Journal of Wuhan Polytechnic

面向客户行为分析的特征提取算法对比研究

赵珩君[1],张景韶[1],肖　进[2]

(1.四川广播电视大学 经济管理学院,四川 成都 610073;
2.四川大学 商学院,四川 成都 610064)

摘　要: 客户特征提取是整个客户行为分析过程中的重要环节。由于客户特征提取时获得的数据具有多共同特征及大噪声等特点,使得在客户行为分析中进行客户特征提取存在较大误差。采用 UCI 机器学习数据库中有多个共同特征的数据集分别对典型特征提取算法进行实验对比及分类规则提取结果分析,验证了 FC-GMDH 算法在特征提取精度和抗干扰方面具有明显的优势,在客户行为分析时取得满意的特征提取效果。

关键词: 客户行为;特征提取;噪声特征;对比分析

中图分类号: F274　　**文献标识码:** A　　**文章编号:** 1671-931X(2016)05-0032-06

一、文献回顾

用于客户行为分析的特征提取是在客户细分后的目标客户群中发现并确定能够区分客户特征的过程。特征提取是客户识别的重要内容,能够帮助企业制定有效的营销策略。特征提取一般的定义为[1]:已知一个特征集,从中选择一个子集使评价判据最优。特征提取的基本任务是从大量基本特征中去掉不相关以及冗余的特征获取对目标函数有利的特征[2]。客户特征提取是整个客户行为分析过程中的一个重要环节。特征提取算法可以去除目标客户数据中的冗余特征、无关特征甚至噪声特征,从而得到一个无冗余、无噪声的样本集,有助于提高客户的识别率以及数据的挖掘速度。现有的特征提取方法在定性数据及噪声数据的处理上存在局限性,而定性数据及带噪声数据在客户行为分析建模过程中是不可避免的[3]。

目前用于高维客户数据规则提取的机器学习分类算法有很多,但通过文献分析发现,这些算法仍然存在很多问题,例如在处理定性数据上存在局限性并且普遍对噪声敏感,这就限制了算法的应用范围[4]。Kira 和 Rendell 提出的基于距离的 Filter 方法提高了计算速度,但其中的距离指标只适用于定量数据[5];Relief 系列算法是公认的分类效果较好的 filter 式特征提取算法[6],能够处理离散和连续的数据,但该算法不能辨别冗余特征。C4.5 决策树算法简洁、高效、易于理解、能够处理大型数据库连接和多种数据类型,是一个在研究分类算法时常用的参考基准[7],但随着数据的重复划分,该算法会产生过多的无统计意义的分支,使得分类过拟合,最终降低分类精度[8]。而 Kalousis[9]和 Riyaz Sikora[10]等人通过模拟实验证明了大多数特征选择方法对数据噪声比较敏感,难以保证得到最优特征。要得到较好的特征提取效果,要充分考虑样本的选择与转换,离散化和噪声干扰等问题[11]。然而,企业客户信息来自不同信息源这就

收稿日期: 2016-07-14

基金项目: 国家自然科学基金委面上项目"大数据环境下基于 GMDH 的客户分类半监督集成模型研究"(项目编号:71471124);四川省青年基金"大数据环境下客户价值区分半监督集成模型研究"(项目编号:2015RZ0056);四川省社科规划项目"类别不平衡环境下客户流失预测半监督集成模型研究"(项目编号:SC14C019)。

作者简介: 赵珩君(1978-),女,博士,四川广播电视大学经济管理学院副教授,研究方向:商务智能数据挖掘及建模、客户关系管理等。

决定了数据的异构性，表现为既包含定性数据，又包含定量数据，并且数据中多带噪声的高维数据，这就为客户特征的提取带来了难度。因此，数据类型和对噪声的容忍能力是衡量特征选择方法优劣首先应该考虑的问题。

二、算法的检验及评价

检验可以说不算特征提取的一部分，然而它在解决实际问题中是必不可少的。检验通常用选定的特征子集对人工的或实际的数据集进行训练和预测，将训练和预测的结果进行分类精度、特征约简等指标的比较，从而验证算法的有效性。

一个性能优良的特征提取模型的设计应追求达到以下三个最佳目标：(1)最少的模糊规则数量；(2)最短的模糊规则长度，即规则的前件要尽量短；(3)具有最高的识别率。

自检验(Resubstitution test)和交叉检验(Cross-validation test)是检验特征提取算法性能的常用方法。自检验可以用来验证模型的自相容能力，该方法在验证过程中使用同一个样本作为训练集和测试集，因此自检测方法受“记忆影响”，不足以说明模型的性能。交叉检验可以用来验证模型的泛化能力。抽样子集分析、独立测试集分析和 Jackknife 检验是三种最常见的交叉检验方法[12]。抽样子集分析将一个数据集分为训练集和测试集，由于数据集的划分具有任意性，除非测试集的规模足够大，否则得到的结果就不能真正用来评价一种模型的优劣。Jackknife 检验是一种可以消除“记忆影响”的检验方法。假设数据集共有 N 个样本，将他们逐个取出作为测试集，其余 N-1 个样本作为训练集。显然这种方法的缺点是时间开销太大。

本文进行算法检验所采用的方法是在实际应用中常采用 Jackknife 检验方法的变形 K-重交叉检验(K-fold cross-validation)，即首先将样本数据随机分为数目大体相等的 K 个子集，依次取出一个子集作为测试集，而其余 K-1 个子集作为训练集，交替进行 K 次后，将各次的准确度求平均值，一般 K 的取值为 5 到 10 间的整数。同时，综合考虑模糊规则提取的三个目标：最少的模糊规则数量，最短的模糊规则长度以及最高的分类精度进行算法的综合评价。

三、共同特征的影响

首先对比当各个样本具有的多个共同特征对特征提取算法的影响。由于细分就是将具有共同特征的客户聚为一类，这就使客户群内的样本具有很多共同特征。但从另外一个角度考虑，这种情况就使得群内的少数特征淹没在共同特征的“信息海洋中”，从而最终使得规则提取算法的精度降低。

(一)研究数据

实验采用的 2 个样本集数据来自 UCI 数据库。该数据集库是由 California 大学 Irvine 分校建立的一个用于各种机器学习算法和数据挖掘算法的基准数据实验平台。为了减少不必要的影响因素，本节实验中均选用无缺失的数据样本集进行算法对比分析。

1.数据集 Wine

实验数据 wine 来自 UCI 机器学习数据库。数据集 N 包含 178 组数据，13 个属性，分别代表来自意大利同一地区 3 种类型植物酿造的酒。其中，第一类 59 个样本，第二类 71 个样本，第三类 48 个样本。表 1 所示为 13 个属性，其中 X_1 为酒的类别标识(取值范围 1~3)作为决策变量，其余 12 个属性作为输入变量，既包含定性数据如 X_{10}，X_{11}，又包含定量数据如 X3，X5 等。

为了对比共同特征较多时，FC-GMDH 与 FRI 算法的精度，用随机发生器分别产生 178 个样本包含 3 个属性(X_{14}~X_{16})的共同属性样本集 W1 和178 个样本包含 5 个属性(X_{14}~X_{18})的共同属性样本集 W_2。其中，要求所产生的 W_1 和 W_2 中的属性是均值为 0，方差为 1 且符合高斯分布的随机特征。在各个

表 1 wine 数据集变量

变量	含义	变量	含义
X_1	Alcohol	X_2	Malic acid
X_3	Ash	X_4	Alcalinity of ash
X_5	Magnesium	X_6	Total phenols
X_7	Flavanoids	X_8	Nonflavanoid phenols
X_9	Proanthocyanins	X_{10}	Color intensity
X_{11}	Hue	X_{12}	OD 280/OD315 of diluted wines
X_{13}	Proline		

表 2 用于比较的 UCI 数据集的性质

序号	数据集	样本个数	特征个数	类别个数	有无缺失	特征类型
1	Vehicle silhouettes	946	18	4	N	定量
2	Acute Inflammations	120	6	4	N	定性、定量
3	Image Segmentation	2310	19	7	N	定量
4	Zoo	101	16	7	N	定性、定量
5	Ionosphere	351	34	2	N	定量
6	Blood Transfusion	748	4	2	N	定量

样本中的取值均相同才能作为三类酒的共同特征加入样本。此时数据样本的属性分别被扩展到16 维和18 维。

2.多属性 UCI 数据集

为了全面测试算法的性能，所以接下来利用 FC-GMDH 算法在 UCI 数据样本集的分类问题上做了测试。根据问题的类型，样本的大小等因素选择以下六个数据集，如表 2 所示。

Vehicle silhouettes 是由英国格拉斯哥大学统计学系 Alistair Sutherland 提供的车辆外形轮廓辨别数据库。包括 946 个样本，根据 18 个外观轮廓参数，将样本分类 4 类，分别表示：欧宝汽车、绅宝汽车、公共汽车和轻便旅行车。

Acute Inflammations 数据集是由波兰国家科学院系统研究所提供的膀胱炎和肾炎诊断数据库。包含 120 个样本，6 个属性，其中有 31 例诊断为肾炎患者，40 例为膀胱炎患者，非膀胱炎和肾炎的 30 例，同时患有膀胱炎和肾炎的为 19 例。目标任务是根据这 6 个指标来判断样本属于哪一类。

Image Segmentation 数据集是马萨诸塞大学视觉研究小组提供的图像分割数据集，包括 2310 个图像样本，由 7 类图像集组成，其样本数均为 330。

Zoo 是由 Richard Forsyth 提供的动物特征数据库，包括 101 个样本，根据 16 个特征属性，将样本分为 7 个类别，分别对应 7 种不同的动物。

Ionosphere 是由约翰霍普金斯大学应用物理实验室提供的根据电离层反射的雷达回波参数来判定雷达性能好坏的数据集。包括 351 个样本，依据 34 个回波参数判定，结果有 225 个样本属于性能优良的雷达，126 个属于劣质雷达。

Blood Transfusion 数据集是由台湾血液中心提供的献血者献血频率统计数据，包含 748 个样本，其中 178 位是主动献血者，570 位是被动献血者。通过 4 个属性变量来判断献血者的类型。

样本是机器学习的数据基础。训练样本集是否具有代表性，决定了机器学习的效果[13]。机器学习时，一般将样本划分为训练集和测试集。训练集的代表性越强，所得模型的质量越好。当总样本确定时，训练集的质量则取决于样本的抽取方法[14]。当训练样本数量达到一定程度时，它对识别率的影响不大。但当样本数量较少时，训练集的抽取就很难保证。本节实验采用的数据样本个数不等，因此，采取了最常用的随机样本分割方法和 10 重交叉法进行实验。随机样本分割法（Randomly splitting sample，RSS）操作简单，易于理解，随机性强，可以保证实验中样本的代表性。

（二）方案与结果分析

上一节中介绍了目前特征提取算法中广泛采用的标准数据库的来源和构成。为保证本文的实验结果和算法对比分析的科学性与公正性，本节针对以上标准数据集进行研究和分析。研究方案的设计及分析如下：

表 3 特征提取结果对比（共同特征比率约 20%）

数据集	共同特征数及比率（%）	FC-GMDH 分类正确率（%）	FRI 分类正确率（%）
训练集 A_1	3（20%）	95.36%	79.21%
训练集 B_1	3（20%）	93.25%	62.46%

表 4 特征提取结果对比（共同特征比率约 30%）

数据集	共同特征数及比率（%）	FC-GMDH 分类正确率（%）	FRI 分类正确率（%）
训练集 A_1	5（30%）	98.49%	69.34%
训练集 B_1	5（30%）	96.26%	59.29%

1.实验一：Wine 数据集规则提取

由于 FC-GMDH 算法与 FRI 算法都是基于自组织数据挖掘（SODM）理论的特征提取算法，因此，有必要将两者在数据集中包含较多共同特征时的特征提取精度做对比。实验方案设计如下：

采用 10 重交叉法[15]对样本进行规则提取的验证。设总样本集为 T，将添加了共同特征的 178 个样本近似平均分为 10 份，取其中的一份为测试样本集 TS，剩余的九份为训练样本集 TR，其中 $T=TR\cup TS$，$TR\cap TS=\varnothing$，10 次轮换交替的分类精度作为算法的检验指标进行比较。实验结果为总体分类精度，取 10 次实验的平均值。

系统输入分明变量为 X_2~X_{16} 及 X_2~X_{18}，输出均为 X_1，用分类正确率来衡量算法特征提取的精度。对于具有 3 个共同特征和 5 个共同特征（共同特征分别约占总特征的 20%和 30%）的样本数据，实验结果对比见表 3 和表 4。

当共同特征占总特征比率 20%时，在 $N\cup W_1$ 上采用 FC-GMDH 进行规则提取，样本训练集与检验集上的平均分类正确率为 95.36%和 93.25%，高于采用 FRI 方法时的正确率。并且当共同特征数增加时（表 4），FRI 的分类正确率明显下降，而 FC-GMDH 算法的分类正确率却略有提高。可见 FRI 方法不适于做共同特征较多时的规则提取。由于 FC-GMDH 算法在特征提取前首先对样本进行了模糊特征子集的划分，使得样本子集间的差异明显且所对应的隶属函数更加准确，从而可以获得精度较高的分类结果。通过数值仿真实验的比较，FC-GMDH 算法在群间共同特征较多时，能够达到理想的分类正确率，符合该算法的理论分析。

2.实验二：多值分类数据集特征提取比较

为了使实验结果更加全面、公正，我们将特征提取算法 FC-GMDH 与有代表性的传统特征提取算法 Relief、C4.5 决策树、朴素贝叶斯分类器（NBC）在 UCI 数据集上的规则提取结果进行了对比。这些数

表 5 特征提取精度对比(%)

数据集	Relief			C4.5			NBC			FC-GMDH		
	①	②	③	①	②	③	①	②	③	①	②	③
1				84.95	84.68	81.28	85.26	83.45	82.64	89.54	91.05	93.45
2		—		84.56	83.12	79.48	88.15	92.69	90.47	89.65	91.56	92.64
3				77.59	76.25	75.12	76.54	71.45	72.38	81.56	82.15	82.65
4				79.18	72.35	71.54	79.45	76.84	72.54	84.65	85.94	86.15
5	73.26	69.65	74.28	68.15	66.25	60.28	72.65	65.35	61.25	85.96	84.18	83.65
6	84.59	80.42	76.25	81.26	80.87	79.45	83.26	85.14	84.99	84.95	86.42	85.98

注:①共同特征比率 5%;②共同特征比率 10%;③共同特征比率 20%。

据集的特征维数从数个到数十个不等；数据类型不同,即包含离散型或连续型的单一类型数据,又包含了混合型定性、定量数据样本集；类别数也各不相同。以上这些数据集广泛的代表了实际无噪声下的分类问题数据，可以较好地验证特征提取算法在实际数据集上的性能。

仍然采用 10 重交叉法进行实验,10 次轮换交替的分类精度作为各个算法的检验指标进行比较。实验结果为总体分类精度，取 10 次实验的平均值。首先，为了比较在相同的数据环境下的特征提取精度,逐步增加(5%,10%,20%)共同特征的比例。实验结果见表 5。

由于 Relief 一般只能用于二值类别数据的规则提取，所以表 5 中无法给出算法在数据集 1-4 上的对应结果。从表 5 可以看出,随着共同特征比率的增加,大多数算法所对应的分类精度都有所下降。具体来看,在特征维度较低的样本二和样本六上,由于特征维度的相关性较低,NBC 算法的精度值接近同比率下 FC-GMDH 算法精度值,近似于最优分类。在高维大数据样本 Image Segmentation 和 Ionosphere 上,FC-GMDH 算法的分类精度仍然维持在较高水平,而 C4.5 算法由于自身在特征提取过程中需要对数据不断进行划分,从而造成了碎片数据的过拟合,使得在这些样本上的分类精度始终最小，远远低于其他特征提取算法。特别值得一提的是,大多数算法在数据集样本中共同特征增加时，分类精度都有所下降，但我们提出的 FC-GMDH 算法由于在特征提取前预先进行了特征子集的划分，使得模糊特征建模中采用的隶属函数更加精确，所以该算法的分类精度不会像其他算法那样随共同的增多而降低。相反,在我们选择的六个数据集中，前四个二值类别数据样本的特征提取精度反而在共同特征比率增大时也同时增大,达到理想分类精度。总体来看,FC-GMDH 算法在样本共同特征增加时分类精度较传统算法最优,验证了新算法在“信息海洋”中强大的特征提取能力,验证了我们的算法设计是有效的。

四、抗干扰性评价

(一)研究数据及方案

为了说明噪声数据对特征提取算法结果的影响,本节采用 UCI 数据库的真实数据样本,对比特征提取算法的分类规则提取结果。实验数据样本包括:

1.人造噪声数据集

实验选用两个人造噪声数据集。Art 数据集采用 Isabelle Guyon 的程序产生的 Art 类数据[16],该程序源代码可以从 NIPS2001 的特征选择研讨主页上下载。所采用的数据集名为 Art(a1,p,a2,a3,a4)格式,其中 a1 表示用以产生类别属性的特征个数,属于分类必需的属性;p 表示全部样本个数;a2 表示独立特征的个数;a3 表示无关特征的个数;a4 表示冗余特征的个数。我们用该数据集考察算法去冗余特征的性能。

用于实验的第二个数据集是在上一节 Wine 数据集基础上通过人工添加噪声形成的。添加噪声的方法如下：从每个类别的样本中随机抽出 20%的数据，再把这些数据加入到另外的任意类别的样本集中,作为噪声数据。在实验中,采用 10 重交叉检验法,对比 FC-GMDH 与 FRI 算法的抗干扰性。使用分类准确率的平均值、绝对误差总和(Summarized Absolute Error,SAE)、平均绝对百分比误差(Mean Absolute Percentage Error,MAPE)、近似方差(Approximation Error Variance,AEV)就算法的抗干扰性能进行对比。

2.有噪声数据集

在 UCI 数据集中，我们选取了三个已知噪声水平的数据集，用来测试 FC-GMDH 算法在真实噪声数据上的性能,如表 6 所示。

表 6 UCI 数据集的基本信息

数据集	样本数	特征数	类别数	噪声水平(%)	变量类型
Credit Approval	690	15	2	5%	定性、定量
Mammographic Mass	961	5	2	16.85%	定量
Horse Colic	368	26	4	30%	定性、定量

三个样本数据集中包含的噪声水平分别为5%，16.85%，30%。文献分析表明了传统特征提取算法Relief、C4.5和NBC朴素贝叶斯各自的优缺点，在实验中我们将FC-GMDH算法与他们做对比，验证该算法在不同的噪声水平下抗干扰性的优劣。

（二）结果分析

1.实验一：人工噪声数据下抗干扰性对比

由于FC-GMDH算法与FRI算法都是基于自组织数据挖掘（GMDH）理论的特征提取算法，因此，有必要将两者在噪声数据集下的特征提取精度做对比。

实验结果如表7和表8所示。

对于人造的噪声数据样本，产生了以下的最优模糊规则，见表8。通过对比表7的结果，我们发现本文提出的FC-GMDH算法对带有噪声的样本数据进行规则归纳，产生规则的绝对误差总和，平均绝对百分比误差以及近似方差的平均值均小于用FRI方法时的各项平均值，同时FC-GMDH算法仅利用3条模糊分类规则就达到了95.63%的分类精度。可见，在人造噪声数据样本下，新算法的抗干扰性更强，能够获得较高的规则提取精度。

表7 抗干扰性比较

算法及规则		SAE	MAPE	AEV	正确分类率（%）
FC-CMDH	R_1	1.23	6.21%	0.328	95.63%
	R_2	2.39	5.46%	0.432	
	R_3	1.96	4.36%	0.425	
FRI	R_1	4.12	7.35%	0.536	81.27%
	R_2	3.26	9.25%	0.582	
	R_3	2.41	13.14%	0.612	
	R_4	2.19	11.02%	0.496	
	R_5	1.98	6.98%	0.533	

接下来，我们在人造噪声数据样本Art数据集上，对比算法FC-GMDH和Relief的去冗余性（表9）。

表8 规则提取

Rules in feature selection by FC-GMDH algorithm	
R_1	If x1 is high and x7 is middle an x9 is middle and x13 is high then output is class1
R_2	If x1 is high and x3 is low and x7 is low and x10 is low and x13 is middle then output is class2
R_3	If x1 is high and x6 is low and x10 is high and x12 is low then output is class3
Rules in feature selection by FRI algorithm	
R_1	If x1 is high and x7 is middle an x13 is high then output is class1
R_2	If x1 is low and x7 is low and x8 is high and x13 is high then output is class2
R_3	If x1 is high and x6 is low and x10 is high and x11 is low then output is class3
R_4	If x1 is high and x3 is low and x7 is low and x12 is middle then output is class2
R_5	If x1 is middle and x3 is high and x9 is high and x10 is low and x13 is high then output is class1

表9 规则提取精度及去冗余性

样本集	Relief		FC-GMDH	
	分类精度（%）	冗余特征个数	分类精度（%）	冗余特征个数
Art(3,300,5,3,3)	82.60	1	89.26	0
Art(3,300,5,3,6)	81.69	3	88.24	0
Art(6,1000,10,3,10)	76.25	6	84.15	1
Art(6,1000,20,3,20)	63.59	6	83.59	2
Art(12,1000,24,3,30)	59.15	18	86.14	1

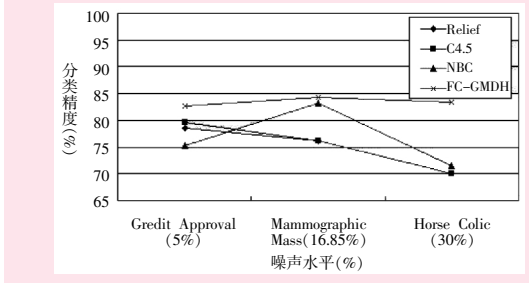

图1 不同算法在UCI数据集上的分类精度

分析表9的结果，随着样本冗余特征的增多，Relief的分类精度明显下降，而FC-GMDH算法的分类精度仍然维持在一个较高的水平，平均分类精度在各个样本上都保持在80%以上。从特征提取后的特征中，我们发现，FC-GMDH算法能剔除绝大部分冗余特征，在前两个样本上的提取结果，可以达到100%的剔除冗余特征。可见，通过模糊划分与自组织数据挖掘算法相结合，FC-GMDH算法在保证较高分类精度的同时，还能有效剔除冗余特征。

2.实验二：真实噪声数据下抗干扰性对比

在不同噪声水平的真实数据样本集上，特征提取算法的实验结果见图1。

从图1所示曲线中看出，由于Relief算法一般只用于二元决策，所以在Horse Colic数据集上没有给出对应的规则提取精度。对比三个不同的噪声水平数据样本下的分类精度，FC-GMDH的分类精度基本保持在85%左右。而传统算法的分类精度，随着噪声水平的增大，其对应分类精度均有所下降。其中，由于C4.5算法在特征提取过程中随着数据的重复划分会产生过多的无统计意义的分支，从而分类

过拟合，导致其在几种算法中精度最低。对于 NBC 算法，只要满足了属性独立性假设成立或相关性较小的时候就可以获得近似最优的分类效果。因此当噪声水平增大时，在数据集 Mammographic Mass 上的分类精度反而提高，接近 FC-GMDH 的算法精度水平。由于 Mammographic Mass 数据集较之其他两个数据集所包含的特征数少，仅有 5 个，降低了属性相关的可能性。而其他两个数据集众多的特征中难免有不少相关的属性，从而导致该算法在这两个数据集上的精度较低。

总体来看，FC-GMDH 算法在真实数据集上的抗噪声干扰能力最优，验证了算法的理论分析与建模的期望一致。

五、结论

模糊特征提取是客户行为分析研究的一个热点。有多种算法用来特征提取，其中 Relief 算法能够处理离散和连续的数据，但它对冗余属性的去除大多无能为力；神经网络算法通过结果比较来选择特征，计算量大，可解释性差。FRI 算法是近些年来提出的一种基于模糊规则归纳的特征提取，该算法在理论和实验上证明较之上两种算法表现稍好。但是 FRI 的建立虽然基于模糊规则归纳原理，却忽略了对隶属函数的准确描述，使得在样本噪声较大时其分类精度较低。所以，本文采用了一种基于模糊划分的特征提取算法 FC-GMDH，通过模糊划分形成的特征子集可以提高确定隶属函数的准确性，并且有效去除噪声特征的干扰，从而提高划分后样本特征子集的差异度，提高模型的分类精度。FC-GMDH 在 UCI 分类数据集的实验的对比，证明达到了预期效果，在精度上也较之传统方法有所提高。实验证明了 FC-GMDH 算法在特征提取精度和抗干扰方面具有明显的优势。

参考文献：

[1] Langley P.Selection of relevant features in machine learning [M].AAAI Fall Symposium on Relevance,1994:140-144.

[2] Blum A L,LangelyP. Selection of relevant features and examples in machine learning [J].Artificial Intelligence, 1997,(12):245-271.

[3] 柯赟.新浪微博信息传播的影响因素分析与效果预测[J].现代情报,2016,(3):22-26.

[4] 汤铃,李建平,孙晓蕾,等.基于模态分解的国家风险多尺度特征分析[J].管理评论,2012,(08).

[5] Almullim H,Dietterich T.Learning with many irrelevant features[A].Proceedings of Ninth National Conference on Artificial Intelligence [C].New Orleans:AAAI Press, 1991:547-552.

[6] Kononenko I.Estimation Attributes:Analysis and extensions of RELIEF [A].In Proceedings of the 1994 Europen Conference on Machine Learning [C].Catania:Springer Verlag,1994:171-182.

[7] 张志宏,寇纪淞,陈富赞,等.基于遗传算法的顾客购买行为特征提取[J].模式识别与人工智能,2010,(02).

[8] Ruggieri S.Efficient C4.5 [J].IEEE Transactions On Knowledge and Data Engineering,2002,14(2):438-444.

[9] Erhard R.,Hong H.D..Data cleaning:problem and current approaches [J].IEEE Data Engineering Bulletin, 2000,23(4):3-13.

[10]Riyaz Sikora,Selwyn Piramuthu.Framework for efficient feature selection in genectic altorithm based data minging [J].European Journal of Operational Research,2007, (180):723-737.

[11] 吉顺权，周毅. 产品用户评论在企业竞争情报中的应用——基于产品特征的关联规则数据挖掘[J].现代情报, 2015,(6):114-121.

[12]Chou K C,Liu W,Maggiora G M,Zhang C T. Prediction and classification of domain structural classes [J].Proteins:Structure,Function,and Bioinformatics,1998, (31):97-103.

[13]Lindenbaum M,Markovitch S,Rusakov D.Selective sampling for nearest neighbor classifiers [J].Machine Learning,2004,54(2):125-152.

[14]Swift D K,Dagli C H.A strdy on the network traffic of connection by boeing:Modeling with artificial neural networks[J].Engineering Applications of Artificial Intellitence, 2008,21(8):1113-1129.

[15]Xiaoming Huo,Seoung Bum Kim,Kwok-Leung Tsui, Shuchun Wang.FRP:A frontier-based tree-pruning algorithm [J].INFORMS Journal on Computing,2006,18 (4): 494-505.

[16]NIPS 2001 workshop on Variable and Feature Selecton [EB/OL].http://clopinet.com/isabelle/,2009-03-21.

[责任编辑：张　磊]

（下转第 42 页）

（上接第 37 页）

Comparing Study on Feature Extraction Algorithms for Customer Behavior Analysis

ZHAO Heng-jun[1] ZHANG Jing-shao[1] XIAO Jin[2]
（1.Sichuan Radio & Television University, Chengdu 610073, China;
2.School of Business, Sichuan University, Chengdu 610064, China）

Abstract: Customer feature extraction is an important part of the customer analysis. Because the data obtained from the customer feature extraction has many characteristics such as common features and large noise, so that there is a large error in the customer analysis of customer features extraction. The feature extraction and classification rule extraction experiment were done by using the UCI machine learning database respectively, and the experiments verified that the FC-GMDH algorithm has obvious advantages in feature extraction accuracy and anti-interference.

Key words: customer behavior; feature extraction; noise characteristic; comparative analysis