特征选择

维基百科，自由的百科全书

在机器学习和统计学中，**特征选择**（英语：feature selection）也被称为**变量选择**、**属性选择** 或**变量子集选择** 。它是指：为了构建模型而选择相关特征（即属性、指标）子集的过程。使用特征选择技术有三个原因：

- 简化模型，使之更易于被研究人员或用户理解，[1]
- 缩短训练时间，
- 改善通用性、降低过拟合[2]（即降低方差[1]）

要使用特征选择技术的关键假设是：训练数据包含许多*冗余* 或*无关* 的特征，因而移除这些特征并不会导致丢失信息。[2] *冗余* 或*无关* 特征是两个不同的概念。如果一个特征本身有用，但如果这个特征与另一个有用特征强相关，且那个特征也出现在数据中，那么这个特征可能就变得多余。[3]

特征选择技术与特征提取有所不同。特征提取是从原有特征的功能中创造新的特征，而特征选择则只返回原有特征中的子集。 特征选择技术的常常用于许多特征但样本（即数据点）相对较少的领域。特征选择应用的典型用例包括：解析书面文本和微阵列数据，这些场景下特征成千上万，但样本只有几十到几百个。

目录

介绍

特征选择算法可以被视为搜索技术和评价指标的结合。前者提供候选的新特征子集，后者为不同的特征子集打分。 最简单的算法是测试每个特征子集，找到究竟哪个子集的错误率最低。这种算法需要穷举搜索空间，难以算完所有的特征集，只能涵盖很少一部分特征子集。 选择何种评价指标很大程度上影响了算法。而且，通过选择不同的评价指标，可以把特征选择算法分为三类：包装类、过滤类和嵌入类方法。[3]

- 包装类方法使用预测模型给特征子集打分。每个新子集都被用来训练一个模型，然后用验证数据集来测试。通过计算验证数据集上的错误次数（即模型的错误率）给特征子集评分。由于包装类方法为每个特征子集训练一个新模型，所以计算量很大。不过，这类方法往往能为特定类型的模型找到性能最好的特征集。

- 过滤类方法采用代理指标，而不根据特征子集的错误率计分。所选的指标算得快，但仍然能估算出特征集好不好用。常用指标包括互信息[3]、逐点互信息[4]、皮尔逊积矩相关系数、每种分类特征的组合的帧间/帧内类距离或显著性测试评分。[4][5] 过滤类方法计算量一般比包装类小，但这类方法找到的特征子集不能为特定类型的预测模型调校。由于缺少调校，过滤类方法所选取的特征集会比包装类选取的特征集更为通用，往往会导致比包装类的预测性能更为低下。不过，由于特征集不包含对预测模型的假设，更有利于暴露特征之间的关系。许多过滤类方法提供特征排名，而非显式提供特征子集。要从特征列表的哪个点切掉特征，得靠交叉验证来决定。过滤类方法也常常用于包装方法的预处理步骤，以便在问题太复杂时依然可以用包装方法。

- 嵌入类方法包括了所有构建模型过程中庸道德特征选择技术。这类方法的典范是构建线性模型的LASSO方法。该方法给回归系数加入了L1惩罚，导致其中的许多参数趋于零。任何回归系数不为零的特征都会被LASSO算法“选中”。

LASSO的改良算法有Bolasso[6]和FeaLect[7]。Bolasso改进了样本的初始过程。FeaLect根据回归系数组合分析给所有特征打分。另外一个流行的做法是递归特征消除（Recursive Feature Elimination）算法，通常用于支持向量机，通过反复构建同一个模型移除低权重的特征。这些方法的计算复杂度往往在过滤类和包装类之间。

传统的统计学中，特征选择的最普遍的形式是逐步回归，这是一个包装类技术。它属于贪心算法，每一轮添加该轮最优的特征或者删除最差的特征。主要的调控因素是决定何时停止算法。在机器学习领域，这个时间点通常通过交叉验证找出。在统计学中，某些条件已经优化。因而会导致嵌套引发问题。此外，还有更健壮的方法，如分支和约束和分段线性网络。

参见

- 群集分析
- 降维
- 特征提取
- 数据挖掘

参考文献

1. Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani. An Introduction to Statistical Learning. Springer. 2013: 204.
2. Bermingham, Mairead L.; Pong-Wong, Ricardo; Spiliopoulou, Athina; Hayward, Caroline; Rudan, Igor; Campbell, Harry; Wright, Alan F.; Wilson, James F.; Agakov, Felix; Navarro, Pau; Haley, Chris S. Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci. Rep. 2015, **5**.
3. Guyon, Isabelle; Elisseeff, André. An Introduction to Variable and Feature Selection. JMLR. 2003, **3**.
4. Yang, Yiming; Pedersen, Jan O. A comparative study on feature selection in text categorization. ICML. 1997.
5. Forman, George. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research. 2003, **3**: 1289–1305.
6. Bach, Francis R. Bolasso: model consistent lasso estimation through the bootstrap. Proceedings of the 25th international conference on Machine learning. 2008: 33–40. doi:10.1145/1390156.1390161
7. Zare, Habil. Scoring relevancy of features based on combinatorial analysis of Lasso with application to lymphoma diagnosis. BMC Genomics. 2013, **14**: S14. doi:10.1186/1471-2164-14-S1-S14

扩展阅读

- Feature Selection for Classification: A Review (Survey, 2014)
- Feature Selection for Clustering: A Review (Survey, 2013)
- Tutorial Outlining Feature Selection Algorithms, Arizona State University
- JMLR Special Issue on Variable and Feature Selection
- Feature Selection for Knowledge Discovery and Data Mining (Book)
- An Introduction to Variable and Feature Selection (Survey)
- Toward integrating feature selection algorithms for classification and clustering (Survey)
- feature subset selection and subset size optimization.pdf Efficient Feature Subset Selection and Subset Size Optimization (Survey, 2010)
- Searching for Interacting Features
- Feature Subset Selection Bias for Classification Learning
- Y. Sun, S. Todorovic, S. Goodison, Local Learning Based Feature Selection for High-dimensional Data Analysis *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. {0}8.{/0} {1}{/1}

外部链接

- A comprehensive package for Mutual Information based feature selection in Matlab
- Infinite Feature Selection (Source Code) in Matlab
- Feature Selection Package, Arizona State University (Matlab Code)
- NIPS challenge 2003 (see also NIPS)
- Naive Bayes implementation with feature selection in Visual Basic (includes executable and source code)
- Minimum-redundancy-maximum-relevance (mRMR) feature selection program

- FEAST (Open source Feature Selection algorithms in C and MATLAB)

取自“https://zh.wikipedia.org/w/index.php?title=特征选择&oldid=47786163”