基于 GA-CFS 属性选择的个人信用评估模型①

柳亚琴，石洪波

(山西财经大学 信息管理学院，太原 030031)

摘 要：属性选择可以有效减少数据的冗余度和降低数据的维度，将 GA-CFS 属性选择方法引入个人信用评估中，利用 CFS 评价得到的启发式“价值”作为 GA 的适应度函数来对个人信用指标体系优化，建立了基于 GA-CFS 属性选择的个人信用评估模型。在 Australian 数据集上比较了 ID3、NB、Logistic、SMO 与 GA-CFS 属性选择方法和这四种分类算法分别结合执行的结果。实验结果表明，基于 GA-CFS 属性选择的个人信用评估模型降低了个人信用指标的维度，减少了学习所需的数据量，而且比基于单分类器的个人信用评估模型具有更高的分类准确率。

关键词：个人信用评估；属性选择；遗传算法；CFS；10 次 10 重交叉验证

Personal Credit Evaluation Model Based on Attribution Selection of GA-CFS

LIU Ya-Qin, SHI Hong-Bo

(Department of Information Management, Shanxi University of Finance & Economics, Taiyuan 030031, China)

Abstract: Attribution selection method could reduce data redundancy and the data dimension degree effectively and efficiently. This paper applies attribution selection of GA-CFS method to personal credit evaluation, and uses by the heuristic "merit" as GA fitness function to optimize personal credit index system through constructing a personal credit evaluation model based on attribution selection of GA-CFS. In addition, we compares with ID3, NB, Logistic, SMO and combination of GA-CFS attribute selection methods and the four classification algorithms in Australian data sets. Experiment results show that this model not only reduces the dimension of personal credit index and the amount of training data but also has higher classification accuracy than the personal credit evaluation model based on single classifier.

Keywords: personal credit evaluation; attribution selection; GA; CFS; 10 times 10-fold cross validation

近年来，我国个人消费信贷业务得到了快速发展，住房贷款、汽车按揭、教育贷款、信用卡等各种个人消费贷款的规模迅速扩大。据中国人民银行统计数据显示 2009 年的居民消费贷款总额达 55333.65 亿元，同比增长 18123.36 亿元，增长率为 48.7%。目前，我国商业银行个人信贷业务的主要矛盾表现为迅速发展的个人信贷规模与风险管理水平较低之间的矛盾，该矛盾严重制约着我国个人信贷业务的进一步发展。建立一套有效的个人信用评估系统将有利于商业银行扩大信贷规模，规避信贷风险，确保资金安全。

个人信用评估方法大体分为统计方法和非统计方法两大类，统计方法包括决策论方法、逻辑回归、线性回归、非线性回归、K 近邻估计等，非统计方法包括线性规划、整数规划、神经网络、分类树、专家系统等[1]。个人信用评估本质上就是模式识别中的一类分类问题，根据客户的若干信用历史资料，建立个人信用评估模型，利用这些规则将客户分为两类或多类。目前，已经有许多将分类学习算法应用到个人信用评估领域的文献，其中，文献[2]利用贝叶斯网络建立了一种个人信用评估模型；文献[3]利用 Logistic-RBF 组

① 基金项目:国家自然科学基金(60873100);山西省自然科学基金(2009011017-4)
收稿时间:2010-09-08;收到修改稿时间:2010-10-15

© 中国科学院软件研究所 http://www.c-s-a.org.cn

合模型对个人信用进行了评估；文献[4]利用决策树算法对个人住房贷款信用风险进行了评估；文献[5]利用支持向量机理论建立了一个新的个人信用评估预测方法。

从以上的工作中可以看出，前人主要着重于建立个人信用评估模型，对于指标体系的选取大多是将评估机构的指标体系直接使用或只是进行简单的筛选，这样得到的指标体系中含有冗余数据或无关数据，使得评估模型效率及准确率得不到保证。属性选择可以有效减少数据的冗余度和降低数据的维度，因此，在各种信用评估技术应用到模型之前，属性选择问题是一个关键性并且具有挑战性的问题。文献[6]提出了一种有效的属性选择方法：在用遗传算法（Genetic Algorithm，GA）搜索的同时利用相关性（Correlation based Feature Selection，CFS）进行评价，从而得到最优属性子集。基于以上考虑，本文将基于 GA-CFS 属性选择的分类器应用于个人信用评估，采用 GA 和 CFS 评价方法相结合进行属性选择，在得到的最优属性子集上再利用分类器进行分类。该方法降低了个人信用指标的维度，减少了学习所需的数据量，具有泛化能力强、可以找到全局最优解等优点，实验结果表明，以该方法为基础的个人信用评估系统提高了分类的准确率，具有广阔的应用前景。

1 个人信用评估

个人信用评估(Personal Credit Evaluation)主要依据客户的信用历史资料以及个人、家庭的内外客观条件，对客户信用等级进行评估分类，再根据分类结果为授信者（银行）的信贷决策提供科学依据。个人信用评估的完整过程应包括：数据收集和清理、个人信用评估指标体系的建立、个人信用评估模型的确定和评估模型的应用及评价。

进行个人信用评估的首要工作是建立个人信用信息数据库，主要采集和保存个人的基本信息以及个人在商业银行的借还款、信用卡、担保等信用信息，为个人信用评估提供原始数据。这些原始数据往往由于人为错误或数据收集过程中的漏洞，导致数据不合格，因此还需要进行数据清理，对那些不符合要求、空缺值多、有误差的数据进行清理或修正，最终得到比较整齐的、干净的，具有标准格式的可以用作数据处理和模型开发的基础数据。

个人信用评估指标体系是将个人信用信息数据库中的字段转换为一系列指标构建的。由于个人信用信息数据库中存在大量的冗余、不相关信息，往往导致评估指标相互重叠、干扰，使得评估模型效率和准确率得不到保证。因此对原始指标体系的优化有利于简化分类规则，缩短数据计算时间，降低数据中噪声的干扰，提高指标体系的科学性，从而使评估模型能够处理大规模的样本数据。

除了建立个人信用信息数据库、选取合适的指标外，还必须确立科学客观的个人信用评估方法。不同的个人信用评估方法有各自的优缺点和适用范围，应全面综合评价各种评估模型，根据需求选择适合的评估模型确定个人信用等级，得到个人信用报告，从而为银行贷款决策提供依据。

2 GA-CFS属性选择方法

2.1 属性选择

属性选择是数据预处理过程中广泛使用的技术。一方面，如果删除不相关和冗余的属性并降低噪声，会使得许多数据挖掘算法的效果更好。另一方面，属性选择可能导致最后生成的模型能保持或提高泛化能力。

属性选择的目的是去除一些不相关和冗余的属性，得到一个满足特定标准的最小的属性子集。属性选择过程的形式化描述如下:设原始数据集对应的变量集 $X=\{A_1, A_2, \cdots, A_n, C\}$，其中 $S=\{A_1, A_2, \cdots, A_n\}$是属性变量的集合，C 是有 l 个取值的类变量。属性选择的过程是从 S 中选择 m 个属性得到一个属性子集 $B=\{A_{i1}, A_{i2}, \cdots, A_{im}\}$（$m \leq n$），B 中属性个数最少，而且 B 可以达到最好的分类效果。

选择一个好的属性子集，通常采用两种方法。一种为过滤（Filter）法，它独立于分类学习方法，在分类学习开始之前，先过滤属性集产生一个最优属性子集；另一种称为包装（Wrapper）法，在属性选择过程中使用特定的学习算法对属性子集进行评价。从属性选择的实现方面来讲，属性选择是由属性子集的搜索和属性评价两部分构成的。属性选择的搜索方法主要可以分为穷举搜索[7]、启发式搜索[8]和随机搜索[9]三类；评价过程可以分为：距离法、信息熵法、相关性方法、一致性方法和精确度方法五类[10]。

2.2 遗传算法

遗传算法是模拟达尔文生物进化论的自然选择和

© 中国科学院软件研究所 http://www.c-s-a.org.cn

遗传学机理的生物进化过程的计算模型，是一种通过模拟自然进化过程搜索最优解的方法。遗传算法首先将问题假设解按某种形式进行编码，通过编码组成初始群体后，按照适者生存和优胜劣汰的原理，逐代演化产生出越来越好的近似解，在每一代，根据问题域中个体的适应度大小选择个体，并借助于自然遗传学的遗传算子进行组合交叉和变异，产生出代表新的解集的种群。末代种群中的最优个体经过解码，可以作为问题近似最优解。

2.3 CFS 评价方法

属性间相关性的度量是数据挖掘的一个基本问题，广泛应用于属性选择、数据清理、贝叶斯分类器等方面。计算属性之间相关度的方法有：线性相关系数法，信息增益法和对称不确定性法。文献[11]提出的 CFS 评价方法是一种基于相关性的属性子集评价方法，计算各子集中每个属性与类属性的关联度及属性之间的冗余度，关联度越大、冗余度越小则评价值越高。在 CFS 算法中，利用信息增益来计算属性之间的相关性大小，下面首先给出信息增益的定义。

根据信息论知识，计算属性 A_i 的熵公式和计算在已知属性 A_j 的情况下属性 A_i 的熵公式分别如下公式(1)和公式(2)所示。

$$H(A_i)=-\sum_k p(a_{ik})\log_2\left(p(a_{ik})\right) \quad (1)$$

$$H(A_i\mid A_j)=-\sum_t p(a_{jt})\sum_k p(a_{ik}\mid a_{jt})\log_2\left(p(a_{ik}\mid a_{jt})\right) \quad (2)$$

差值 $H(A_i)-H(A_i|A_j)$即属性 A_i 的熵的减少量反映了属性 A_j 提供给属性 A_i 的附加信息，被称为信息增益，其值越大，A_i 与 A_j 的相关性越强。信息增益是一种对称性测量方法，其缺点是趋向于拥有更多信息的属性。因此，信息增益应该规格化，以确保它们是可比较的和具有相同的效果。对称不确定性方法弥补了信息增益的缺点，且规格化其值为区间[0,1]，可以描述为如下公式(2.3)。

$$SU=2.0\times\left[\frac{H(A_i)-H(A_i\mid A_j)}{H(A_i)+H(A_j)}\right] \quad (3)$$

CFS 算法是一种过滤算法[11]，它根据基于相关性的启发式评价函数来选择属性子集。评价函数的特点是选择包含这些属性的属性子集：属性与类属性高度相关，而属性之间不相关。不关联属性应该去除，因为它们与类属性相关性很低；冗余属性应该筛选出，因为它们高度相关于一个或更多的已选属性。在属性选择遗传算法搜索中每个染色体(基因子集)的适应度用 CFS 评价，根据上述原理，对于种群中的每个个体 h，其适应度函数可以描述为如下公式(4)。

$$Fitness(h)=\frac{m\cdot\overline{r_{ca}(h)}}{\sqrt{m+m(m-1)\overline{r_{aa}(h)}}} \quad (4)$$

其中，Fitness(h)是一个包含 m 个属性的属性子集 B 的启发式“价值”；$m\cdot\overline{r_{ca}(h)}$ 是属性与类之间相关性的均值($a\in S$)，$\overline{r_{aa}(h)}$ 是属性间交互相关性的均值，$\overline{r_{ca}(h)}$ 和 $\overline{r_{aa}(h)}$ 可根据公式(3)中 SU 关于属性之间相关性大小度量的定义计算得到。公式(4)的分子可以看作属性子集预测类标签能力的度量，分母可以看作子集内属性间冗余性大小的度量。

3 基于GA-CFS属性选择的个人信用评估模型算法

本文首先采用 CFS 对每个可能选择的属性子集进行评价，并利用得到的启发式“价值”作为适应度进行迭代，得到最优解；在最优属性子集上训练所选择的分类器，然后利用测试数据检验模型。算法过程描述如下：

算法：GA-CFS 属性选择的个人信用评估模型算法

输入：个人信用评估数据集 D；

p：群体中包含的假设数量；

c：交叉算子；

mu：变异算子；

g：迭代次数；

一种分类学习方案；

输出：一个 GA-CFS 属性选择的个人信用评估分类器。

方法：

(1) 初始化群体 P：P←随机产生 p 个属性子集；

(2) 评估：对于 P 中的每个属性子集 h，计算适应度 Fitness(h)；

(3) do {

产生新的一代 PS:

1) 选择：父代中最优的两个个体直接遗传到下一代；

2) 交叉和变异：根据上面给出的 Fitness(hi)，从 P 中按轮盘赌方法选择（p-2）/2 对属性子集。对于每对属性子集<h1，h2>，应用交叉算子 c 得到<s1，s2>，对<s1，s2>应用变异算子 mu 随机选择变异位进行变异，产生两个个体，把所有的后代加入 PS；

© 中国科学院软件研究所 http://www.c-s-a.org.cn

3) 更新：P←PS；

4) 评估：对于 P 中的每个 h，计算适应度 Fitness(h)；

} while (性能不再提高或达到迭代次数 g)

(4) 在选取的最优属性子集上训练分类器。

图 1 GA-CFS 属性选择的个人信用评估模型算法

算法中基于 GA-CFS 属性选择的分类器产生的时间不仅与基分类器种类有关，也与属性子集的选择时间有关。其中原始属性集中属性个数为 n，已选择属性子集中属性个数为 m（m≤n），属性子集的选择时间是一个关于 n 的函数 $F_{sel}(n)$，基分类器的训练时间是一个关于 m 的函数 $F_{train}(m)$，则基于 GA-CFS 属性选择的分类器训练的时间为 $F_{sel}(n)+F_{train}(m)$，单个基分类器的训练时间为 $F_{train}(n)$。在模型训练时间上，基于 GA-CFS 属性选择的分类器与单个基分类器对于不同的数据集和不同的基分类器可能会有不同的结果，但是基于 GA-CFS 属性选择的分类器算法能达到较高的准确率，所以它适用于更关注准确率的情况。

4 实验

4.1 数据描述及实验方法

本文利用 Australian 某商业银行的个人信用贷款数据作为研究数据集。数据集中共有 690 个样本。定义了两类人，第一类(Good Credit)样本 383 个，第二类(Bad Credit)样本 307 个，每个样本含有 14 个属性变量，1 个类别变量，两类人的数量相对比较均衡。

根据 GA-CFS 属性选择方法，采用 10 次 10 重交叉验证方法进行分类，分类算法分别采用有代表性的决策树、朴素贝叶斯、逻辑回归、支持向量机，将 GA-CFS 属性选择方法和分类算法结合执行的结果与使用单个分类算法的结果进行比较。本文所做的实验都是在新西兰怀卡托大学开发的 Weka 平台上进行的，决策树、朴素贝叶斯、逻辑回归、支持向量机分别采用 Weka 中的 ID3、Naive Bayes（简称 NB）、Logistic 和 SMO 算法。

4.2 实验结果与分析

为了客观地评价分类器的性能，最小化数据间相关性的影响，改进计算结果的可靠性，样本按比例随机分成 10 个等份，每次保留独立的一份作为测试集，取其余的 9 份作为训练集，每次的测试集均不相同，轮换计算 10 次，其均值作为 1 次 10 重交叉验证中准确率的估计。重复这一过程 10 次，用 10 次准确率的平均值作为最终分类准确率。

在 GA-CFS 属性选择算法中，设置算法参数：p=20，c=0.6，mu=0.033，最大迭代次数 g=100，算法终止条件为达到最大迭代次数或性能不再提高。表 1 给出了 ID3、NB、Logistic、SMO 以及本文采用的 GA-CFS 属性选择方法和这四种分类算法分别结合执行的结果。

表 1 实验结果

分类器	属性数	属性数（GA-CFS）	准确率（%）	准确率（GA-CFS）（%）
ID3	14	5	73.38	83.93
NB	14	5	85.17	85.78
Logistic	14	5	83.83	85.01
SMO	14	5	85.07	85.43

实验结果表明，在单分类器模型中分类准确率最高的是 NB，准确率为 85.17%。与原数据集相比，采用 GA-CFS 属性选择方法后，属性子集维度都降低了 64.29%，并且基于 GA-CFS 属性选择的分类器模型比 ID3、NB、Logistic 和 SMO 的单分类器模型都具有更高的分类准确率。准确率虽然只是提高了几个百分点，但是在实际应用中准确率的微小提高就有可能使授信者避免巨额的损失。

5 结论

利用数据挖掘技术来实现信用评估是目前的一个研究热点，本文将 GA-CFS 属性选择方法引入个人信用评估中，建立了基于 GA-CFS 属性选择的个人信用评估模型。实验结果表明，基于 GA-CFS 属性选择的个人信用评估模型比基于单分类器的个人信用评估模型具有更高的分类准确率。在信用评估领域，分类准确率即使只有微小的提升，就有可能给授信者带来很大的收益。除了本文应用的 GA-CFS 属性选择方法，属性选择方法还有许多。今后的工作将围绕这方面展开，使其分类器的性能进一步提高。

参考文献

1 Crook JN, Edelman DB, Thomas LC. Recent developments in consumer credit risk assessment. European Journal of Operational Research, 2007,183(3):1447－1465.

2 郭春香,李旭升.贝叶斯网络个人信用评估模型.系统管理学

（下转第 161 页）

© 中国科学院软件研究所 http://www.c-s-a.org.cn

据为由 ECC 生成的 D0=d*k*p；干扰数据为由 D0 经过不断的 SHA-1 运算生成的数据。然后与声音数据进行合成后经 UDP 协议将数据发送出去。在接收端接到数据后，要先根据事先约定的 ECC 参数计算出 D0 值，然后根据数据的信息头进行相应的 SHA-1 计算，然后将数据进行分解得到压缩后的声音数据，然后进行 G729a 运算还原出声音，通过 DirectSound 将声音播放出来。加密流程如图 2 所示。

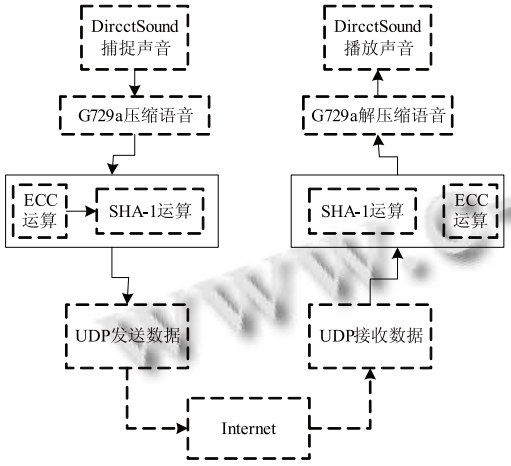

图 2 加密流程

4 结论

本文致力 VoIP 的应用及安全方面的研究，保证通话安全性。深入研究安全散列算法，论述了其实现的过程，并对其进行了 VC 的实现。使用椭圆曲线密码与 SHA 结合，加密语音数据，实现语音的实时加密。

参考文献

1 杨波.现代密码学.北京:清华大学出版社,2007.
2 斯托林斯.孟庆树,王丽娜,傅建明译.密码编码学与网络安全——原理与实践.北京:电子工业出版社,2006.
3 斯皮尔曼,叶阮健,曹英,张长富.经典密码学与现代密码学.北京:清华大学出版社,2005.
4 Stinson DR.密码学原理与实践.北京:电子工业出版社,2003.
5 孙淑玲.应用密码学.北京:清华大学出版社,2004.
6 卢开澄.计算机密码学.北京:清华大学出版社,2003.
7 Sun LH. Study of Authentication Based on Smart Card and Fingerprint Dynamic Password ICICCI 2010.

（上接第 213 页）

报,2009,18(3):249－254.
3 姜明辉,谢行恒,等.个人信用评估的 Logistic-RBF 组合模型.哈尔滨工业大学学报,2007,39(7):1128－1130.
4 刘军丽,陈翔.基于决策树的个人住房贷款信用风险评估模型.计算机工程,2006,32(13):263－265.
5 肖文兵,费奇.基于支持向量机的个人信用评估模型及最优参数选择研究.系统工程理论与实践,2006,26(10):73-79.
6 Ni ZW, Li FG, Yang SL, Liu X, Zhang WL, Luo Q. Attributes reduction based on GA-CFS method. LNCS, 2007(4505):868－875.
7 Narendra PM, Fukunaga K. A branch and bound algorithm for feature subset selection. IEEE Trans. on Computers, 1977, 26(9):917－922.
8 Zhou R, Hansen E. Breadth-First heuristic search. Artificial Intelligence, 2006,170(4-5):385－408.
9 Gheorghies O, Luchian H, Gheorghies A. A study of adaptation and random search in genetic algorithms. Proc. of the 2006 IEEE Congress on Evolutionary Computation (CEC). 2006:2103－2110.
10 Almuallim H, Dietterich TG. Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence. 1994,69(1-2):279－305.
11 Hall MA. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. Proc. of the 17th International Conference on Machine Learning (ICML), 2000:359－366.

© 中国科学院软件研究所 http://www.c-s-a.org.cn