



上海对外经贸大学

SHANGHAI UNIVERSITY OF INTERNATIONAL BUSINESS AND ECONOMICS

## 题目：机器学习之特征降维技术研究

系：统计与信息学院

课程名称：人工智能

任课教师：胡光

学生姓名：陶盛皿


学号：19024075

专业：大数据 1901

## 考试诚信承诺书

本人郑重承诺：在 2021-2022 学年第二学期课程期末考试中，严格遵守学校《学生考试规定》，独立完成考试（论文、报告、作业等），不违纪，不作弊，如有违反，按学校规定接受处理。

学生签名：



日期：2022 年 5 月 26 日

## 摘要

本文对机器学习中的特征降维技术展开研究与论述。特征降维技术分为特征选择和特征提取算法两类，针对特征选择算法以分枝定界法为例的完全搜索算法、以增 1 减 r 法为例的启发式搜索算法、以遗传算法为例的随机搜索算法分析算法流程和效率，并以多维尺度算法为例对特征提取算法论述。接着，对以上算法进行应用分析。最后论述机器学习领域特征降维技术发展现状和趋势，得出即使在深度学习时代，该技术仍十分重要的结论。

**关键词：**特征选择，特征提取，分枝定界算法，增 1 减 r 算法，遗传算法，多维尺度法，深度学习

# 题目：机器学习之特征降维技术研究

## 引言

机器学习是人工智能的核心之一，其核心研究内容是：通过对数据的模式识别和学习来获取经验和知识，由此模拟人类的智能实现或执行特定问题上的任务，在此过程中机器通过不断迭代计算以更新模型或策略参数，使之不断改善自身算法的性能<sup>1</sup>。从机器学习定义可知，对于一个机器学习问题，智能行为的决策以及结果取决于模式识别过程中学习到的知识，与元数据和数据特征直接相关，故数据和特征决定了机器学习的上限，而算法或策略使其逼近机器学习上限。由此可见，数据和特征在模型的整个开发过程中十分重要。对于数据和特征，特征降维是主要方法，特征降维的方法分为两类：特征选择和特征提取。本文从两类实现特征降维的两个方法的算法展开论述，引入算法包括：以分枝定界法为例的完全搜索算法、以增 1 减 r 法为例的启发式搜索算法、以遗传算法为例的随机搜索算法和特征提取的多尺度算法，并进行算法应用分析。最后，分析机器学习中特征降维（包括：特征选择和特征提取）在深度学习时代发展现状与趋势。

## 一、概念和理论

### （一）机器学习

机器学习是人工智能的一个分支，旨在通过数据和经验中计算最大后验概率进行决策从而模拟人类智能。发展至今机器学习包含：传统机器学习和深度学习两

个分支，学术界统称为机器学习，基于机器学习发展出的技术目前已广泛应用于模式识别、计算机视觉、自然语言处理等计算机科学领域，航空航天、气象卫星、工农业等工业界，金融、互联网等服务业，以及生物医学领域。<sup>2</sup>该领域有三个重点：以提高算法在特定任务或应用场景为目标的工程，以通过计算来实现人机交互和决策的认知模拟，以及从理论层面的机器学习算法研究。<sup>3</sup>

## （二）特征降维

数据特征的质量直接影响机器学习模型准确度，模式识别是实现机器学习的基础，在模式识别中特征降维是重点之一。尤其针对大数据的海量、高速、多样、价值的“4V”高维特征，需要使用特征降维技术避免“维度灾难”问题<sup>4</sup>。特征降维技术分为特征提取和特征选择两类方法。

### 1. 特征选择

特征选择是指从给定全部特征中选择若干特征以最优化机器学习任务指标，提高算法性能,是机器学习任务中数据预处理阶段的降维技术的方法之一。

### 2. 特征提取

特征提取从元数据中提取非冗余的特征（信息），以提高机器学习任务更好的可解释性和模型的泛化能力。

## 二、降维技术的典型算法

### （一）特征选择

特征选择算法分为基于数据层面的特征选择算法和基于传统机器学习划分的特征选择算法。基于传统机器学习层面，特征选择算法分为：监督、无监督、基于策略的包装、过滤、嵌入方法<sup>5</sup>。

#### 1. 完全搜索算法：以分枝定界法为例

假定当特征空间为  $D$  维，一个有效的特征选择算法需要通过某种判定准则，选择具有  $d$  个特征的最优组合，通常  $d < D$ ，故该问题可归为搜索问题，利用穷举法可知一共有： $C_D^d = \frac{D!}{(D-d)!d!}$  种可能性，故易知暴力搜索算法效率太低。分枝定界算法是一种优化搜索的算法，由于自顶向下地从所有候选特征开始逐步去除特征，故是回溯的方法<sup>6</sup>。基本思想是：将候选特征集构建一个树形结构，并在树搜索过程融入贪心的思想进行剪枝。

### 算法流程：

0. 初始化层数  $L=0$ ，结点位置  $i=0$ ，特征候选集包含全部特征记为  $D$ 。

#### **Repeat:**

1. 递归建树，对于第  $L$  层的第  $i$  个结点，记候选集为  $D_{li}$ 。在该层，若去掉某个特征  $a(a \in D_{li})$  准则函数损失最大，则认为这个特征是最不可能去除的。
2. 根据准则函数计算对应特征的损失值，在第  $L$  层进行特征排序，选取若干特征该结点的子结点，记该结点的第  $j$  个子结点的候选集为  $D_{l+1,j}$ 。  
( $D_{l+1,j} = D_{li} - \text{第}j\text{个结点的特征} - \text{第}l\text{层}i\text{结点左侧所有特征}$ )。

**Until:**  $D_{l+1} = \emptyset$ 。

3. 计算最后一层第  $k$  个叶子结点的准则函数值，记为界限值  $B$ 。

#### **Repeat:**

4. 向上回溯若遇到分枝结点则向下搜索分枝结点左侧的孩子结点，计算结点准则函数值  $B'$ ，若  $B' < B$ ，则剪枝，将  $B'$  作为新界限值。

**Until:** 回溯至 root，且根据  $B'$  不能在向下搜索，算法停止。

即根据  $B'$  选择保留的最优特征。

算法效率：最好情况是二叉树结构  $O(\log n)$  是下界，故分枝定界法时间复杂度为  $\omega(\log n)$ 。

## 2. 启发式搜索算法：以增 l 减 r 法为例

分枝定界法是完全搜索算法，虽然使用到贪心思想进行剪枝算法效率由于暴力搜索算法，但时间复杂度还是较大。基于启发式搜索算法的增 l 减 r 法进一步对特征选择算法进行优化。

该算法折中了顺序前进法（sequential forward selection, SFS）和顺序后退法（sequential backward selection, SBS），融合了回溯和局部最优解的思想，分为自底向上和自顶向下两种搜索模式<sup>7</sup>。

### 算法流程：

0. 初始化：设选择特征值  $S=0$ ，人为规定最终特征个数  $S'$  判断自顶向下还是自底向上，if 自底向上，then  $l > r$ 。

### Repeat:

1.  $S += 1$ ，逐步增选 1 个特征，再  $S -= r$  逐步剔除  $r$  个与其他特征组合起来准则函数值最小的特征。

**Until:**  $S == S'$ , 算法终止。（若自顶向下，则反之初始化  $S = |D|$ ,  $l < r$ 。）

该算法相较于分枝定界法的改进之处在于一次选取多个特征，并考虑了特征之间的相关性。算法效率：时间复杂度  $O(\log(\binom{l-r}{r}\sqrt{n}))$ ，最差情况为  $O(\log(\sqrt[n]{n}))$ 。

## 3. 随机搜索算法：以遗传算法为例

遗传算法缘起于生物学中染色体具有一定概率（随机性）的变异。基于随机搜索的遗传算法属于全局寻优算法，根据最目标函数匹配到最优解，故算法效率由于前述两种特征选择算法。基于遗传算法的特征选择是一种包装方法(wrapper)，该算

法是二分类作为特征选择判断依据。在遗传算法中，特征选择有 0-1 分布的等概率性，故对所选择的特征用二进制字符串来初始化，

人为设定选取  $d$  ( $d < D$ ) 个特征，假设 0 代表特征未选中，1 代表特征选中，即染色体  $m$  为长度为  $|D|$  的 0-1 字符串。并定义目标函数：适应度函数  $fitness$ ，适应度值为  $f(m)$ ，选择概率记为  $p(f(m))$ 。

### 算法流程：

0. 初始化：  $t=0$ ，染色体  $m$  为长度为  $|D|$  的全 0 字符串，适应度值为  $f(m_0)$ 。

随机产生包含  $L$  条染色体的种群  $M(0)$ 。

### Repeat:

1. 计算每一条染色体的适应度值  $f(m_t)$

2. 按照概率  $p(f(m))$  进行采样，  $t += 1$ , 更新染色体  $m_t$ ，更新种群  $M(t)$ 。

3. Return to 1，计算适应度值  $f(m_t)$

Until:  $m$  中 1 的个数为  $d$  或者  $f(m_t)$  达到阈值，算法终止。

## (二) 特征提取：以多维尺度法为例

多维尺度法核心思想简单概括就是将高维样本按比例缩放或映射到低维空间生成对样本的低维表示，是融合了线性代数投影矩阵的思想的统计学研究方法。

### 基本步骤：

0. 根据数据类型：定量或定性数据，选择度量型多维尺度分析或非度量型多维尺度分析。

1. 根据评价标准，确定合适的维数，目标：以较少的维数空间为前提，去优化数据拟合水平。



2. 结果评估：计算拟合优劣度、压力值、梯度收敛值和单调收敛值。（以论述步骤为主，具体计算过程复杂，不展开赘述。）
3. 若为度量型多维尺度分析，如此迭代直到算法收敛。

### 三、应用分析

#### （一）分枝定界法：以在 Graph-SLAM 技术中的应用为例

分枝定界法适用于运筹学中的约束优化问题。SLAM (simultaneous localization and mapping) 是实时定位与地图构建,指智能体(Agent)从未知起始点出发,利用多种传感器探测周围环境,并在行进过程中实时 2D 或者 3D 建模渲染地图,以确定当前位置。Graph-SLAM 是图数据结构,结点代表 Agent 采样时刻位置,有向边表示路径。通过可以采用分枝定界算法可以进行后端优化,提升 Agent 实时建模效率,最优提升可达 50%<sup>8</sup>。

#### （二）遗传算法：以智能体路径规划为例

机器人路径规划是遗传算法传统应用场景,智能体(Agent)根据自身传感器对环境的感知,自行规划路线。智能体(Agent)路径规划之前首先要建立地图,采用栅格法建立智能体(Agent)行走空间。设智能体(Agent)工作空间为 2D 空间,使用栅格划分该空间,形成具有坐标的图数据结构。坐标即为栅格设定了序,以此坐标形式优化传统遗传算法的二进制字符串编码形式,引入插入算子和删除算子以保证染色体的概率变异<sup>9</sup>,在此基础上迭代适应度目标函数,结果表明使用了遗传算法的智能体(Agent)路径规划应用场景中,算法稳定性和收敛效率都得到了大幅度提升。

### （三）多维尺度法：以计算机视觉中应用为例

目前多尺度在图像处理方面有很多的应用，诸如图像融合，细节增强，SIFT 算法等。所谓多尺度即多种粒度，是对模式识别信号进行多粒度的采样，因为通常在不同的尺度下可以提取到不同的特征。在数字图像处理中，会为了增强图像的特征以及提高对于小目标的识别能力，通常采用的图像金字塔方式是基于多尺度法的进行图像叠加实现“特征融合机制”<sup>10</sup>，增强深度网络模式识别图像语义信息（特征）的提取，经过实验，融入了多维尺度法的计算机视觉任务能够显著提升目标检测定位的精度，以及算法准确度。

## 四、发展现状和趋势：特征降维与深度学习

从 2000 年初起深度学习兴起以来，机器学习领域从传统机器学习迈向了深度学习，领域以文献数量与专利数量呈逐年上升趋势，故以此标准划分发展三阶段：2001 年前的萌芽期、2001 至 2010 年的平稳成长期、2011 年至 2018 年的快速成长期，整个领域发展迅速<sup>11</sup>。深度学习从繁复冗杂的人工特征工程解放了人力，同时又在特征任务上有效减轻“维度灾难”问题。在于深度学习能够通过空间网状结构的特性发现数据的分布式特征表示，即自动组合低维特征形成高阶属性与类别，某种程度上实现特征自动选择与提取<sup>12</sup>。此外，被广泛使用的注意力机制能对局部区域投入注意力提高该区域注意力权重，提取该区域的信息，并抑制对其他区域的注意力，从而无需明显的训练信号也能跟踪注意力目标实现“自动”特征提取<sup>13</sup>。在此前提下，引出了一个问题：在深度学习时代是否还需要特征降维（包括：特征选择和特征提取）？

首先，深度学习优化了特征降维的线性映射拟合，以多层非线性映射的算法能拟合复杂函数，但对隐藏层参数敏感并且隐藏层作为“黑盒”存在可解释性不强，存在潜在的缺点：过拟合，尤其是对于稀疏特征形成的高秩矩阵<sup>14</sup>，而特征降维能

提高可解释性并提升模型泛化能力。针对该问题有学者提出将特征选择算法之一的遗传算法融入深度学习，利用其特征选择全局寻优特性优化了深度学习参数<sup>15</sup>。故深度网络融入特征降维逻辑单元的解决方案经过实验能够显著改善过拟合并提高算法性能<sup>16</sup>。综上，特征降维具有两个显著优点：能更好的表示业务逻辑的可解释性和提高算法性能<sup>17</sup>，在深度学习时代特征降维技术仍被需要，在未来面对更为海量的稀疏大数据时，该技术尤为重要。

## 五、参考文献

- 
- <sup>1</sup> Mitchell, T.M. Machine Learning[DB/OL]. McGraw-Hill, New York,1997.
  - <sup>2</sup> El Naqa, I., Murphy, M.J. What Is Machine Learning?[J]. Machine Learning in Radiation Oncology. Springer, Cham, 2015. [https://doi.org/10.1007/978-3-319-18305-3\\_1](https://doi.org/10.1007/978-3-319-18305-3_1).
  - <sup>3</sup> Jaime G. Carbonell, Ryszard S. Michalski, Tom M. Mitchell. AN OVERVIEW OF MACHINE LEARNING[M]. Machine Learning, Morgan Kaufmann, 1983: 3-23. <https://doi.org/10.1016/B978-0-08-051054-5.50005-4>.
  - <sup>4</sup> Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: Data mining, inference and prediction[J]. Math, 2005: 83–85.
  - <sup>5</sup> Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu.Feature Selection: A Data Perspective[J]. ACM Comput, 2017. <https://doi.org/10.1145/3136625>.
  - <sup>6</sup> 拉塞尔, S.J.), 诺维格,等. 人工智能:一种现代的方法(第 3 版)[M]. 清华大学出版社, 北京, 2011: 68 – 73.
  - <sup>7</sup> 张学工. 模式识别: 第 3 版[M]. 清华大学出版社, 北京, 2021: 154 – 156.
  - <sup>8</sup> 李敏,王英建,刘晓倩.基于深度优先搜索分支定界法的 Graph-SLAM 后端优化算法改进[J].自动化技术与应用,2018,37(09):4-8.
  - <sup>9</sup> 孙树栋,曲彦宾.遗传算法在机器人路径规划中的应用研究[J].西北工业大学学报,1998(01):85-89.

- 
- <sup>10</sup> 王平, 江雨泽, 赵光辉. 目标检测的多尺度定位提升算法[J]. 西安电子科技大学学报, 2021, Vol.48, Issue(3): 85 -90. doi: 10.19665/j.issn1001-2400.2021.03.011.
- <sup>11</sup> 黄鲁成, 薛爽. 机器学习技术发展现状与国际竞争分析[J]. 现代情报, 2019, 39(10):165-176.
- <sup>12</sup> 马利星, 胡敏. 特征工程研究领域发展趋势的可视化分析[J]. 北京信息科技大学学报(自然科学版), 2020, 35(04):32-37.DOI:10.16508/j.cnki.11-5866/n.2020.04.006.
- <sup>13</sup> Mnih V, Heess N, Graves A. Recurrent models of visual attention[J]. Advances in neural information processing systems, 2014, 27.
- <sup>14</sup> Fang, H., Guo, G., Zhang, D., Shu, Y. Deep Learning-Based Sequential Recommender Systems: Concepts, Algorithms, and Evaluations[J]. Web Engineering. ICWE, 2019. Lecture Notes in Computer Science(), vol 11496. [https://doi.org/10.1007/978-3-030-19274-7\\_47](https://doi.org/10.1007/978-3-030-19274-7_47).
- <sup>15</sup> 陈珍, 夏靖波, 柏骏, 徐敏. 基于进化深度学习的特征提取算法[J]. 计算机科学, 2015, 42(11):288-292.
- <sup>16</sup> Uddin M F, Riizvi S, Razaque A. Proposing logical table constructs for enhanced machine learning process[J]. IEEE Access, 2018, 6: 47751 – 47769.
- <sup>17</sup> 产品经理的 AI 知识库. 特征工程 – Feature Engineering[EB/OL]. (2021-03-02)[2022-05-24]. <https://easyai.tech/ai-definition/feature-engineering/>.