基于光量子计算的信用评分特征筛选研究报告

文 凯[1],马 寅[1],王 鹏[2],朱德立[2]
(1.北京玻色量子科技有限公司,北京 100016;2.光大科技有限公司,北京 100083)

摘 要:随着科学技术的进步,量子计算突破了传统的算力瓶颈,在各个领域发挥着越来越重要的作用。在金融领域中,信用评分场景是贷款行业的重中之重。特征选取是一种十分高效的数据预处理策略。特征选择可以构建更简单、更容易理解的模型,提高数据挖掘性能,从中提取有效的特征,降低数据维度,为金融业提供有效的贷款参考信息。主要讨论量子计算在信用评分场景下的应用,改进了金融数据预处理的方式,创新性地使用量子计算机来求解特征选择的 QUBO 模型,与 one-hot 转码相比较,所使用的 WOE 分箱处理策略可以直接解析出特征,筛选结果可以进行直接对比。基于量子计算的特征选取与传统的基于相关性的特征选取策略相比,差距很小,并且由于量子计算机的先天优势,此策略速度更快,更具前景。

关键词:量子计算;特征选择;相关性;信用评分
中图分类号:TP38 **文献标识码**:A DOI:10.20044/j.csdg.2097-1788.2022.03.002
引用格式:文凯,马寅,王鹏,等. 基于光量子计算的信用评分特征筛选研究报告[J].网络安全与数据治理,2022,41(3):13-18.

Research report on feature screening of credit scoring based on photonic quantum computing

Wen Kai[1], Ma Yin[1], Wang Peng[2], Zhu Deli[2]
(1.Beijing Qboson Quantum Technology Co., Ltd., Beijing 100016, China;
2.Everbright Technology Co., Ltd., Beijing 100040, China)

Abstract: With the progress of science and technology, quantum computing has broken through the traditional bottleneck of computing power and is playing an increasingly important role in various fields. In the financial field, the credit scoring scenario is the top priority of the loan industry. Feature selection can build simpler and easier-to-understand models, improve data mining performance, prepare concise and understandable data to extract effective features and reduce data dimensions, and provide effective loan reference information for the financial industry. This paper mainly discusses the application of quantum computing in the credit scoring scenario, improves the financial data preprocessing method, and creatively uses quantum computer to solve the QUBO model of feature selection. Compared with one-hot encoding, the WOE strategy used in this paper can directly analyze the features, and the screening results can be directly compared. Compared with the traditional feature selection strategy based on correlation, the feature selection strategy based on quantum computing is little difference. Moreover, due to the inherent advantages of quantum computer, this strategy is faster and more promising.

Key words: quantum computing; feature selection; correlation; credit evaluation

0 引言

目前,量子计算是未来的计算发展趋势,全球各主要研究机构和公司选用不同的物理方案来制造量子计算机,主流的技术路线包括超导量子计算、光量子计算等。超导量子计算系统对环境要求苛刻,要求在绝对零度附近的超低温下才能工作;光量子计算其原理是使用光量子的叠加态对组合优化问题进行指数级求解加速。基于光量子系统的相干伊辛计算架构(Coherent Ising Machine, CIM)[1],具有光量子常温下编码操控和其在相干时间、室温

工作、全联接等方面的技术优势。目前,国内北京玻色量子科技有限公司等企业,已完成第一台全国产光量子计算原型机的设计制造。

CIM 可以充分利用光量子常温下编码操控的技术优势,实现 100~100 000 量子比特的量子计算的有效应用和算法优越性验证[2],并且可以广泛地应用于生物制药、交通、人工智能[3-7]等领域。在金融风控领域,特别是在信贷业务场景下,需要利用客户多维度的特征,对客户未来的违约行为做出预测,从而进行风险控制决策。因此好的风控评估模型能为银行风控业务提供从资产负债、信用风险、反欺诈、反洗钱等全方位完整的风险控制方案。在建立风控模型的过程中,随着大数据时代的到来,客户数据维度呈指数型增长,传统的特征筛选方法需要人工经验的参与,对大维度数据的处理显得较为吃力,亟需创新式的解决方案。量子计算作为超强算力的代表,在此领域拥有极大的潜力。

在信用评分的建模场景中[8],特征选择在整个过程起着至关重要的作用,通过筛选后续入模的特征从而提高模型的准确率和效率,并具有更好的泛化能力。尤其是在特征数较大时,不同特征的选择将决定最后信用评分模型的整体效果。本文将采用传统信用评分的建模逻辑,对于特征筛选这一环节,采用量子计算的方式进行优化,从而对整体模型效果进行提升(并与传统方式的特征选择进行对比)。通过建立相应的二次无约束二值优化(Quadratic Unbounded Binary Optimization,QUBO)[9]模型来实现特征选择,该模型理想情况下选择既独立又有影响力的特征。此次研究主要通过量子计算解决 QUBO 模型来实现特征选择,相比传统信用评分的特征选择,在不牺牲准确率的前提下,量子计算效率更高而且人工干扰更少,并在特征数很大时,解决了人工筛选难度大的问题。

1 数据及预处理

本文采用的数据是德国信用数据,其中包括 20 个特征(7 个数字特征,13 个分类特征)和 1 个二元分类特征(良好信用或不良信用)。在此基础上,本文采用了两种数据预处理的方式。

方式 A:将分类特征进行 one-hot 编码[10],使得特征数增加为 48 个;

方式 B:采用传统信用评分业务中的建模逻辑,对原始数据进行 WOE 分箱处理,不改变原有的特征数。

将处理后的数据作为 QUBO 模型的输入,用量子计算机求解 QUBO 模型,输出选择后的特征子集。

经过预处理后,得到一个 m 行,n 列的矩阵 $\boldsymbol{U}$,每一列代表一个特征,每一行表示信用申请人的相应数据值。

$$\boldsymbol{U}=\begin{bmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{m1} & \cdots & u_{mn} \end{bmatrix} \tag{1}$$

历史信用记录表示为 m 个元素的向量 $\boldsymbol{V}$:

$$\boldsymbol{V}=\begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} \tag{2}$$

其中原始数据中代表信用 credit 的数据值(v_i)为 01 变量,0 表示接受,1 表示拒绝信贷申请。

在建立 QUBO 模型时,需要计算特征之间的相关性及每个特征对信用 $\boldsymbol{V}$ 的相关性,而实验 A、B 也采用了不同的处理方式:

实验 A:用斯皮尔曼相关性计算方法

实验 B:沿用斯皮尔曼相关性计算特征之间的相关性,用信息变量(Information Value,IV)值替换特征与信用数据之间的相关性。

2 特征选取

特征选取作为一种数据预处理策略,已被证明可以适用在各种数据挖掘和机器学习问题上,且对最终模型效果起到显著的作用。特征选择的目标包括构建更简单、更容易理解的模型,提高数据挖掘性能,以及准备干净、可理解的数据。从方法论上讲,为了强调传统数据现有特征选择算法的异同,一般分为四类[11]:基于相关性[12]、基于信息理论[13]、基于稀疏学习和基于统计的方法[14]。本文主要讨论了两种特征选取策略:基于相关性的传统特征选取;基于量子计算的特征选取。

2.1 传统特征选取

假设从 n 个特征的原始集合中想要选择具有 m 个特征的一个子集,用于做出信用决策。首先,通过 IV 值筛选掉对结果影响不大的冗余特征,在此基础上选择出相关性较高的特征对。

2.2 量子特征选取

从数学上讲,特征选取的目标将是找到与向量 $\boldsymbol{V}$ 相关,但彼此不相关的矩阵 $\boldsymbol{U}$ 的列。令 ρ_{ij} 表示矩阵 $\boldsymbol{U}$ 的第 i 列与第 j 列的相关性,ρV_j 表示 $\boldsymbol{U}$ 的第 j

列与 V 的单列的相关性。为了找到“最佳”子集，本文引入了 n 个二进制变量 x_j，它们具有如下数学含义：

$$x_j=\begin{cases}1, & \text{特征 } j \text{ 在子集中}\\ 0, & \text{特征 } j \text{ 不在子集中}\end{cases} \tag{3}$$

将这些元素共同组成向量 X，形如：

$$X=\begin{bmatrix}x_1\\ \vdots\\ x_m\end{bmatrix} \tag{4}$$

筛选最佳特征子集，求解最小化目标函数对应的 X 的值，目标函数由两部分组成：第一个部分表示特征对被标记的类的影响为：

$$H_1=\sum_{j=1}^{n}x_j|\rho V_j| \tag{5}$$

第二个组成部分代表了独立性为：

$$H_2=\sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n}x_jx_k|\rho_{jk}| \tag{6}$$

引入参数 $\alpha(0\leqslant\alpha\leqslant1)$ 以表示独立性(在 $\alpha=0$ 时最大)和影响性(在 $\alpha=1$ 时最大)的相对权重并得到如下的目标函数为：

$$f(x)=-\alpha H_1+(1-\alpha)H_2 \tag{7}$$

QUBO 模型的数学表达式为：

$$f(x)=-\sum_{i\leqslant j}q_{ij}x_ix_j \tag{8}$$

其中 x_i 为待求二进制变量，取值为 $\{0,1\}$，q_{ij} 为二次项系数，为已知量，当 $i=j$ 时，将 x_i^2 简化为 x_i。将 $f(x)$ 写成线性代数的形式：

$$f(x)=-X^{\mathrm{T}}QX \tag{9}$$

通过 CIM 求解向量 X^*，从而得到筛选后的特征子集为：

$$X^*=\arg\min_{X}[-X^{\mathrm{T}}QX] \tag{10}$$

固定超参数 α 的值后筛选的特征结果如下：

(1)超参数 α 的值为 0.977 时，特征选择从 48 个特征中得到的特征数量是 24 个，使得模型的预测准确率达到极大值。由于其中的分类特征经过 one-hot 编码之后没有直观的意义，在此不再与传统筛选的特征进行比对，只在后续的准确率计算中进行比对。

(2)超参数 α 的值为 0.97 时，特征选择从 20 个特征中选取 12 个特征，统手工筛选出 13 个特征，如表 1 所示。

3 评估指标及实验结果

信用评分模型的评估是通过未加权精度，即正确分类的数量除以分类的总数，对训练集和测试集的预测结果进行评分。

零规则：德国信用数据有 700 个 0 类样本(“良好信用”)和 300 个 1 类样本(“不良信用”)。因此，将所有样本分配给 0 类的“盲猜模型”将获得 70% 的成功率。

本文希望量子特征选择比零规则和随机选择的子集更好，结果可以媲美甚至超过传统的特征选择模型。在进行特征选择之前，首先确定逻辑回归模型在整个特征集上的表现，平均精度取决于数据被打乱的次数，以及数据如何在训练集和测试集之间进行分割。

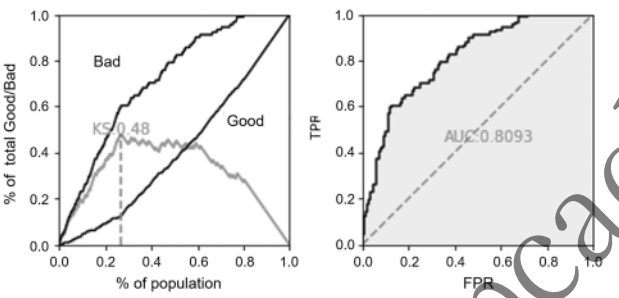
表 1 传统筛选和量子筛选特征结果对比

传统方法筛选后的特征	量子计算筛选后的特征
status.of.existing.checking.account_woe	status.of.existing.checking.account_woe
foreign.worker_woe	foreign.worker_woe
present.employment.since_woe	present.employment.since_woe
credit.amount_woe	credit.amount_woe
savings.account.and.bonds_woe	savings.account.and.bonds_woe
other.installment.plans_woe	other.installment.plans_woe
purpose_woe	purpose_woe
duration.in.month_woe	duration.in.month_woe
property_woe	property_woe
housing_woe	housing_woe
credit.history_woe	credit.history_woe
age.in.years_woe	age.in.years_woe
installment.rate.in.percentage.of.disposable.income_woe	

选择 1 000 次洗牌和 20%的测试份额的组合作为初始性能比较的标准。其他研究表明在德国信用数据上使用传统的特征选择准确性得分通常在 70%~75%之间,标准差在 5%左右。以下的实验结果均是基于 1 000 次洗牌和 20%的测试份额的初始设置进行,并且根据 K-S、ROC 以及 LR 评判模型判断算法的好坏。

实验 A:用 one-hot 编码对原始数据处理后获得的实验结果

图 1 中,图 1(a)展示了 K-S 指标,其表示随着样本数(% of population)的增加,样本数中好的百分比和坏的百分比之间的差值的最大值;图 1(b)展示了 ROC 曲线,阴影部分为 AUC 面积,代表了随着FPR 的增加 TPR 的变化,AUC 越接近 1 越好。这两个值经常用来评判模型区分样本好坏的程度。表 2 为具有 48 个特征的 LR 模型的准确率,表 3 为不同的超参数进行量子特征选择的结果。



(a)K-S (b)ROC

图 1 48 个特征的 LR 模型的 K-S 和 ROC

表 2 具有 48 个特征的 LR 模型的准确率

	平均精度	标准差
训练集	0.786 2	0.008 1
测试集	0.754 8	0.027 2
全部	0.779 9	0.004 9

表 3 不同的超参数进行量子特征选择的结果

超参数 α	训练集平均精度	测试集平均精度	平均精度
0.90	0.699 6	0.701 2	0.699 9
0.91	0.699 7	0.701 3	0.700 0
0.92	0.699 5	0.701 1	0.699 8
0.93	0.713 5	0.713 1	0.713 4
0.94	0.737 4	0.733 0	0.736 5
0.95	0.737 9	0.730 4	0.736 4
0.96	0.749 6	0.738 0	0.747 3
0.97	0.760 2	0.746 3	0.757 4
0.98	0.777 6	0.761 0	0.774 3
0.99	0.780 0	0.755 6	0.775 1

不同的超参数进行量子特征选择的测试集结果如图 2 所示,考虑 $\alpha \geqslant 0.9$,精度高于零规则结果,从图 2 可以看到测试集的效果在 $\alpha=0.98$ 时达到较好的结果之后开始下降。$\alpha=0.98$ 时模型的 K-S 和 ROC 如图 3 所示,$\alpha=0.98$ 时进行量子特征选择后的模型准确率如表 4 所示。

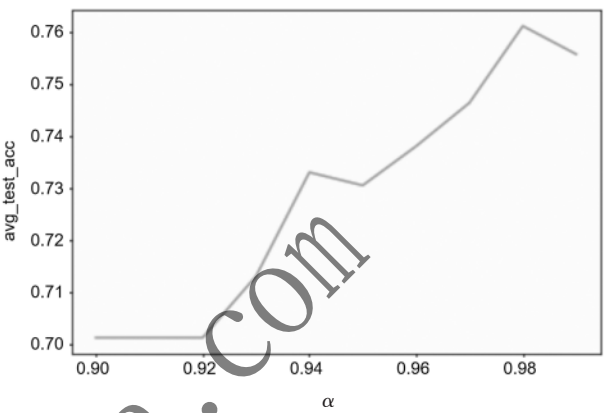


图 2 不同的超参数进行量子特征选择的测试集结果

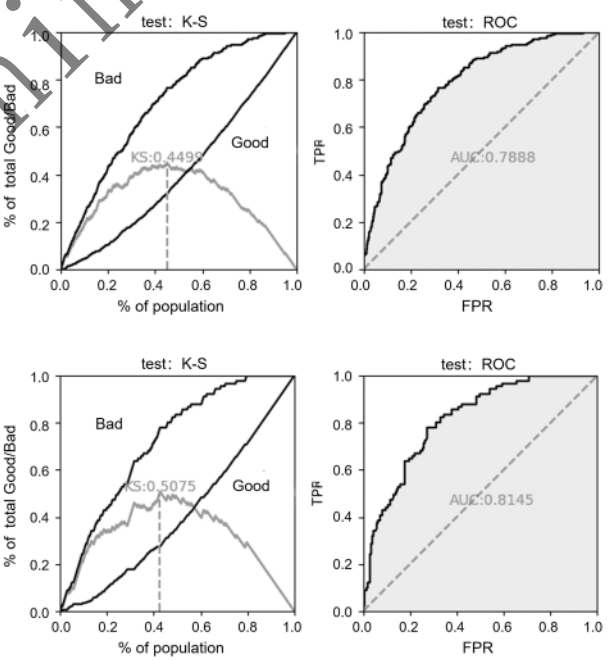


图 3 $\alpha=0.98$ 时模型的 K-S 和 ROC

表 4 $\alpha=0.98$ 时进行量子特征选择后的模型准确率

	平均精度	标准差
训练集	0.777 6	0.007 9
测试集	0.761 0	0.026 7
全部	0.774 3	0.004 8

实验 B：用 WOE 分箱策略预处理数据，获得的实验结果如图 4 所示，全 20 个特征代入 LR 模型的模型准确率如表 5 所示，不同的超参数进行量子特征选择的结果如表 6 所示。

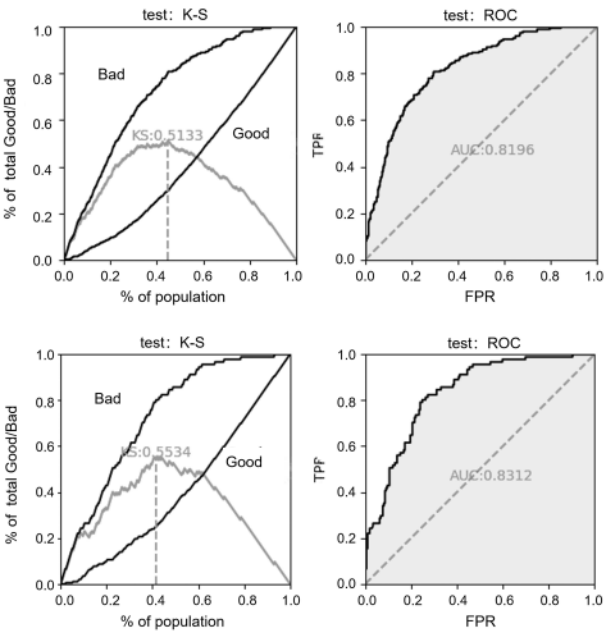


图 4 全 20 个特征代入 LR 模型的 K-S 和 ROC

表 5 全 20 个特征代入 LR 模型的模型准确率

	平均精度	标准差
训练集	0.773 5	0.007 7
测试集	0.760 8	0.026 5
全部	0.771 0	0.004 2

表 6 不同的超参数进行量子特征选择的结果

超参数 α	训练集平均精度	测试集平均精度	平均精度
0.7	0.699 7	0.701 3	0.700 0
0.8	0.727 4	0.722 8	0.726 5
0.9	0.761 5	0.756 7	0.760 5
1.0	0.773 5	0.760 8	0.771 0

更进一步得到 $\alpha=0.98$ 时，测试集的结果表现令人满意(如图 5 所示)，之后的精度增长趋于平缓。将选择的特征放入 LR 模型进行训练，结果如图 6 所示，20 个特征用量子计算特征选择之后的模型准确率如表 7 所示。

4 结论

在与传统的特征筛选方式进行对比后发现，本文采用的 WOE 策略与传统的 one-hot 编码相比，结果展示更为直观。通过量子计算方法筛选得到的特征与传统方法筛选的特征相比差别极小，在不降低准确率的情况下，基于量子计算的特征选取策略可以减少人为的参与，提高效率并降低对业务人员的依赖，从而减少操作风险。而在 K-S 以及 ROC 这两个评价模型中，量子计算策略是优于传统筛选策略；在 LR 评价模型中，量子计算策略和传统筛选策略效果近似。本文展示了量子计算应用于特征筛选该类特定问题上的可行性，尤其是面对特征数巨大的

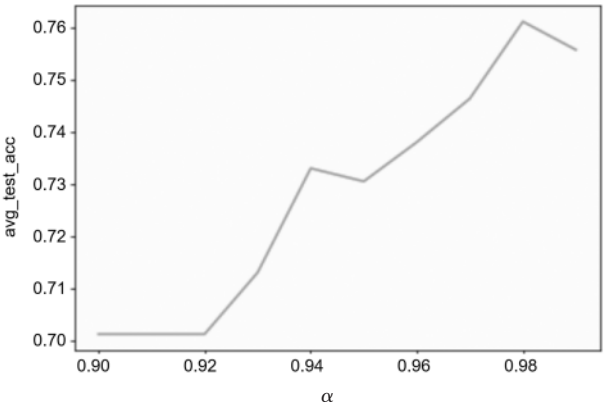


图 5 不同的超参数进行量子特征选择的结果

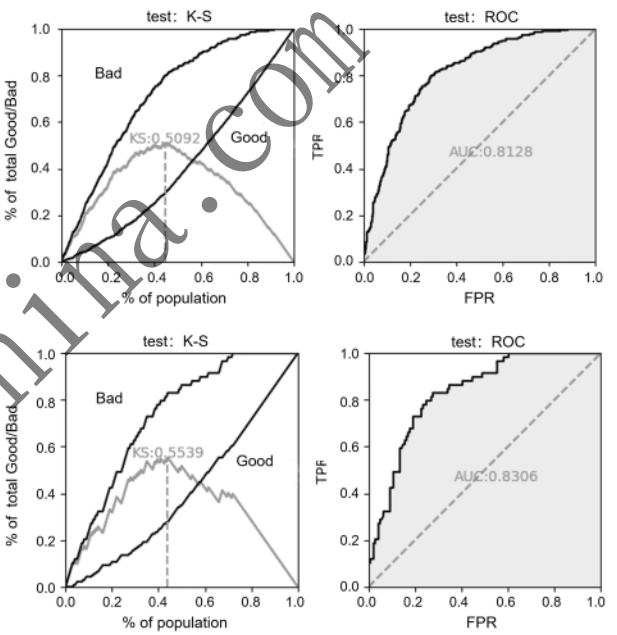


图 6 $\alpha=0.98$ 时特征筛选后模型的 K-S 和 ROC

表 7 $\alpha=0.98$ 时进行量子特征选择的结果

	平均精度	标准差
训练集	0.772 7	0.007 7
测试集	0.761 6	0.026 6
全部	0.770 5	0.004 4

情况下,量子计算更显优势,其超越并替代传统方法的潜力巨大。

随着量子计算机和量子计算算法的发展,传统业务中的一些难题将迎来新的技术解决方案,例如计算成本较大、传统计算机的并行计算能力不高以及问题最优解优化不够等问题,都可以通过量子计算来解决。将量子计算运用到金融传统业务场景中的特定问题上,将是现阶段重点探讨和未来努力的方向。

参考文献

[1] WANG Z, MARANDI A, WEN K, et al. Coherent Ising machine based on degenerate optical parametric oscillators[J]. Phys. Rev. a, 2013, 88(6): 3869-3876.

[2] HONJO T, SONOBE T, INABA K, et al. 100,000-spin coherent Ising machine[J]. Science Advances, 2021, 7(40): eabh0952.

[3] SARKAR A, AL-ARS Z, BERTELS K. QuASeR: quantum accelerated de novo DNA sequence reconstruction[J]. PLoS ONE, 2021, 16(4): e0249850.

[4] MARCHAND D J J, NOORI M, ROBERTS A, et al. A variable neighbourhood descent heuristic for conformational search using a quantum annealer[J]. Scientific Reports, 2019, 9(1): 1-13.

[5] MAGUIRE J B, GRATTAROLA D, MULLIGAN V K, et al. XENet: using a new graph convolution to accelerate the timeline for protein design on quantum computers[J]. PLOS Computational Biology, 2021, 17.

[6] FELD S, ROCH C, GABOR T, et al. A hybrid solution method for the capacitated vehicle routing problem using a quantum annealer[J]. Frontiers in ICT, 2019, 6: 13.

[7] ARTHUR D, PUSEY-NAZZARO L. QUBO formulations for training machine learning models[J]. Scientific Reports, 2021, 11(1): 1-10.

[8] 石庆焱,靳云汇.个人信用评分的主要模型与方法综述[J].统计研究,2003,8(82):36-39.

[9] GLOVER F, KOCHENBERGER G, HENNIG R, et al. Quantum bridge analytics I: a tutorial on formulating and using QUBO models[J]. Annals of Operations Research, 2022: 1-43.

[10] RODRÍGUEZ P, BAUTISTA M A, GONZALEZ J, et al. Beyond one-hot encoding: lower dimensional target embedding[J]. Image and Vision Computing, 2018, 75: 21-31.

[11] LI J, CHENG K, WANG S, et al. Feature selection: a data perspective[J]. ACM Computing Surveys (CSUR), 2017, 50(6): 1-45.

[12] EL AKADI A, ELOUARDIGHI A, ABOUTAJDINE D. A powerful feature selection approach based on mutual information[J]. International Journal of Computer Science and Network Security, 2008, 8(4): 116.

[13] VIDAL-NAQUET M, ULLMAN S. Object recognition with informative features and linear classification[C]// ICCV, 2003(3): 281.

[14] YANG S, YUAN L, LAI Y C, et al. Feature grouping and selection over an undirected graph[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012: 922-930.

(收稿日期:2022-07-24)

作者简介:

文凯(1982-),男,博士,主要研究方向:量子计算、人工智能、凝聚态物理。

马寅(1986-),男,硕士,主要研究方向:人工智能、精密仪器、光电信息。

王鹏(1986-),男,硕士,工程师,主要研究方向:大数据、人工智能、金融科技。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部