综述. 特征选择方法

MF1733062 万晨 weanl_jc@163.com

2018 年 4 月 8 日

1. 介绍

FS 在 ML 中能解决那些问题? 已经解决得怎么样了? 还有那些问题? feature construction = FS + FE 特征选择 (Feature, Variable and Attribution Selection)，是机器学习中 feature construction 的重要组成部分。在筛选原始数据，构造有用的特征集合方面，特征选择不同于特征提取 (Feature Extraction)：后者会通过线性或非线性的方式从原始数据中构造出全新的特征，具有特征学习和表示学习的能力 ["Representation Learning: A Review and New Perspectives"]；特征选择通过设计一些简单高效或者精致巧妙的方法，实现从原始特征集合中选出最优的特征子集，能够保持特征对应的原始物理意义。所谓最优特征子集，理论上定义为没有信息丢失的最小特征子集，以 Markov blanket 的形式给出 [D. Koller, Toward optimal feature selection] [C.Aliferis, Local causal and markov blanket induction for causal discovery and feature selection]；实际中理论的 ground-truth 很难找，所以经验上一般我们用预测器性能（如分类器的精度）来评估特征子集的选择结果。特征选择一直以来是一个重要的课题：为了分析特征间相关性，早期 [Blum and langley,1997,Kohavi and John,1997] 等在 1997 年提出了特征选择方面课题研究，当时大多数应用领域下特征维数还不超过 40。后续基因序列分析和 web 文本分类等典型应用不断推动特征选择课题研究：[2001, Feature selection for high-dimensional genomic microarray data] 针对高维的染色体序列数据提出了给予特征选择的基因分析方法，[2015, Deep Feature Selection: Theory and Application to Identify Enhancer and Promoters] 将深层神经网络应用到特征选择中，实现了基因 Enhancer 和 Promoter 的有效分析。目前特征选择研究有两大趋势：第一，应对各种结构化和非结构化的数据设计出一套较为通用的方法，决策树类和深层神经网络类在特征选择方面的改进是不错的解决方法，[Feature Selection via Regularized Trees] 就是通过修改单棵树的构造算法实现基于随机森林的 Ensemble 类的特征选择方法；第二，应对 curse of dimensionality，设计复杂度较低的算法，有效地处理高维数据，改进现有的算法以及组合使用一些简单的算法都是很好的思路。

在数据处理中，应用特征选择方法概括起来有如下优势：

- 可以过滤无关特征：采集过程可能引入数据噪声，从而影响后续的数据处理；同样与任务显著的无关特征 (irrelevant) 也可以认为是噪声，特征选择方法一定程度上可以过滤这一部分噪声；
- 可以剔除冗余特征：相当部分特征虽然与任务相关，但互相之间存在显著的冗余关系 (redundant)，特征选择方法可以依据实际需求选出代表性的特征，降低冗余；

- 可以实现特征重要性的评估：一些带有指标（如相关系数、权值）或其他“得分”的特征选择方法，可以在选出的特征子集中按指标或“得分”对特征进行重要性排序。

2. 过滤式与包裹式