(1): 676-689

[15] Fan J, Jiang Y, Liu Y. Quick attribute reduction with generalized indiscernibility models. Information Sciences, 2017, s397: 15-36

[16] Zhao Z, Wang L, Liu H, Ye J. On similarity preserving feature selection. IEEE Transactions on Knowledge & Data Engineering, 2013, 25(3): 619-632

[17] Liu J, Lin Y, Lin M, et al. Feature selection based on quality of information. Neurocomputing, 2017, 225: 11-22

[18] Wu X H, Zhou J J. Fuzzy discriminant analysis with kernel methods. Pattern Recognition, 2006, 39(11): 2236-2239

[19] Xu J, Gu Z, Xie K. Fuzzy local mean discriminant analysis for dimensionality reduction. Neural Processing Letters, 2015, 44(3): 1-18

[20] Zhao M, Chow T W S, Zhang Z, Random walk-based fuzzy linear discriminant analysis for dimensionality reduction. Soft Computing, 2012, 16(8): 1393-1409

[21] Zhang J, Zhang Y, Du C, et al. Prevalence and risk factors of posttraumatic stress disorder among teachers 3 months after the Lushan earthquake: A cross-sectional study. Medicine, 2016, 95(29): e4298

[22] Lee S, Ma Y L, Tsang A. Psychometric properties of the Chinese 15-item patient health questionnaire in the general population of Hong Kong. Journal of Psychosomatic Research, 2011, 71(2): 69-73

[23] Qu Y, Jiang H, Zhang N, et al. Prevalence of mental disorders in 6 — 16-year-old students in Sichuan province, China. International Journal of Environmental Research & Public Health, 2015, 12(5): 5090-5107

[24] Candes E, Li X, Ma Y, Wright J. Robust principal component analysis? Journal of the ACM, 2009, 58(3): 11

[25] Sharma A, Paliwal K K. A deterministic approach to regularized linear discriminant analysis. Neurocomputing, 2015, 151(1): 207-214

XIONG Xi, born in 1983, Ph. D., lecturer. His current research interests include data mining, machine learning and social computing.

QIAO Shao-Jie, born in 1981, Ph. D., professor. His current research interests include big data, moving objects databases and complex networks.

HAN Nan, born in 1984, Ph. D., lecturer. Her current research interests focus on data mining.

YUAN Chang-An, born in 1964, Ph. D., professor. His current research interests focus on data mining.

ZHANG Hai-Qing, born in 1986, Ph. D., lecturer. Her current research interests include data mining and decision-making systems.

LI Bin-Yong, born in 1982, Ph. D., lecturer. His current research interests include big data and cloud computing.

Background

This work is a part of the "Research on Influence Factors and Propagation Mechanism of Users' Emotion in Social Networks", which is mainly supported by the National Natural Science Foundation of China under Grant Nos. 61772091 and 61802035 and the Youth Foundation for Humanities and Social Sciences of Ministry of Education of China under Grant No. 17YJCZH202. These projects focus on the analysis and modeling of emotion in social media. The current research of sentimental analysis mainly utilizes the method of natural language processing, which does not consider the influence of various factors on users' complicated emotion generated in their real lives, which is more trustful and helpful to this study.

The offline emotion data can be collected by psychometric tests. However, the intrinsic fuzziness of clinical psychology and the psychometric data has hindered the application of artificial intelligence in the analysis of users' emotion. Currently, the information extraction method on psychometric data with fuzziness has not been extensively studied. This research focuses on extracting valuable information and obtaining key attributes from fuzzy psychometric data, which can facilitate the analysis of users' emotion. This paper proposes a fuzzy-option based attribute discriminant method. As the core of the whole approach, the fuzzy option reduction algorithm can merge fuzzy options into other reserved options in order to reduce the fuzziness of psychometric data. The experimental results on real clinical data sets demonstrate that the prediction accuracy can be improved by 3.3%—14.1% without increasing the computational complexity.