Statistics and Application 统计学与应用, 2026, 15(2), 31-48
Published Online February 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.152032

Hans汉斯

基于无监督特征选择的半导体质量检测方法

余青青

重庆工商大学数学与统计学院，重庆

收稿日期：2025年12月30日；录用日期：2026年1月21日；发布日期：2026年2月2日

摘　要

半导体质量检测数据存在高维冗余、类别不平衡及标签获取成本高的特性，导致传统有监督检测方法工业适用性受限，而传统无监督特征选择方法往往忽略全局结构或缺乏冗余量化机制，难以适配高维强非线性耦合的数据需求。基于此，本文提出一种面向半导体高维制造数据的无监督特征选择方法FSSC-DCOR (Feature Selection by Spectral Clustering and Distance CORrelation coefficient)。该方法结合谱聚类、距离相关系数与贪心策略三种技术，以特征为聚类对象，通过谱聚类挖掘特征内在关联结构并筛选高信息量候选特征，利用距离相关系数矩阵量化非线性冗余，最终通过贪心策略保留低冗余、高区分度的核心特征子集。该方法无需依赖标注标签即可完成高维数据有效降维，适配半导体场景标签稀缺的现实需求。实验结果表明，在SECOM半导体数据集上，所提方法的性能度量指标均优于传统特征选择方法。

关键词

半导体质量检测，谱聚类，距离相关系数，无监督特征选择

Semiconductor Quality Inspection Method Based on Unsupervised Feature Selection

Qingqing Yu

School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing

Received: December 30, 2025; accepted: January 21, 2026; published: February 2, 2026

Abstract

Semiconductor quality inspection data exhibits characteristics of high-dimensional redundancy, class imbalance, and high cost of label acquisition, resulting in limited industrial applicability of traditional supervised detection methods. In contrast, conventional unsupervised feature selection methods either ignore the global structure or lack redundancy quantification, making it difficult to

文章引用: 余青青. 基于无监督特征选择的半导体质量检测方法[J]. 统计学与应用, 2026, 15(2): 31-48.
DOI: 10.12677/sa.2026.152032

meet the requirements of high-dimensional and strongly nonlinearly coupled data. To address this issue, an unsupervised feature selection method named FSSC-DCOR (Feature Selection by Spectral Clustering and Distance CORrelation coefficient) is proposed for high-dimensional semiconductor manufacturing data. This method combines three techniques: spectral clustering, distance correlation coefficient, and greedy strategy. Taking features as clustering objects, it mines the intrinsic correlation structure of features through spectral clustering to select high-information candidate features, quantifies nonlinear redundancy using a distance correlation coefficient matrix, and finally retains a core feature subset with low redundancy and high discriminability via the greedy strategy. Without relying on labeled data, the method can achieve effective dimensionality reduction of high-dimensional data, adapting to the practical demand of label scarcity in semiconductor scenarios. Experimental results demonstrate that on the SECOM semiconductor dataset, the performance metrics of the proposed method are all superior to those of traditional feature selection methods.

Keywords

Semiconductor Quality Inspection, Spectral Clustering, Distance Correlation Coefficient, Unsupervised Feature Selection

1. 引言

在全球半导体产业向高精度、高集成度转型的进程中，半导体晶圆制造已发展为包含数百道精密工序、涉及多类型设备与复杂工艺参数的超复杂生产系统，产品质量直接决定电子信息产业的发展上限与终端设备可靠性[1]。质量检测作为半导体生产全流程的核心管控环节，其效能却受限于检测数据的固有特性：一方面，单一样本常包含数百个质量特性指标(如 SECOM 半导体数据集含 590 维特征)，高维数据中的冗余信息易引发“维度灾难”，增加模型计算复杂度并掩盖关键质量关联[2]；另一方面，合格产品与不合格产品样本比例悬殊(如 SECOM 半导体数据集的样本比例为 14.07:1)，类别不平衡问题易导致传统模型偏向多数类，显著提升“不合格产品误判为合格”的消费者风险[3]；更为关键的是，半导体生产中标签数据的获取需额外投入检测成本与时间成本，导致实际场景中无标签数据大量积累，而有标签数据占比极低，这一现状严重制约了依赖标签的检测方法的应用效果。这些痛点叠加，使得依赖人工经验或简单统计模型的传统检测方法，已无法满足智能制造对精准化、高效化质量管控的需求。

在半导体质量检测的实际应用场景中，相关研究已积累了较为丰富的成果，为该领域的技术优化与效率提升奠定了坚实基础，例如，Nuhu 等[1]提出融合数据插补、合成数据生成(SMOTE/BSMOTE-SVM/ADASYN)与特征选择(PCA/UFS)的故障诊断框架，经数据预处理(清洗、标准化)、多机器学习模型(RF/SVC/MLP 等)分类，结合准确率、召回率等多指标评价性能，优化半导体制造过程的故障检测与缺陷诊断，有效解决数据高维、类别不平衡及缺失值问题；程云飞等[2]提出 GA-LightGBM 方法，通过 PCA 提取特征、SMOTE 处理不平衡数据，并借助遗传算法优化参数，综合特征有效性与分类性能实现合格品与不合格品识别；柳嘉昊[3]提出基于 *K*-Means 聚类欠采样的改进随机森林算法(KMUS-RF)，经数据预处理(填补缺失值、标准化)、*K*-Means 聚类欠采样生成平衡数据集、RF 分类，结合特征重要性评价，精准识别半导体生产过程中的关键质量特性(CTQ)，有效降低第二类错误率，优化半导体质量检测与生产管控；Gómez-Sirvent 等[4]提出基于穷举搜索(ES)的晶圆缺陷分类特征选择方法，通过计算机视觉提取尺寸、形

状等四类特征构建向量，遍历所有特征子集并结合 SVM 与加权 F1-score 筛选最优子集，实现缺陷高精度分类；He 等[5]提出 FD-kNN 故障检测方法，利用 k 近邻非线性特性，仅以正常操作数据建模，通过计算待检测样本与 k 个最近邻的平方距离和并对比阈值判断故障，适配过程非线性与多模态特性；Baek 等[6]提出面向多元时间序列的故障检测与主因识别方法，将数据转化为签名矩阵，经 CAE 检测异常、MLP 预测待更换部件，再通过 SHAP 算法溯源故障主因；Qian 等[7]人提出 MI-DNS-WkNN 方法，经 KPCA 特征提取与加权，通过多步邻域分离指数校正动态邻域尺度，融合静态与动态分量设置阈值，适配复杂工况；Kuo 等[8]人提出融合 ML、DES 与 LCA 的故障检测框架，经 PCA 降维、RF 预测缺陷，结合多指标评价特征重要性，优化光刻工艺检测。Rosa 等[9]提出两阶段方法，以几何变换平衡数据、SqueezeNet 提取缺陷特征，通过网格搜索优化超参数，综合特征区分度与分类性能实现晶圆缺陷高效分类；Jiao 等[10]提出双模态成像+轻量化网络的检测方法，通过 DPConv、GSConv 与 EMA 机制构建 SWC-ResEMA-Net，综合特征细节、跨通道相关性与分类性能，检测六种混合微缺陷；闫伟等[11]提出基于 IG 的关键质量特性识别方法，度量特征与质量类别的相关程度，删除弱相关与冗余特征；李岸达等[12]提出 ReliefF-W 混合算法，通过 ReliefF 计算特征权重、FNO 确定最优特征数，综合权重与分类精度识别 CTQs；Lee 等[13]提出数据驱动方法，以 EM 插补缺失值、SMOTE 平衡数据，通过改进 MeanDiff 量化工艺步骤差异，结合与晶圆良率的关联度选择关键步骤；这些特征选择与数据处理方法本质上均为有监督分类模型服务，仍需依赖标注数据完成最终的模型训练与效果验证。

这些现有方法在标注数据充足的场景下展现出良好的检测精度，但核心共性局限在于：均以有监督学习为核心框架，对高质量标注数据存在强依赖。而在半导体实际生产中，标签数据的获取需要额外的检测设备投入、专业技术人员判定与较长的检测周期，导致有标签数据稀缺而无标签数据大量闲置，直接限制了这些有监督方法的工业适用性。与此同时，传统无监督特征选择方法(例如相关系数法[14])虽不依赖标签，但仅关注特征两两关联，忽视全局结构与非线性冗余；谱聚类等无监督技术[15]虽能挖掘数据潜在结构，却未与冗余量化有效结合，难以适配半导体高维、强非线性耦合的数据特性，降维后特征的区分性与冗余控制效果不佳。

基于此，本文提出一种面向半导体高维制造数据的无监督特征选择方法 FSSC-DCOR (Feature Selection by Spectral Clustering and Distance CORrelation coefficient)，核心目标是适配半导体场景标签获取成本高、稀缺性强的现实需求。该方法以特征为聚类对象，通过谱聚类挖掘特征间内在关联结构并筛选高信息量候选特征，再利用距离相关系数矩阵量化非线性冗余，采用贪心策略保留低冗余、高区分度的核心特征子集，无需依赖标注标签即可完成高维数据有效降维。在 SECOM 半导体数据集上的对比实验表明，该方法的性能度量指标均优于传统特征选择方法。

2. 相关算法

2.1. 谱聚类

谱聚类[16]是一种基于图论与谱分析的无监督学习算法，核心思想是将数据聚类问题转化为图的划分问题——通过对数据的相似度矩阵(或亲和矩阵)进行特征分解(谱分解)，提取低维嵌入特征后，再采用传统聚类算法(如 *K*-Means)完成最终聚类，适用于高维数据、非凸分布数据或复杂结构数据的聚类任务，在特征选择、图像分割、社区发现等领域广泛应用。

2.1.1. 谱聚类的相关定义

谱聚类的核心依赖图论中的图划分准则与线性代数中的谱分解理论，关键定义如下：

(1) 数据图构建：将每个样本视为图的顶点 v_i，构建无向加权图 $G=(V,E,W)$，其中：

$V=\{v_1,v_2,\cdots,v_n\}$ 为顶点集合(n 为样本/特征数量)；

$E=\{(v_i,v_j)\mid i\neq j\}$ 为边集合，存在边表示顶点 v_i 与 v_j 存在关联；

$W=[w_{ij}]_{n\times n}$ 为权重矩阵(亲和矩阵)， $w_{ij}\geq 0$ 表示顶点 v_i 与 v_j 的相似度($w_{ij}=0$ 表示无关联)。

(2) 相似度度量：常用度量方式包括：

径向基函数(RBF 核，FSSC-DCOR 采用)：$w_{ij}=\exp\left(-\gamma\cdot\|x_i-x_j\|^2\right)$ 其中 $\gamma>0$ 为核参数，$\|x_i-x_j\|$ 为样本 x_i 与 x_j 的欧氏距离；

最近邻权重：仅保留每个顶点的 k 个最近邻顶点的边权重(其余设为 0)，即 $w_{ij}>0$ 当且仅当 v_j 是 v_i 的 K-近邻(或反之)。

(3) 图的核心矩阵定义：

度矩阵 D：对角矩阵 $D=\mathrm{diag}(d_1,d_2,\cdots,d_n)$，其中 $d_i=\sum_{j=1}^{n}w_{ij}$ 为顶点 v_i 的所有关联边的权重和；

拉普拉斯矩阵 L：核心矩阵，定义为 $L=D-W$，其具有半正定、对称的性质，且特征值非负；

规范化拉普拉斯矩阵：为避免数据分布偏差，常用两种规范化形式：

对称规范化：$L_{\mathrm{sym}}=D^{-1/2}LD^{-1/2}$；随机游走规范化：$L_{\mathrm{rw}}=D^{-1}L$。

2.1.2. 谱聚类的核心步骤

谱聚类的核心步骤如下：

构建亲和矩阵 W：根据数据类型选择相似度度量(如 RBF 核)，计算所有顶点间的权重，形成 $n\times n$ 维亲和矩阵。

计算拉普拉斯矩阵 L：基于亲和矩阵 W 计算度矩阵 D，进而得到拉普拉斯矩阵(或规范化拉普拉斯矩阵)。

谱分解(特征值与特征向量求解)：对拉普拉斯矩阵进行特征分解，得到特征值 $\lambda_0\leq\lambda_1\leq\cdots\leq\lambda_{n-1}$ 及对应的特征向量 $u_0,u_1,\cdots,u_{n-1}$；选取前 k 个最小特征值对应的特征向量 $u_0,u_1,\cdots,u_{k-1}$，组成 $n\times k$ 维特征矩阵 U (低维嵌入空间)。

聚类(K-Means 等)：将特征矩阵 U 的每行视为一个低维样本，采用 K-Means 等传统聚类算法对其聚类，得到最终的聚类标签。

2.2. 距离相关系数

距离相关系数(dCor)研究两个变量之间的独立性，距离相关系数为 0 表示两个变量是独立的。克服了皮尔逊相关系数的弱点，距离相关系数可以描述非线性相关性[17] [18]。估计距离相关系数的步骤如下(以包含 n 个观测值的连续型随机变量 $X=(x_1,x_2,\cdots,x_n)$、$Y=(y_1,y_2,\cdots,y_n)$ 为例)：

(1) 计算变量内元素的两两距离分别计算 X、Y 内部任意两个元素之间的绝对距离：

对 X，第 i 个与第 j 个元素的距离：

$$a_{ij}=\|x_i-x_j\|\quad(i,j=1,2,\cdots,n)$$

对 Y，第 i 个与第 j 个元素的距离：

$$b_{ij}=\|y_i-y_j\|\quad(i,j=1,2,\cdots,n)$$

(2) 构建中心距离矩阵

对上述距离做“中心化”调整(消除行列均值的影响)，得到中心距离矩阵：

X 对应的中心距离 A_{ij}：

$$A_{ij}=a_{ij}-\frac{1}{n}\sum_{k=1}^{n}a_{ik}-\frac{1}{n}\sum_{k=1}^{n}a_{kj}+\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}a_{kl}$$

Y 对应的中心距离 B_{ij}：

$$B_{ij}=b_{ij}-\frac{1}{n}\sum_{k=1}^{n}b_{ik}-\frac{1}{n}\sum_{k=1}^{n}b_{kj}+\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}b_{kl}$$

(3) 计算样本距离协方差的平方将中心距离矩阵 A 与 B 的对应元素相乘，求和后除以 n^2，得到 X 与 Y 的样本距离协方差的平方：

$$\mathrm{dCov}_n^2(X,Y)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}B_{ij}$$

(4) 计算变量自身的距离方差的平方

分别计算 X、Y 与自身的距离协方差(即距离方差)：

X 的样本距离方差的平方：

$$\mathrm{dVar}_n^2(X)=\mathrm{dCov}_n^2(X,X)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}^2$$

Y 的样本距离方差的平方：

$$\mathrm{dVar}_n^2(Y)=\mathrm{dCov}_n^2(Y,Y)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}B_{ij}^2$$

(5) 计算距离相关系数

用 X 与 Y 的样本距离协方差，除以二者距离方差乘积的平方根，得到最终的距离相关系数：

$$\mathrm{dCor}_n(X,Y)=\frac{\mathrm{dCov}_n(X,Y)}{\sqrt{\mathrm{dVar}_n(X)\mathrm{dVar}_n(Y)}}$$

3. 基于谱聚类与距离相关系数的无监督特征选择方法

基于谱聚类与距离相关系数的无监督特征选择方法(FSSC-DCOR)以文献[19]提出的无监督特征选择思想为基础，针对半导体高维数据的强耦合、高冗余特性，引入特征冗余度量机制，通过“谱聚类分组–候选特征初选–冗余剔除优化”的三级框架，实现“簇内高信息量、簇间低冗余”的特征子集筛选。其核心思路是：首先以特征为聚类对象，通过谱聚类对高维半导体制造数据的特征进行分组，挖掘特征间的内在关联结构并超量筛选簇内高信息量候选特征；其次构建距离相关系数矩阵量化特征间的非线性冗余性，采用贪心策略仅保留与已选特征最大冗余值低于阈值的特征，筛选出兼具高区分度与低冗余性的核心特征子集。

为消除量纲差异对聚类结果的影响，需要对给出的数据集执行 Min-Max 标准化：

$$x'_{ij}=\frac{x_{ij}-\min\left(X_{:,j}\right)}{\max\left(X_{:,j}\right)-\min\left(X_{:,j}\right)}$$

式中，x_{ij} 为第 i 个样本在第 j 个特征的原始取值，$X_{:,j}$ 表示特征矩阵第 j 列，x'_{ij} 为标准化后特征值。

3.1. 核心定义

定义 1：特征信息量(Feature Information Content)。特征对样本类别状态的区分能力，采用选定的信

息量指标量化，如标准差、四分位距、综合指标等，记为$I\left(f_j\right)$，$I\left(f_j\right)$越大，该特征包含的特征信息量越丰富。

对每个特征簇，计算簇内各特征的信息量指标，筛选出信息量最大的特征作为簇代表特征，支撑候选特征集构建。为适配半导体数据的噪声特性，提供 3 类信息量评估指标：

(1) 标准差(std)：表征特征值离散程度，离散度越高，对晶圆合格/不合格状态的区分能力越强；

(2) 四分位距(iqr)：对异常值鲁棒性更强，适用于含传感器噪声的半导体数据；

(3) 综合指标：融合标准差(权重 0.4)、四分位距(权重 0.4)、中位数绝对偏差(MAD，权重 0.2)，兼顾信息量与抗噪性；

定义 2：特征冗余度(Feature Redundancy)。特征与已选特征子集间的非线性相关性强度，采用距离相关系数(dCor)量化，对于特征 f_j 和已选特征子集 F_{final}，其冗余度为该特征与 F_{final} 中所有特征的最大距离相关系数，记为 $\max\_\mathrm{dCor}\left(f_j,F_{final}\right)$，计算公式为：

$$\max\_\mathrm{dCor}\left(f_j,F_{final}\right)=\max_{f_s\in F_{final}}\mathrm{dCor}\left(f_j,f_s\right)$$

式中，$\mathrm{dCor}\left(f_j,f_s\right)$为特征 f_j 与 f_s 的距离相关系数，取值范围$[0,1]$；$\max_\mathrm{dCor}\left(f_j,F_{final}\right)$越大表示特征 f_j 与已选特征子集的非线性冗余性越高。

3.2. 谱聚类特征分组

谱聚类旨在将相似特征聚为一簇，为后续冗余剔除提供分组基础；候选特征初选通过超量筛选确保高信息量特征不被遗漏，二者共同构成 FSSC-DCOR 的第一阶段核心流程。

以特征为聚类对象，对标准化后特征矩阵转置得到 $X^{\mathrm{T}}\in\mathbb{R}^{d\times n}$，采用径向基函数(RBF)构建特征相似度矩阵 $S\in\mathbb{R}^{d\times d}$，量化特征间相似性：

$$S_{pq}=\exp\left(-\gamma\left\|X_{:,p}-X_{:,q}\right\|_2^2\right)$$

式中，$\gamma=0.1$为核函数参数，$X_{:,p}$、$X_{:,q}$分别为第 p、q 个特征的列向量，$\|\cdot\|_2$为欧氏范数。

基于相似度矩阵构建归一化拉普拉斯矩阵 $L=D^{-1/2}\left(D-S\right)D^{-1/2}$，其中 D 为相似度矩阵的度矩阵 $D_{ii}=\sum_j S_{ij}$ 对 L 进行特征分解，选取前 $k_{candidate}$ 个最小特征值对应的特征向量构建投影矩阵，采用离散化标签分配方式生成特征簇标签，最终将 d 个原始特征划分为 $k_{candidate}$ 个特征簇 $C=\left\{C_1,C_2,\cdots,C_{k_{candidate}}\right\}$，其中 $k_{candidate}=1.5k$ （k 为目标特征子集规模），确保候选特征数量充足。

3.3. 基于距离相关系数的冗余特征删除

针对半导体数据的非线性冗余特性提出基于谱聚类与距离相关系数的特征选择思想 FSSC-DCOR，放弃仅能捕捉线性关联的皮尔逊相关系数，采用距离相关系数(dCor)量化候选特征间的非线性冗余，并通过“最大冗余值 < 阈值则保留”的核心规则，剔除高冗余特征，得到目标规模的特征子集。

基于距离相关系数矩阵，采用贪心策略对候选特征集 $F_{candidate}$ 执行冗余剔除，核心规则为“仅当特征与已选特征子集的最大冗余值小于冗余阈值时保留”，最终得到规模为 k 的特征子集 F_{final}，步骤如下：

初始化：已选特征子集 $F_{final}=\varnothing$，设置冗余阈值 $\tau=0.8$；

遍历候选特征：按特征信息量 $I\left(f_j\right)$从高到低遍历 $F_{candidate}$ 中的特征 f_j：

若 $F_{final}=\varnothing$，直接将直接加入聚类标签为 0 的簇特征加入 F_{final}；

若 $F_{final}\neq\varnothing$，计算 f_j 与 F_{final} 的最大冗余值 $\max\_\mathrm{dCor}\left(f_j,F_{final}\right)$；当 $\max\_\mathrm{dCor}\left(f_j,F_{final}\right)<\tau$，则将 f_j 加入 F_{final}；

终止条件：当 $F_{final}=k$ 时停止遍历；若遍历完所有候选特征仍未达到 k，则补充剩余未选中特征中信息量 $I\left(f_j\right)$ 最大的特征，直至 $F_{final}=k$；

输出：最终特征子集 F_{final}

3.4. 算法整体描述

输入：训练数据集 $X\in\mathbb{R}^{n\times d}$，$n$ 为训练样本数，d 为特征数；目标特征子集规模 k；冗余阈值 τ；信息量评估方法；冗余度量方式 R。

输出：特征子集 F_{final}

1) 初始化已选特征子集 $F_{final}=\varnothing$ 全部特征集合为 $F_{all}=\left\{f_1,f_2,\cdots,f_d\right\}$，候选特征集规模 $k_{candidate}=1.5k$；

2) 对数据集 X 执行 Min-Max 标准化处理，将标准化后矩阵转置为 $X^{\mathrm{T}}\in\mathbb{R}^{d\times n}$，通过 RBF 核函数构建特征相似度矩阵，采用谱聚类算法将 F_{all} 划分为 $k_{candidate}$ 个特征簇；

3) 基于信息量评估方法 M 计算各特征簇内每个特征的信息量得分，筛选每个簇内信息量得分最大的特征，生成候选特征集 $F_{candidate}$ (满足 $\left|F_{candidate}\right|=k_{candidate}$)；

4) 基于冗余度量方式 R 构建 $F_{candidate}$ 的冗余度量矩阵，遍历 $F_{candidate}$ 中特征，按优先级依次筛选：若 F_{final} 为空则直接加入聚类标签为 0 的簇特征，否则计算当前特征与 F_{final} 中所有特征的最大冗余值 $\max\_\mathrm{dCor}\left(f_j,F_{final}\right)$，仅当 $\max\_\mathrm{dCor}\left(f_j,F_{final}\right)<\tau$ 时将该特征加入 F_{final}，直至 $\left|F_{final}\right|=k$；

5) 输出特征子集 F_{final}。

4. 实验结果与分析

4.1. 数据集说明

本实验采用 SECOM 半导体数据集[20] (UCI 机器学习仓库公开数据集)作为验证载体，该数据集聚焦半导体成品质量二分类任务，包含 1567 个样本，其中合格样本 1463 个、不合格样本 104 个，样本类别分布比例为 14.07:1，属于典型的类别不平衡工业数据集。每个样本涵盖 590 个连续型质量特性特征(涵盖晶圆沉积、蚀刻精度、掺杂浓度等制造流程关键参数)，1 个二值化类别标签(0 表示合格，1 表示不合格)。因此，在进行实验之前需要对数据进行预处理。

首先，填补缺失值。SECOM 半导体数据集中，部分样本缺少某个或某几个质量特性的数据，为便于模型进行预测，本文使用均值填充法(Mean Completer)，用每一质量特性的均值填充缺失值。

接着，标准化数据。为进一步提高模型的收敛速度和预测精度，本文使用(Min-Max Normalization)对数据样本进行无量纲化处理，具体方法如下式所示：

$$x'_{ij}=\frac{x_{ij}-\min\left(X_{:,j}\right)}{\max\left(X_{:,j}\right)-\min\left(X_{:,j}\right)}$$

式中，x_{ij} 为第 i 个样本在第 j 个特征的原始取值，$X_{:,j}$ 表示特征矩阵第 j 列，x'_{ij} 为标准化后特征值，其值将落在 0 到 1 之间。

最后，将原始标签中可能存在的−1 值统一修正为 0，保证二分类标签体系的一致性。SECOM 数据集详细信息见表 1。

Table 1. Information of the SECOM Dataset
表 1. SECOM 数据集信息

名称	类型	样本数	特征数
SECOM	分类	1567	590

4.2. 实验设计

为验证所提 FSSC-DCOR (基于谱聚类与相关系数的无监督特征选择)算法在半导体数据集 SECOM 上的有效性，本文选取 4 类特征选择方法作为对比基准：

基于谱聚类的无监督特征选择(FSSC) [19]：基础谱聚类特征选择算法，通过谱聚类对特征分簇后，从每个簇中基于标准差筛选代表性特征；

方差阈值法(Variance)：以 0.01 为方差阈值过滤低区分度特征，再按目标特征数选取方差最大的特征子集；

皮尔逊相关系数法(Pearson)：计算特征间皮尔逊相关系数，以 0.8 为阈值剔除高冗余特征，最终保留目标数量的核心特征；

主成分分析法(PCA)：基于方差解释率(95%)完成特征降维，若解释率对应特征数不足则补充至目标数量。

FSSC-DCOR 算法核心参数设置：谱聚类采用 RBF 核($\gamma = 0.1$)，特征冗余度量选用距离相关系数(dCor)，冗余阈值设为 0.8，采用 KNN (K = 5)作为分类器，簇内特征筛选采用标准差(std)作为核心指标；对比算法均保持与原始文献或工业实践一致的参数配置，确保对比公平性。

实验以特征选择数量和随机种子为核心控制变量，通过选择不同数量的特征，探究不同特征规模对算法性能的影响规律；其中随机种子选取 35 至 44 共 10 个取值进行重复实验，通过多轮随机验证消除谱聚类标签分配及分类器初始化等随机因素对实验结果的干扰，保障数据结论的统计鲁棒性。评估体系围绕特征选择算法的核心效能构建，采用宏平均精确率(Macro Precision)、宏平均召回率(Macro Recall)与宏平均 F1 分数(Macro F1)作为分类性能核心度量指标，该类指标通过对正负类样本性能的均衡考量，可有效适配 SECOM 半导体数据集的类别不平衡特性，避免多数类样本对评估结果的主导偏差；同时，通过特征数–指标值曲线的近似 AUC 量化各算法在全特征规模梯度下的整体性能差异，形成“单指标精准评估 + 整体效能量化”的双层评估体系，确保对算法性能的全面且客观的衡量。各度量指标计算过程如下：

(1) 宏平均精确率

宏平均精确率表征模型预测为某一类别的样本中，实际属于该类别的样本比例的均值，其计算分为两步：首先计算每个类别的精确率，再对所有类别的精确率取算术平均。

单类别精确率公式：

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

宏平均精确率公式：

$$\text{Macro Precision} = \frac{1}{C}\sum_{i=1}^{C}\text{Precision}_i$$

(2) 宏平均召回率

宏平均召回率表征各类别中被模型正确识别的样本占该类别总样本数比例的均值，反映模型对各类别样本的“查全能力”。

单类别召回率公式：

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

宏平均召回率公式：

$$\text{Macro Recall} = \frac{1}{C}\sum_{i=1}^{C}\text{Recall}_i$$

(3) 宏平均 F1 分数

宏平均 F1 分数是宏平均精确率与宏平均召回率的调和均值，能够综合权衡模型的查准率与查全率，避免单一指标的局限性。

单类别 F1 分数公式：

$$\text{F1}_i = 2\times\frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

宏平均 F1 分数公式：

$$\text{Macro F1} = \frac{1}{C}\sum_{i=1}^{C}\text{F1}_i$$

上述公式中，C 为数据集的类别总数(本研究中 $C=2$)；TP_i 为第 i 类的真阳性样本数(被正确预测的样本)；FP_i 为第 i 类的假阳性样本数(其他类样本被错误预测为第 i 类)；FN_i 为第 i 类的假阴性样本数(第 i 类样本被错误预测为其他类)。

4.3. 实验结果分析

宏平均精确率(Macro Precision)、宏平均召回率(Macro Recall)与宏平均 F1 分数(Macro F1)是值越大越优的指标，以特征个数为 x 轴，性能度量指标值为 y 轴，将五种特征选择方法的性能结果可视化。观察发现曲线在图中出现交叉点，为了更客观的比较各种方法的优劣，计算出各曲线下的面积(Area Under the Curve, AUC)。

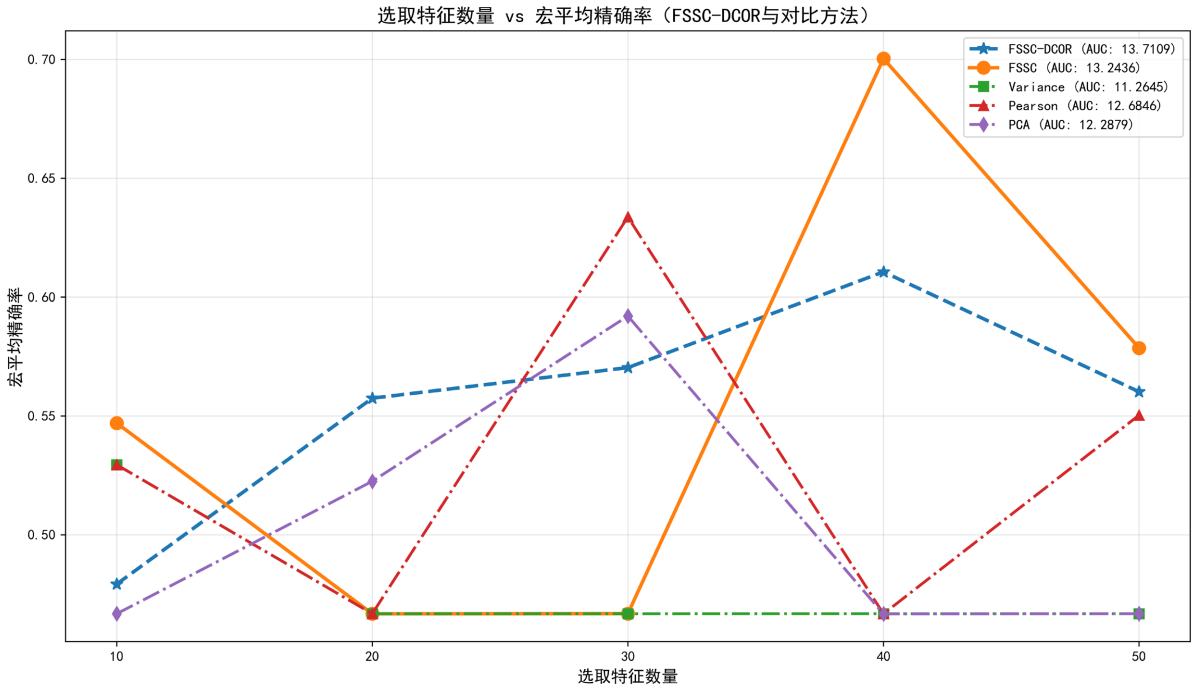


Figure 1. Comparison of macro-averaged precision between FSSC-DCOR and comparative algorithms under different numbers of selected features

图 1. 不同特征选择数量下 FSSC-DCOR 与对比算法的宏平均精确率对比

由图 1 可知，FSSC-DCOR 在宏平均精确率性能上的近似 AUC (13.7109)在所有算法中最高，显著优于 FSSC (13.2436)、Pearson (12.6846)等对比方法；因此 FSSC-DCOR 算法在宏平均精确率指标上展现出更优的综合性能、更强的特征规模适配性与更稳定的表现。

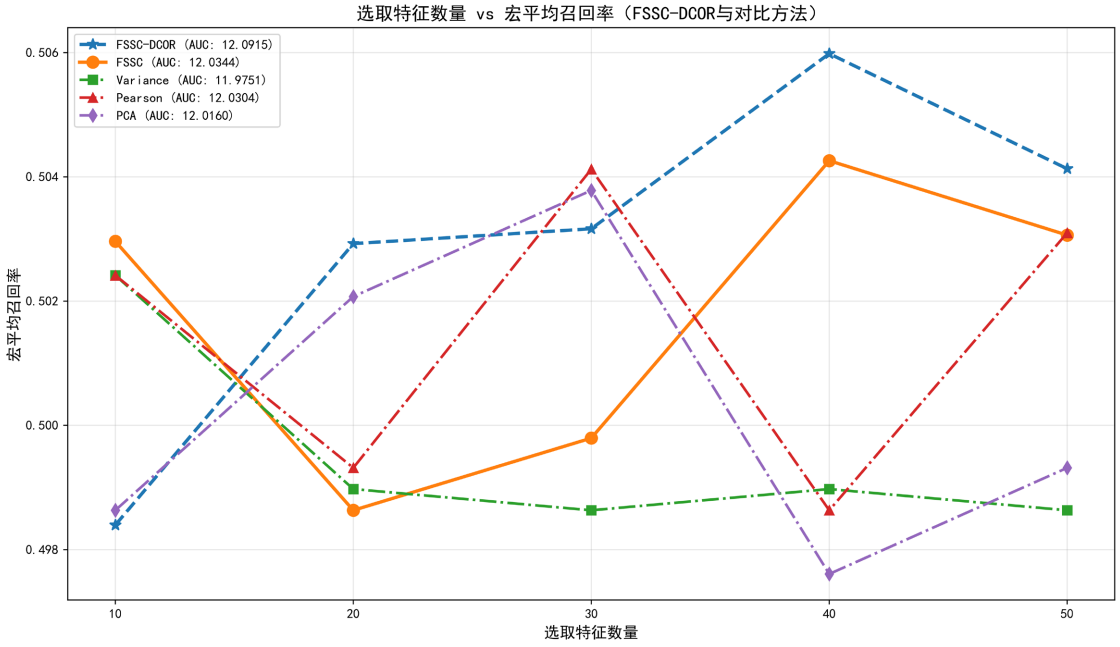


Figure 2. Comparison of macro-averaged recall between FSSC-DCOR and comparative algorithms under different numbers of selected features
图 2. 不同特征选择数量下 FSSC-DCOR 与对比算法的宏平均召回率对比

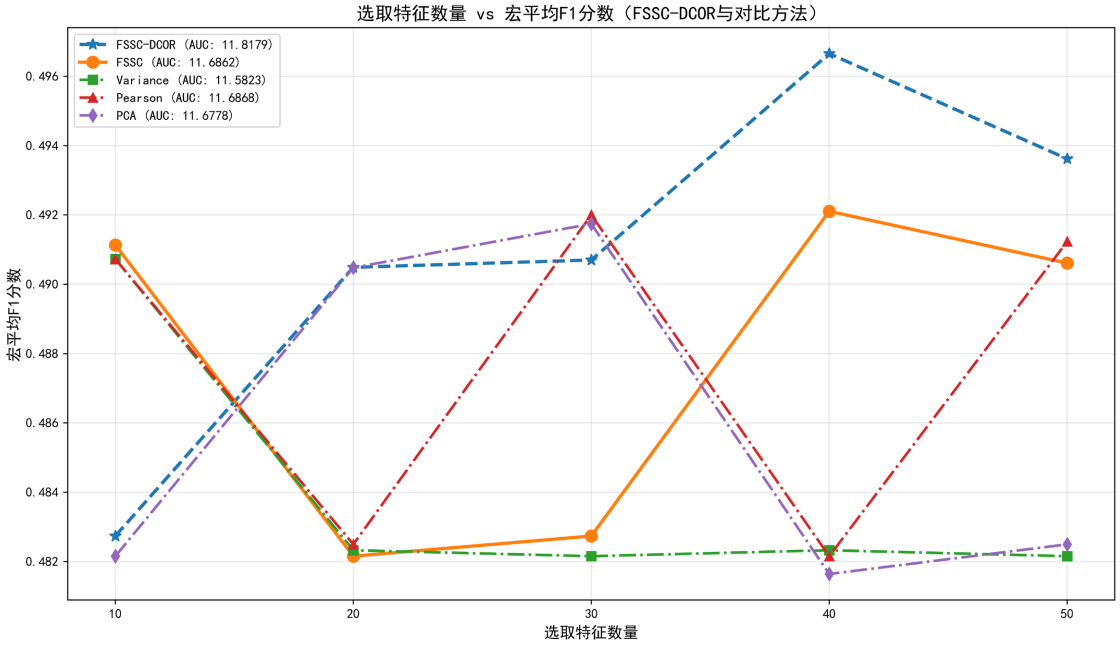
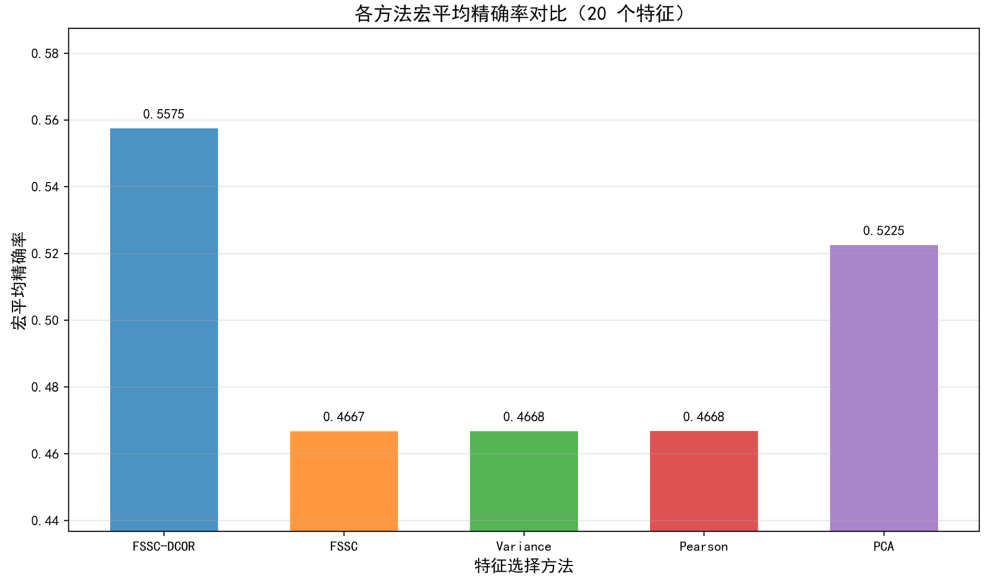


Figure 3. Comparison of macro-F1 scores between FSSC-DCOR and comparative algorithms under different numbers of selected features
图 3. 不同特征选择数量下 FSSC-DCOR 与对比算法的宏 F1 分数对比

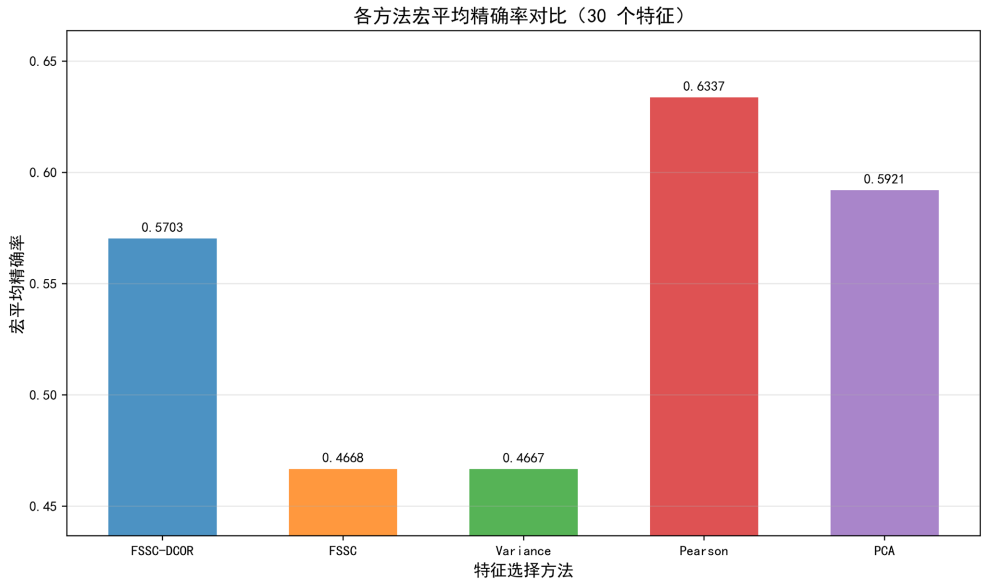
由图2可知，FSSC-DCOR在宏平均召回率性能上的近似AUC (12.0915)在所有算法中处于领先水平，显著优于Variance (11.9751)、PCA (12.0160)等对比方法；因此，本文所提的特征选择算法在SECOM半导体数据集上表现更优，所选择的特征子集效果更好。

由图3可知，FSSC-DCOR算法在宏F1性能上全面优于FSSC、Variance、Pearson及PCA算法：其近似AUC (11.8179)为所有算法中最高，体现了全特征数量梯度下的综合性能领先；同时FSSC-DCOR的性能曲线更稳定，受特征数量变化的干扰小于Pearson等算法。

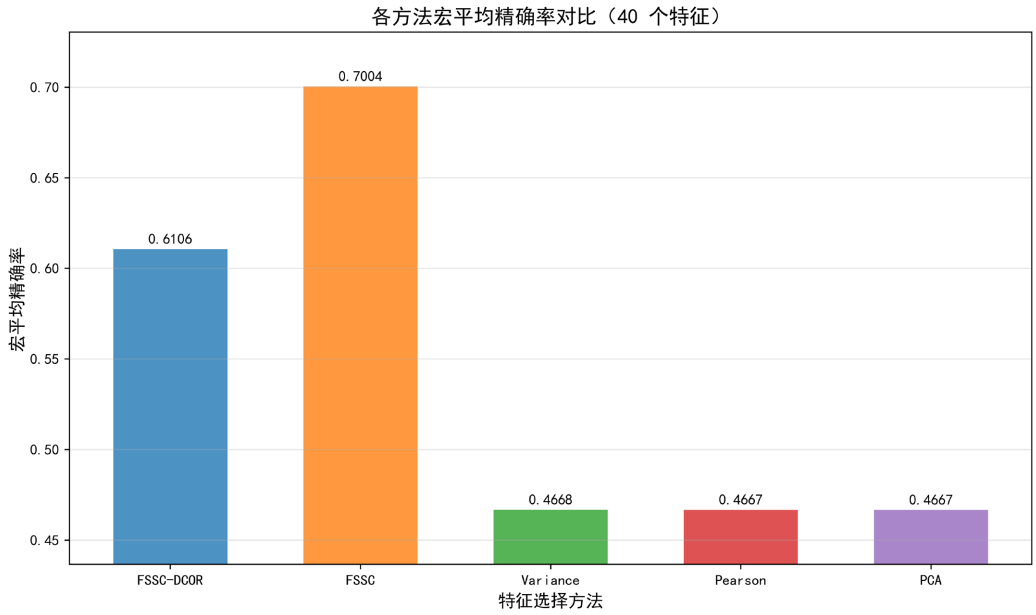
为了进一步验证本文提出的基于谱聚类与距离相关系数的无监督特征选择模型(FSSC-DCOR 算法)在SECOM半导体数据集上筛选的特征子集是否具备更优性能，本研究固定特征选取数量，设置10组不同随机种子并开展10次独立实验，以宏平均精确率、宏平均召回率及宏平均F1分数为核心评价指标，系统对比FSSC-DCOR算法与其他对比特征选择算法在该数据集上的综合性能表现。



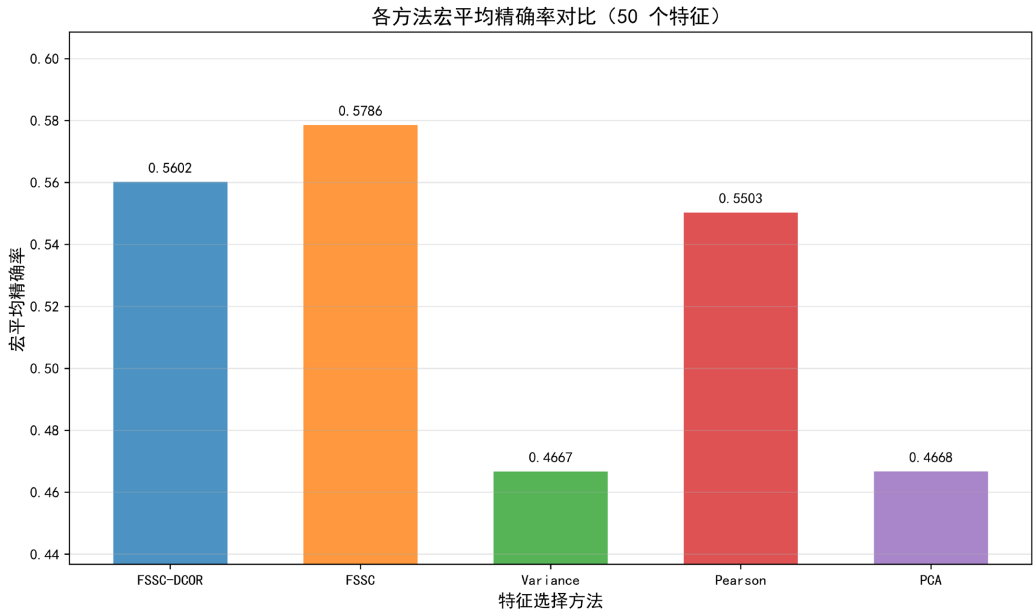
(a)



(b)



(c)



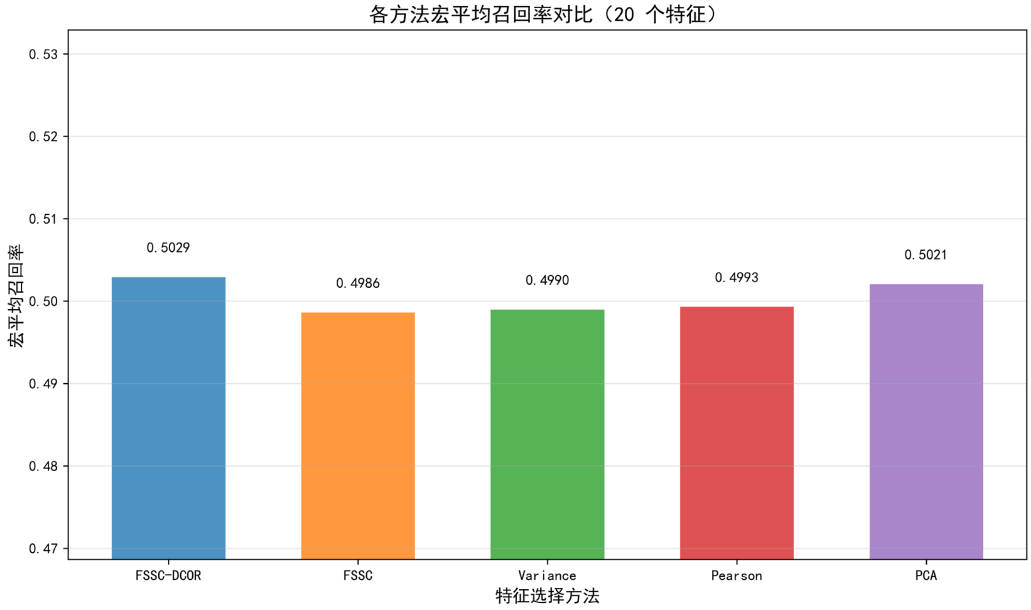
(d)

Figure 4. Performance of macro-averaged precision in the bar chart under a fixed number of selected features
图 4. 固定特征数量下宏平均精确率柱状图表现

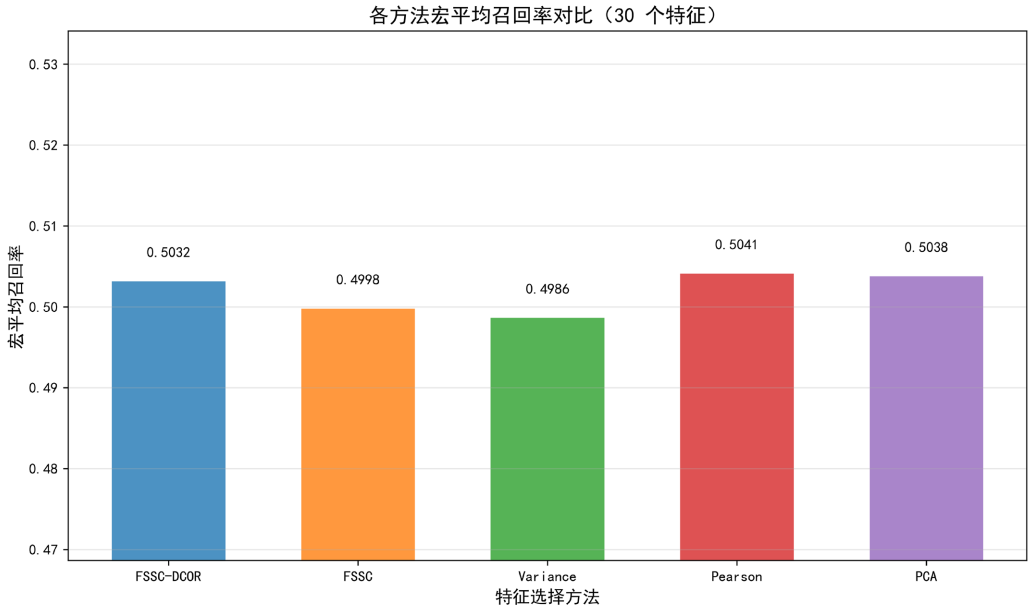
由图 4 可知，通过 4 组柱状图分别呈现了特征数量为 20、30、40、50 时各算法在 SECOM 半导体数据集上的宏平均精确率表现，在特征数为 20 时，FSSC-DCOR 的宏平均精确率(0.5575)显著高于其余算法，是该特征数量下的最优算法；特征数增至 30 时，FSSC-DCOR 的精确率(0.570)虽略低于 Pearson (0.6337)，但仍优于 Variance 与 FSSC；即便是在 FSSC 表现突出的 40、50 特征场景中，FSSC-DCOR 的精确率(0.6106, 0.5602)也稳定处于中等偏上水平，且远高于 Variance、PCA 等算法的低水平表现。整体而言，FSSC-DCOR 在特征数量较少的场景下具备明显的精确率优势，且在特征数量增加时仍能维持稳定且优

于多数对比算法的性能。

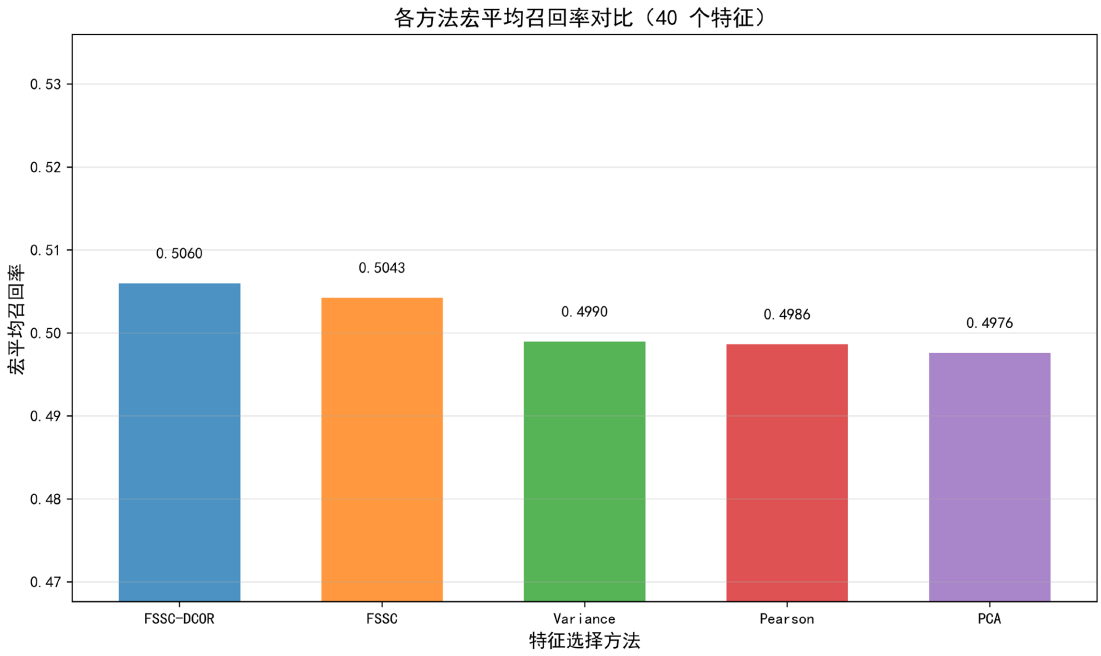
由图 5 可知，通过 4 组柱状图分别呈现了特征数量为 20、30、40、50 时各算法的宏平均召回率在 SECOM 半导体数据集上的表现，在 20 个特征场景下，FSSC-DCOR 的宏平均召回率(0.5029)为所有算法中最高；30 个特征时 FSSC-DCOR 的召回率(0.5032)与 PCA (0.5038)接近，且优于 FSSC、Variance 等算法；40 个特征与 50 个特征场景中，FSSC-DCOR 的召回率(0.5060, 0.5041)均稳定处于最优水平，显著高于 Variance (0.4990, 0.4986)、PCA (0.4976, 0.4993)等算法的表现。整体来看，FSSC-DCOR 在不同特征数量下的宏平均召回率始终保持稳定且领先的水平，体现出其在特征选择任务中对少数类样本的识别能力更优，是各场景下召回率表现最可靠的算法。



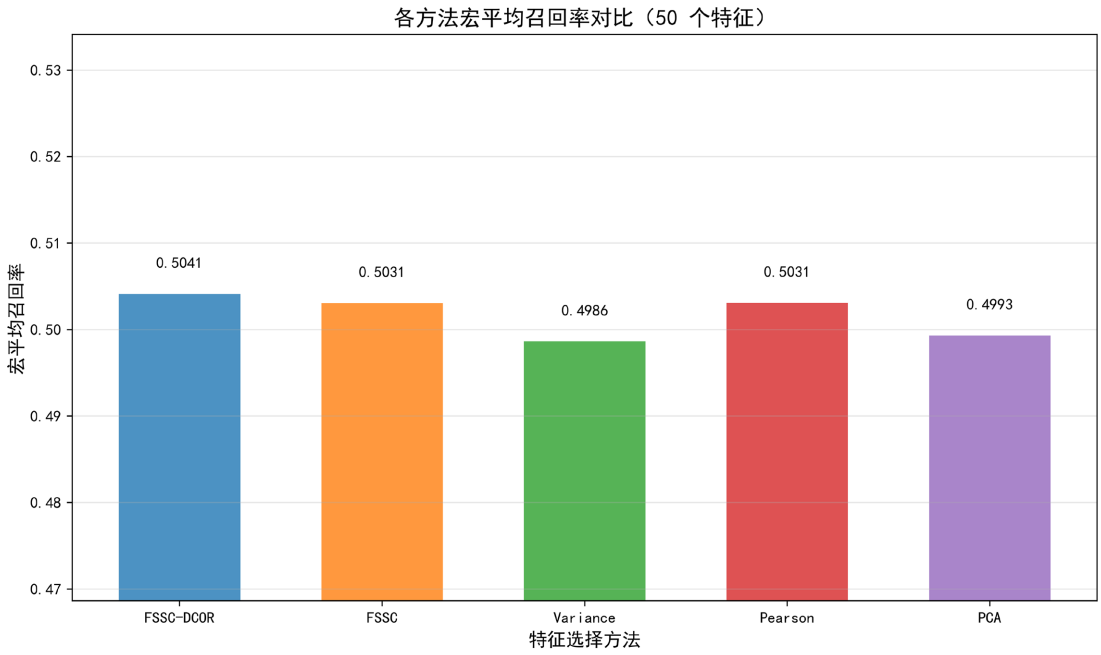
(a)



(b)



(c)

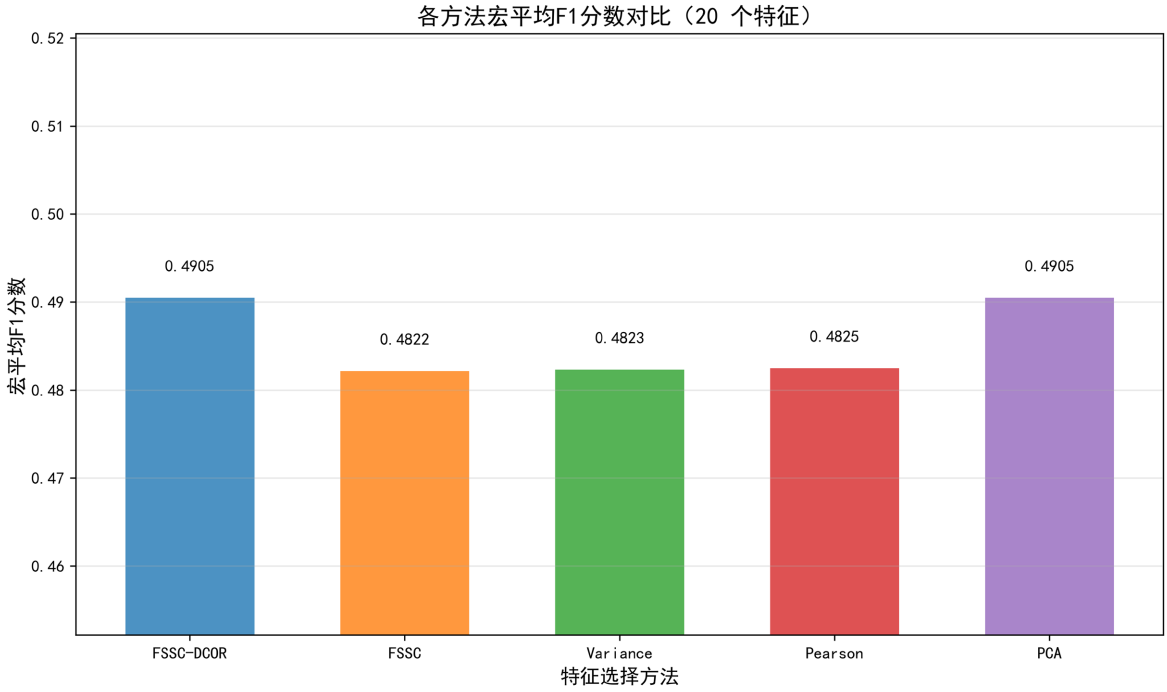


(d)

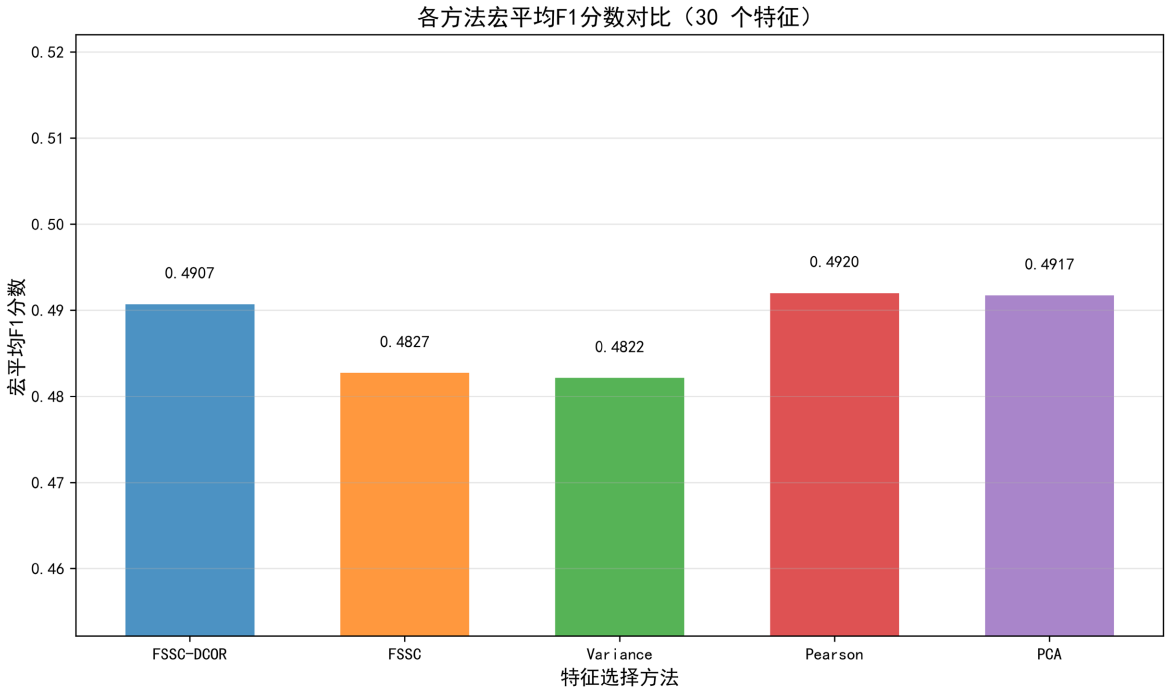
Figure 5. Performance of macro-averaged recall under a fixed number of selected features
图 5. 固定特征数量下宏平均召回率的表现

由图 6 可知，通过 4 组柱状图分别呈现了特征数量为 20、30、40、50 时各算法的宏平均 F1 分数在 SECOM 半导体数据集上的表现，20 个特征下，FSSC-DCOR 的宏平均 F1 分数(0.4905)为所有算法中最高；30 个特征时，其 F1 分数(0.4907)虽略低于 Pearson (0.4920)，但仍优于 FSSC、Variance；40 个与 50 个特征场景中，FSSC-DCOR 的 F1 分数(0.4967, 0.4936)均稳居首位，显著高于其余算法的表现。

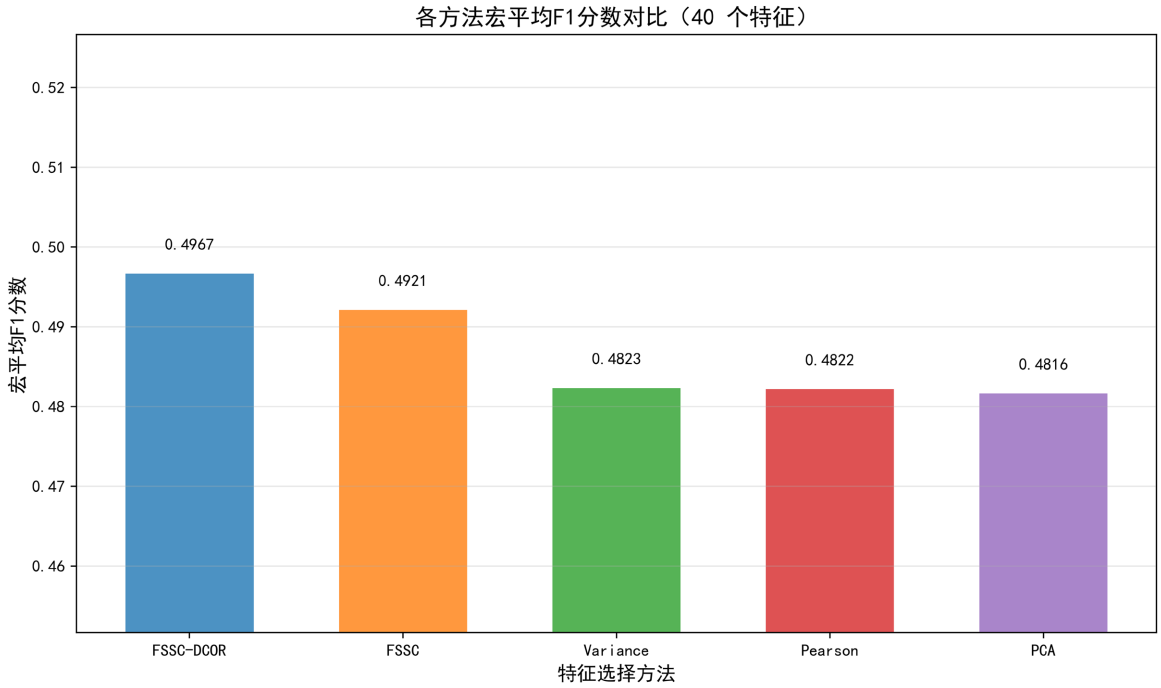
综合图 1~6 的实验结果可知：本文提出的 FSSC-DCOR 算法在 SECOM 半导体数据集上的表现显著优于 FSSC、Variance、PCA 及 Pearson 四种对比算法。从特征子集的性能来看，FSSC-DCOR 筛选得到的特征子集在宏平均精确率、宏平均召回率与宏平均 F1 分数三项核心指标上，均优于其余四种对比算法所生成的特征子集，验证了 FSSC-DCOR 在该数据集特征选择任务中的综合最优性。



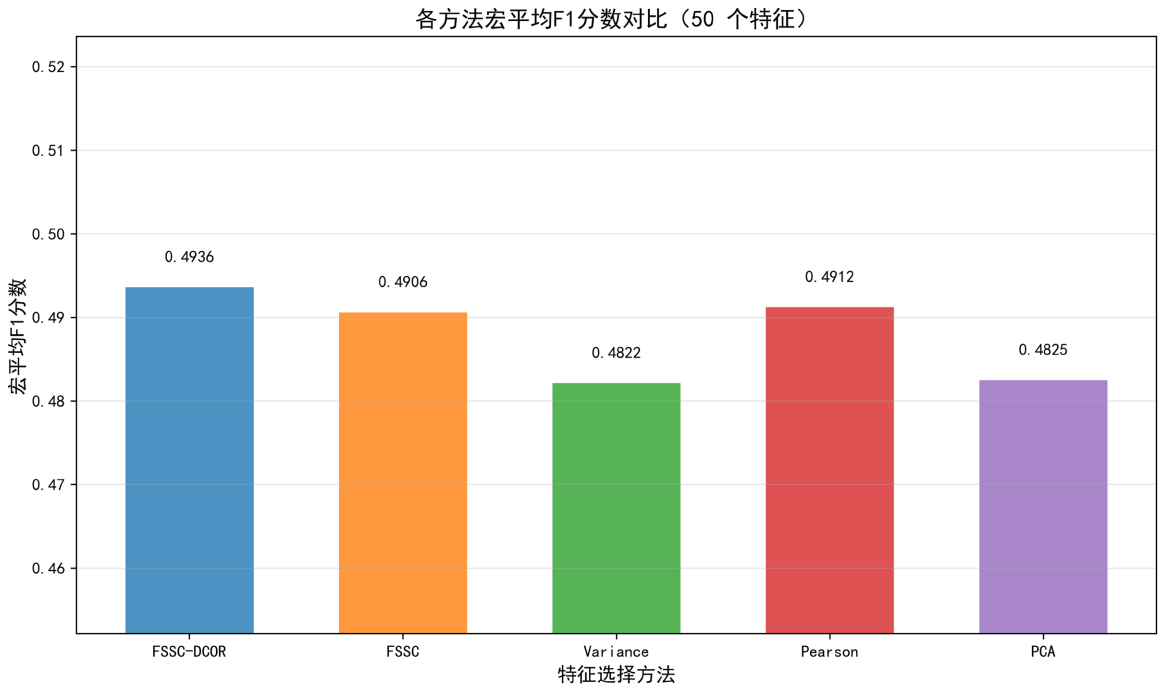
(a)



(b)



(c)



(d)

Figure 6. Performance of macro-F1 scores in the bar chart under a fixed number of selected features
图 6. 固定特征数量下宏平均 F1 分数的柱状图表现

5. 总结

针对半导体高维制造数据高维、强非线性耦合且标签标注成本高的特点，本文采用无监督特征选择

算法 FSSC-DCOR 进行特征筛选。实验结果表明，该算法在半导体相关高维数据集上表现优异：能够筛选出兼具强分类能力与低冗余性的特征子集，有效提升后续数据处理效率；经统计显著性检测验证，该算法与 FSSC、PCA、方差阈值法、皮尔逊相关系数法等对比算法存在显著差异，在宏平均精确率、宏平均召回率及宏平均 F1 分数三项核心分类指标上均展现出更优性能，充分适配半导体高维制造数据的特征选择需求。

本文的研究价值体现在：(1) 聚焦半导体制造场景标签稀缺的核心痛点，构建无监督特征选择框架，无需标注数据即可实现高维冗余数据降维，契合工业实际应用需求；(2) 融合谱聚类的全局结构挖掘能力与距离相关系数的非线性冗余量化优势，弥补传统无监督方法“重结构轻冗余”或“重冗余轻结构”的单一缺陷；(3) 为半导体高维无标签数据的预处理提供高效可靠的方案，助力企业提升后续数据分析效率，降低质量管控的计算成本，优化晶圆制造全流程的数据驱动决策效果。

基金项目

余青青的研究受到重庆工商大学研究生创新型科研项目资助(项目编号：yjscxx2025-269-23)。

参考文献

[1] Nuhu, A.A., Zeeshan, Q., Safaei, B. and Shahzad, M.A. (2022) Machine Learning-Based Techniques for Fault Diagnosis in the Semiconductor Manufacturing Process: A Comparative Study. *The Journal of Supercomputing*, **79**, 2031-2081. https://doi.org/10.1007/s11227-022-04730-x

[2] 程云飞, 周丽芳, 赵波, 等. 特征提取及数据扩充的 GA-LightGBM 半导体质量检测方法[J]. 重庆邮电大学学报(自然科学版), 2024, 36(2): 351-356.

[3] 柳嘉昊. 基于 KMUS-RF 算法的复杂产品关键质量特性识别研究[J]. 中小企业管理与科技(下旬刊), 2021(10): 134-137.

[4] Gomez-Sirvent, J.L., de la Rosa, F.L., Sanchez-Reolid, R., Fernandez-Caballero, A. and Morales, R. (2022) Optimal Feature Selection for Defect Classification in Semiconductor Wafers. *IEEE Transactions on Semiconductor Manufacturing*, **35**, 324-331. https://doi.org/10.1109/tsm.2022.3146849

[5] He, Q.P. and Wang, J. (2007) Fault Detection Using the K-Nearest Neighbor Rule for Semiconductor Manufacturing Processes. *IEEE Transactions on Semiconductor Manufacturing*, **20**, 345-354. https://doi.org/10.1109/tsm.2007.907607

[6] Baek, M. and Kim, S.B. (2023) Failure Detection and Primary Cause Identification of Multivariate Time Series Data in Semiconductor Equipment. *IEEE Access*, **11**, 54363-54372. https://doi.org/10.1109/access.2023.3281407

[7] Qian, X., Sun, T., Wang, B. and Zhang, Y. (2023) A Weighted KNN Fault Detection Based on Multistep Index and Dynamic Neighborhood Scale under Complex Working Conditions. *IEEE Access*, **11**, 49183-49192. https://doi.org/10.1109/access.2023.3272001

[8] Kuo, T., Hong, T. and Chen, L. (2025) Sustainable Fault Detection and Process Simulation in Semiconductor Manufacturing Using Machine Learning and Life Cycle Assessment. *Computers & Industrial Engineering*, **210**, Article ID: 111584. https://doi.org/10.1016/j.cie.2025.111584

[9] López de la Rosa, F., Gómez-Sirvent, J.L., Morales, R., Sánchez-Reolid, R. and Fernández-Caballero, A. (2023) Defect Detection and Classification on Semiconductor Wafers Using Two-Stage Geometric Transformation-Based Data Augmentation and Squeezenet Lightweight Convolutional Neural Network. *Computers & Industrial Engineering*, **183**, Article ID: 109549. https://doi.org/10.1016/j.cie.2023.109549

[10] Jiao, S., Yang, W., Wu, C., Li, Y. and Xue, B. (2025) Mixed-Type Micro-Defect Detection in Semiconductor Wafers: A Dual-Modal Feature Real-Time Detection Approach via Optical Topography and Lightweight Classification Network. *Engineering Applications of Artificial Intelligence*, **160**, Article ID: 111838. https://doi.org/10.1016/j.engappai.2025.111838

[11] 闫伟, 何桢, 田文萌, 等. 基于 IG 的复杂产品关键质量特性识别[J]. 工业工程与管理, 2012, 17(1): 70-74, 83.

[12] 李岸达, 何桢, 何曙光. 基于 Filter 与 Wrapper 的复杂产品关键质量特性识别[J]. 工业工程与管理, 2014, 19(3): 53-59.

[13] Lee, D., Yang, J., Lee, C. and Kim, K. (2019) A Data-Driven Approach to Selection of Critical Process Steps in the Semiconductor Manufacturing Process Considering Missing and Imbalanced Data. *Journal of Manufacturing Systems*, **52**, 146-156. https://doi.org/10.1016/j.jmsy.2019.07.001

[14] 李航. 机器学习方法[M]. 北京: 清华大学出版社, 2022.

[15] von Luxburg, U. (2007) A Tutorial on Spectral Clustering. *Statistics and Computing*, **17**, 395-416. https://doi.org/10.1007/s11222-007-9033-z

[16] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[17] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007) Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, **35**, 2769-2794. https://doi.org/10.1214/009053607000000505

[18] Li, R., Zhong, W. and Zhu, L. (2012) Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association*, **107**, 1129-1139. https://doi.org/10.1080/01621459.2012.695654

[19] 谢娟英, 丁丽娟, 王明钊. 基于谱聚类的无监督特征选择算法[J]. 软件学报, 2020, 31(4): 1009-1024.

[20] Murphy, P. and Aha, D. (2008) UCIML Repository. https://archive.ics.uci.edu/ml/datasets/SECOM