中国科学 : 信息科学

SCIENTIA SINICA Informationis

论文

基于众包学习的交互式特征选择方法

陈长建[1], 姜流[1], 雷娜[2], 刘世霞[1*]

1. 清华大学, 北京 100084

2. 大连理工大学, 大连 116024

* 通信作者. E-mail: shixia@tsinghua.edu.cn

国家重点研发计划 (批准号: 2018YFB1004300), 国家自然科学基金 (批准号: 61672308, 61761136020, 61936002) 资助项目

摘要 集成特征选择算法将多种特征选择方法结果结合在一起, 旨在得到更加有效的特征子集. 然而这些算法通常假设每种特征选择方法是平等的, 没有考虑不同特征选择方法性能的差异性, 导致少数方法选择出的有效特征被忽略. 为解决这一问题, 本文提出一种可以有效地结合不同特征选择方法优势, 并利用专家的知识逐步改善所选特征的交互式特征选择方法. 该方法包括一个基于众包学习的集成特征选择算法和一个基于该算法开发的可视分析系统. 基于众包学习的集成特征选择算法利用众包学习模型对不同特征选择方法的性能进行建模, 计算每种方法的可靠性, 并在此基础上将这些方法的结果有机融合. 可视分析系统提供了丰富的排序方式, 帮助专家理解单个特征选择方法的特征选择结果和特征在分类任务中所起的作用, 从而让专家交互迭代地改善现有特征子集. 在 4 个真实世界数据集上的数值实验表明, 相比于现有的集成特征选择算法, 本文提出的算法能够带来 0.63%-2.85% 分类准确率的提升. 此外, 在文本和图像数据集上进行的两个案例分析表明, 本文提出的可视分析系统能够进一步带来 0.28%-5.24% 的分类准确率提升.

关键词 集成特征选择, 众包学习, 可视分析, 交互式可视化, 排序可视化

1 引言

特征选择指的是从特征全集中选择一个与机器学习任务相关的特征子集, 从而达到降低数据存储需求、减少机器学习模型训练时间以及提高机器学习模型预测能力的目的[1]. 在过去几十年中, 多种特征选择方法被提出. 这些特征选择方法往往基于某些特定假设. 但是这些假设并不总是成立, 导致在某些情况下单个特征选择方法会选择出较多无关特征. 比如特征选择方法 Variance Maximization（VM）假设有效特征的方差较大. 如果存在较多与机器学习任务无关但是方差较大的特征, 这些无效特征会被特征选择方法 VM 选出.

为此, 近年来许多工作[2~7] 将多种特征选择方法的结果结合在一起, 在提升特征选择结果有效性方面取得了比较好的效果. 然而这些方法通常假设每种特征选择方法是平等的, 没有考虑到不同特征

引用格式: 陈长建, 姜流, 雷娜, 等. 基于众包学习的交互式特征选择方法. 中国科学: 信息科学, 在审文章
Chen C, Jiang L, Lei N, et al. An interactive feature selection method based on learning-from-crowds (in Chinese). Sci Sin Inform, for review

选择方法性能的差异性. 这样导致某些有效特征在仅仅被少数特征选择方法选择时, 最终不会被选出. 其次, 使用集成特征选择算法的专家缺乏一个有效的方法帮助他们根据任务, 选择相应的特征选择方法来进行结合. 此外, 如果在得到的集成特征选择结果上训练的机器学习模型效果不理想, 专家需要一个有效的工具来理解不同特征选择方法结果、交互式修改特征选择结果以及构造新的特征. 本文受最大间隔多数表决众包学习模型（max-margin majority voting, 简称为 M^3V）[8, 9] 的启发, 提出了基于众包学习的集成特征选择算法（learning-from-crowds-based ensemble feature selection algorithm, 简称为 Crowd-EFS 算法）, 用于有效地结合不同特征选择方法的结果. 该算法将不同特征选择方法的性能用混淆矩阵和可靠性向量来刻画, 并通过 M^3V 模型对混淆矩阵、可靠性向量和集成特征选择结果进行联合估计. 在联合估计过程中, Crowd-EFS 算法根据可靠性向量调整各个特征选择方法的权值, 从而使得一些只被少数可靠的（权值高的）特征选择方法选择的有效特征最终会被选出, 一定程度上解决了上面所提到的平等对待每个特征选择方法所带来的问题. 本文在文本, 图像, 基因和 2003 年神经信息处理系统进展大会（NIPS）特征选择比赛数据 1) 等 4 类真实世界数据集上验证了该算法的有效性. 数值实验结果表明, 相比于最新的集成特征选择算法, Crowd-EFS 算法能够带来 0.63%-2.85% 的准确率提升.

虽然 Crowd-EFS 算法一定程度上解决了现有集成特征选择算法的问题 (只被少数特征选择方法选择的有效特征最终不被选出）, 但是如果某些有效特征没有被任何特征选择方法选出, 它们也不会被 Crowd-EFS 算法选出. 另外, 专家还需要分析选出的特征, 从而合并冗余特征. 在这两种情况下, 专家需要结合其知识来选择被遗漏的有效特征或构造新的特征, 从而逐步改善 Crowd-EFS 算法的特征选择结果. 为此, 本文在 Crowd-EFS 算法的基础上开发了一个可视分析系统, 名为"FeatureExplorer". 该系统的视频演示链接为：http://visgroup.thss.tsinghua.edu.cn/FeatureSelection/video.mp4. FeatureExplorer 提供了多种特征排序方式, 比如以 Crowd-EFS 算法输出的特征置信度、特征在不同类上分布的熵、多种特征选择方法评分和等为准则的排序方式, 从不同角度展现单个特征选择方法的特征选择结果和特征在机器学习任务中所起的作用. 此外, FeatureExplorer 还提供了一个交互迭代分析环境, 帮助专家交互迭代式地更新 Crowd-EFS 算法以及逐步改善所选特征.

综上所述, 本文的主要贡献有：

（1) 一个基于众包学习的集成特征选择算法. 数值实验表明, 相比于最新的集成特征选择算法, 该算法在文本、图像、基因和比赛等 4 类数据集上能带来 0.63%-2.85% 分类准确率的提升.

（2) 一个可视分析系统, 帮助专家理解不同特征选择方法结果、交互式修改特征选择结果以及构造新的特征.

（3) 两个在真实数据集上的案例分析, 说明本文所提出的可视分析系统能够帮助专家有效地选择特征, 进一步带来 0.28%-5.24% 分类准确率的提升.

2 相关工作

本文工作与集成特征选择算法和交互式特征选择相关. 本节将介绍这两个方向的相关工作.

2.1 集成特征选择算法

近年来, 多种集成特征选择算法被提出, 用以结合不同特征选择方法结果来提高所选特征的有效性. 现有的集成特征选择算法可以分为两大类: 基于评分和的集成特征选择算法和包裹式的集成特征

1) http://clopinet.com/isabelle/Projects/NIPS2003/

选择算法. 基于评分和的集成特征选择算法将不同特征选择方法对特征的评分进行加和, 从而得到每个特征新的评分. Saeys 等人 [2] 使用用户给定的函数将特征在每个特征选择方法上的排序值映射为一个实数值, 数值之和为每个特征新的评分. 评分高的特征优先被选择. 该策略也被用于 [3, 4] 中.

包裹式的集成特征选择算法使用包裹式方法将不同的特征选择结果结合 [5~7]. 比如, Netzer 等人 [6] 提出的 SFR 算法每次在各个特征选择方法评分最高且暂未被选择的特征集合中选择一个特征加入到现有特征选择结果中, 使得在新的特征选择结果上训练的分类器的" 受试者工作特征"(ROC) 曲线下面积 (AUC) 增益最大. 该过程不断重复直到分类器的 AUC 没有提升为止.

这些算法虽然一定程度上提升了所选特征的有效性, 但是没有考虑不同特征选择方法性能的差异性, 因而忽略仅被少部分特征选择方法选择出的有效特征. 本文提出的 Crowd-EFS 算法使用混淆矩阵和可靠性向量对特征选择方法的性能进行建模和估计. 特别地, 该算法根据可靠性向量调整不同特征选择方法的权值, 从而选出被少数可靠的特征选择方法选出的有效特征, 在一定程度上解决现有集成特征选择算法所存在的问题.

2.2 交互式特征选择与构造

交互式特征选择的工作可以分为两大类: 最小化冗余性方法以及最大化相关性方法. 相关性指的是特征与任务目标 (比如分类任务中的类标) 之间的相关程度; 冗余性指的是特征本身的相似性造成的冗余 (比如两个特征线性相关)[10].

最小化冗余性方法主要通过展示特征之间的相互关系, 帮助专家剔除冗余特征. 特征之间相互关系的展现形式一般来说有两种: 相关性矩阵和基于投影技术的二维散点图. Guo[11] 将两两特征之间的相互关系用相关性矩阵表示, 其中相关性用条件熵来衡量. MacEachren 等人 [12] 同样使用相关性矩阵来展示特征之间的条件熵, 并添加了散点图、二元图等以辅助专家选择特征. Ingram 等人 [13] 使用相关性矩阵展示特征之间的相似度以及 Pearson 相关性系数. 而 Yang 等人 [14] 使用投影技术, 比如多维缩放 (MDS) 算法, 将特征投影至二维平面. 特征在平面上的距离大小表示特征之间的相关性大小. Lin 等人 [15] 将特征投影在二维平面上来辅助特征选择, 用以发现某些子空间中存在的异常数据.

最大化相关性方法的主要特点是展示特征与任务目标 (比如说分类任务中的类标) 之间相关性的排序, 从而帮助专家交互式地选择与任务最相关的特征子集. Seo 等人 [16] 提出的 rank-by-feature 框架将所有特征按照方差进行排序, 并使用列表、矩阵等形式展现排序结果. 在 rank-by-feature 的框架下, Piringer 等人 [17] 增加选择数据子集的交互以探究特征在部分数据上的作用. May 等人 [18] 允许专家自定义一个划分, 将数据划分成多个数据子集, 在不同的数据子集上分别进行特征选择. 为结合多个特征选择方法的结果, Johansson 等人 [19] 使用专家设定的权值, 将特征的相关性、离群值和聚类特性进行加权平均以对每个特征进行排序. Krause 等人 [20] 则展示了每个特征在不同特征选择方法、不同交叉验证集上准确率的排序结果, 从而帮助专家选择合适的特征选择方法.

除了交互式特征选择之外, 还有少部分工作使用可视化技术辅助构造新的特征, 从而提高机器学习模型的性能 (比如分类器的准确率). Brooks 等人 [21] 通过分析被分类器误判的数据以及对比不同的特征来构造最能区分不同类别样本的特征. Liu 等人 [22] 通过对比不同特征的分布来选择相似的特征进行合并, 以产生新的特征.

本文所提出的交互式方法属于最大化相关性的交互式特征选择方法. 现有的最大化相关性的交互式特征选择方法大多只关注一种特征选择方法. 部分工作使用多种特征选择方法, 但仅提供单一的排序方式, 对于特征选择任务来说往往不充分. 本文所提出的交互式方法提供多种排序方式, 允许专家从多角度理解单个特征选择方法的特征选择结果和特征在机器学习任务中所起的作用. 此外, 本文提

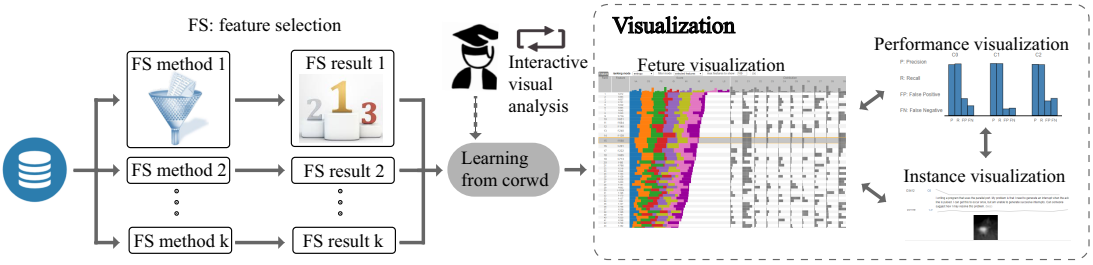

图 1　系统概览: 特征选择算法模块和可视化模块

Figure 1　System overview: a feature selection model module and a visualization module

出的 Crowd-EFS 算法能够使用混淆矩阵和可靠性向量对特征选择方法的性能进行建模和估计, 使得所选特征更加有效.

3　方法概述

本节介绍相关需求的收集以及基于这些需求所设计的交互式特征选择系统.

3.1　需求分析

本工作的想法来源于作者与机器学习 [38] 相关的项目. 在实践过程中, 特别是文本数据相关的项目 [35, 36], 通常会遇到数据维度过高导致算法速度达不到实时要求等问题, 需要进行特征选择以减少数据的维度. 在这些项目中, 集成特征选择方法常用于得到特征选择结果. 但是, 挑选用于集成的特征选择方法只能以不断试错的方式进行——不仅繁琐而且耗时耗力. 当所选特征效果不理想时, 分析特征以及修改特征选择结果需要耗费大量时间和精力.

为了提高上述过程的效率以及所选特征的有效性, 作者咨询了两位机器学习领域的专家（E_1, E_2). 作为高年级的博士生, 两位专家在平时的研究中经常使用集成特征选择方法以降低数据的维度. 为了使与专家之间的交流更加具体和高效, 本文以分类任务为例来设计系统. 作者与专家每周定期讨论所遇到的问题, 以及就原型系统的不足之处进行咨询和反馈收集. 该过程共持续了 14 周. 基于这些讨论和反馈, 本文得到以下 4 个需求:

需求 1（**R1**): 分析每种特征选择方法的差异性, 以提高集成方法的有效性, 从而使专家能在一个相对较好的特征选择结果基础上进行逐步改善.

需求 2（**R2**): 检查分类器在每个类上的分类性能, 并快速定位当前所选特征无法很好区分的类.

需求 3（**R3**): 允许从不同角度了解单个特征选择方法的特征选择结果和特征在分类任务中所起的作用, 从而让专家交互迭代地选择特征.

需求 4（**R4**): 理解特征的具体语义或功能, 辅助专家构造新的特征.

3.2　系统概览

基于以上 4 个需求, 本文提出了一个交互式特征选择系统. 该系统包括一个基于众包学习的集成特征选择算法模块以及一个在该算法基础上设计的交互式可视化模块. 系统概览如图 1 所示. 首先每个特征选择方法均得到一个特征选择结果. 根据各个特征选择方法的结果, 基于众包学习的集成特征选择算法模块利用 M^3V 模型估计出他们的混淆矩阵和可靠性向量, 并根据可靠性向量调整不同特征选择方法的权值, 在此基础上加权地集成这些特征选择结果（**R1**). 专家通过交互迭代分析环境选择

或者删除某些特征选择方法或者特征. 该信息会被反馈给基于众包学习的集成特征选择算法. 算法根据反馈更新特征选择方法的混淆矩阵、可靠性向量以及得到新的特征选择结果.

交互式可视化模块包含一个主可视化——特征可视化；以及两个辅助可视化——性能可视化和样本可视化. 特征可视化用于展示单个特征选择方法的特征选择结果以及特征在分类任务中所起的作用；性能可视化用于展示在当前所选特征上训练的分类器的分类性能；样本可视化用于提供样本的原始信息. 此外, 交互式可视化模块提供一个交互迭代分析环境来帮助专家逐步改善特征选择结果. 专家通过性能可视化查看当前分类器在每个类上的精确率和召回率等性能度量, 从而快速定位当前所选特征无法区分的类（**R2**). 特征可视化展示特征在不同特征选择方法上的评分、在不同类上出现频率的分布以及特征按照不同排序方式的排序结果, 帮助专家从不同角度了解单个特征选择方法的特征选择结果和特征在分类任务中所起的作用, 并在此基础上交互迭代地选择特征（**R3**). 在探索特征可视化时, 专家还可利用样本可视化查看特征在原始样本中的含义, 帮助理解特征在分类任务中的作用（**R4**). 基于对分类器性能和特征的理解, 专家可以交互式地构造出对分类任务有效的新特征（**R4**).

4 基于众包学习的集成特征选择算法

众包是获取标注数据的一种有效的方式 [8]. 众包数据包含多个工人对样本的标注. 众包学习从众包数据中推测出样本的正确类标. 本文提出的集成特征选择算法（Crowd-EFS 算法) 受启发于众包学习领域的最大间隔多数表决众包学习模型（M^3V)[8,9]. M^3V 模型通过混淆矩阵和可靠性向量对工人的标注性能进行刻画, 并通过正则化贝叶斯框架 [23] 对混淆矩阵、可靠性向量和样本的类标进行联合估计. 由于对工人标注性能进行了准确刻画, M^3V 模型的类标估计准确率往往高于将每个工人平等对待的多数表决方法 [8].

类似地, 在本文的问题背景下, 将每个特征选择方法看做是一个工人, 将不同特征选择方法的性能用混淆矩阵和可靠性向量来刻画. 混淆矩阵刻画了一个特征选择方法的特征选择结果与 Crowd-EFS 算法的结果是否一致. 一个典型的混淆矩阵如图 2 所示. 其中行代表算法预测为有效的特征（预测有效特征)/算法预测为无效的特征（预测无效特征), 列代表被该特征选择方法选出的特征（被选特征)/未被该特征选择方法选出的特征（未选特征). 单元格中的值代表预测有效特征或预测无效特征被该特征选择方法选出（被选特征) 或未被该特征选择方法选出（未选特征) 的概率. 比如左上角的 0.6 代表预测有效特征被该特征选择方法选出的概率为 0.6. 可靠性向量是一个 k 维向量. 其中 k 表示特征选择方法个数. 可靠性向量中的值代表特征选择方法性能的综合度量, 越高代表性能越好.

Crowd-EFS 算法首先使用多数表决方法得到初始特征选择结果, 并根据特征选择结果估计每个特征选择方法的混淆矩阵. 在混淆矩阵和特征选择结果的基础上, 该算法推测出特征选择方法的可靠性向量. 算法根据可靠性向量调整每个特征选择方法的权值, 从而得到新的特征选择结果. 根据新的特征选择结果, 算法进行新一轮的混淆矩阵以及可靠性向量的更新. 在这个过程中, 性能好的（可靠性高的) 特征选择方法的权重会不断的提高, 使得一些只被少数性能好的特征选择方法选出的特征最终也能被选出, 一定程度上解决了现有集成特征选择算法平等对待每个特征选择方法所带来的问题.

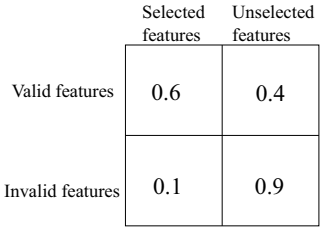

图 2 混淆矩阵刻画特征选择方法的性能特性

Figure 2 A confusion matrix that characterizes the performance of a feature selection method.

为方便算法描述, 本文定义以下符号. 假设训练数据集 $\mathcal{D}$ 包含 n 个样例 $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), .., (\mathbf{x}_n, y_n)$, 以及 m 个特征 $\boldsymbol{f} = (f_1, f_2, \ldots, f_m)$. 单个特征选择方法一共 k 个: $(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k)$. 第 i 个特征选择方法 $\mathcal{F}_i$ 对每个特征选择方法的评分记为 $\boldsymbol{s}_i$, 其中 $\boldsymbol{s}_{ij} \in \mathbb{R}^+$ 为第 i 个特征选择方法结果中第 j 个特征的评分值. 给定需要选择的特征个数 n', 第 i 个特征选择方法 $\mathcal{F}_i$ 的特征选择结果 $\boldsymbol{t}_i$ 包含了 $\boldsymbol{s}_i$ 中评分最高的 n' 个特征.

Crowd-EFS 算法分为两步. 首先其将多个特征选择方法的结果转化为众包数据 (数据转化), 然后使用 M^3V 模型将转化后的众包数据集成得到最终的特征选择结果 (集成).

第一步: 数据转化. 多个特征选择方法的特征选择结果无法直接输入 M^3V 模型中进行集成, 其首先需要转化为众包数据形式. 对于第 i 个特征选择方法的特征选择结果 $\boldsymbol{t}_i$, 其被转化为一个 m 维向量 $\overline{\boldsymbol{v}_i}$. 对于第 j 个特征, 如果其在特征选择结果 $\boldsymbol{t}_i$ 中, 则 $\overline{\boldsymbol{r}_{ij}}$ 等于 1; 否则 $\overline{\boldsymbol{r}_{ij}}$ 等于 0.

第二步: 集成. 令 $\boldsymbol{L} = [\overline{\boldsymbol{r}_1}, \overline{\boldsymbol{r}_2}, \ldots, \overline{\boldsymbol{r}}_k]^T$, 即 $\boldsymbol{L}_{ij}$ 表示第 i 个特征选择方法对第 j 个特征的排序结果. 如果将每个特征选择方法视为一个工人, 排序结果视为工人对特征的标注, 那么 $\boldsymbol{L}$ 可认为是众包学习问题中的众包数据. 此时将 $\boldsymbol{L}$ 输入 M^3V 模型得到每个特征的推测类标 $\boldsymbol{y}$. M^3V 模型引入了间隔 (margin) 的概念, 并将生成式方法和判别式方法的优势结合起来. 一个样本的间隔指的是其预测类标与其他所有类标的最小差距. 该方法最大化所有样本间隔之和, 以及给定众包数据 $\boldsymbol{L}$ 下, 正确类标 $\boldsymbol{y}$, 所有工人的可靠性向量 $\boldsymbol{\eta}$ 和混淆矩阵 $\boldsymbol{\Phi}$ 的后验概率

$$\inf_{q \in \mathcal{P}} \mathcal{L}(q(\boldsymbol{R})) + 2c \cdot \mathbb{E}_q \left[\sum_{i=1}^{M} (\zeta_i)_+ \right]. \tag{1}$$

其中 ζ_i 表示第 i 个样本的间隔. $(x)_+ = \max(0, x)$. $\boldsymbol{R}$ 表示一系列需要估计的变量, 包括正确类标 $\boldsymbol{y}$, 所有工人的可靠性向量 $\boldsymbol{\eta}$ 和混淆矩阵 $\boldsymbol{\Phi}$. $q(\boldsymbol{R})$ 是期望求得的分布. $\mathcal{L}(q(\boldsymbol{R}))$ 衡量了 $q(\boldsymbol{R})$ 和初始贝叶斯后验概率分布之间的 Kullback-Leibler (KL) 散度. $\mathbb{E}_g \left[\sum_{i=1}^{\mu} (\zeta_i)_+ \right]$ 是在 $q(\boldsymbol{R})$ 下对间隔求期望. 该优化问题可使用吉布斯采样 (Gibbs sampling) 求解 [8]. 最终算法可得到正确类标 $\boldsymbol{y}$ 以及每个特征的类标等于 1 (即被选出) 的置信度. 置信度高的特征优先被选出.

5 可视化

虽然 Crowd-EFS 算法一定程度上解决了某些只被少数特征选择方法选出的有效特征最终没有被集成特征选择方法选出的问题, 但是如果一个有效特征没有被任何特征选择方法选出, 那么其也不会被 Crowd-EFS 算法选出. 另外, 专家需要分析算法所选出的特征, 从而合并冗余特征. 针对这两种情况, 本文设计了一个可视分析系统 (FeatureExplorer), 让专家可以结合其知识交互式地选择未被 Crowd-EFS 算法选出的有效特征或构造新的特征, 从而逐步改善 Crowd-EFS 算法的特征选择结果. 单个特征选择方法以及特征是集成特征选择方法的关键. 理解单个特征选择方法的特征选择结果以及特征在分类任务中所起的作用, 有助于专家选取特征选择方法用于集成以及选择特征加入特征子集. 因此 FeatureExplorer 以特征的可视理解和分析为主. 此外, 为了让专家更好地改善所选特征, FeatureExplorer 提供了一个交互迭代分析环境, 方便专家从可视化中获取特征选择方法的特征选择结果和特征在分类任务中所起作用等信息, 并在此基础上交互迭代地对所选特征进行修改.

5.1 特征可视化

特征排序结果的可视化常用于交互式特征选择方法 [16~20], 用以帮助专家选择与分类任务相关的特征. 现有基于排序的交互式特征选择方法中, 结合多个特征选择方法结果的算法 [19, 20] 往往只提供

图 3　FeatureExplorer: (a) 性能可视化; (b) 特征可视化; (c) 样本可视化

Figure 3　FeatureExplorer: (a) performance visualization; (b) feature visualization; (c) instance visualization

单一的排序方式, 比如按照多个特征选择方法评分加权和排序的方式 [19]. 然而专家在进行特征选择时, 往往需要结合集成特征选择方法结果、特征在不同类上的分布情况以及特征在所有方法上的评分和等信息来进行分析. 单一的排序方式往往只能展示某一方面的信息, 而不能让专家从多个角度分析特征. 为此, 特征可视化提供了丰富的排序方式, 包括按照 Crowd-EFS 算法输出的特征置信度排序、按照特征在不同类上出现频率分布的熵排序以及按照特征选择方法评分之和排序等.

视觉编码. 特征可视化 (图 3 (b)) 使用堆叠柱状图 (图 3B) 以及柱状图 (图 3D) 展示特征按照不同排序方式的排序结果以及特征在不同特征选择方法上的评分、在不同类上出现频率的分布. 其中堆叠柱状图 (图 3B) 中的每一个柱 (bar) 对应一个特征. 柱的水平位置代表对应特征在当前排序方式下的排序高低. 每个柱 (图 3C) 由多个带颜色的矩形块堆叠而成, 每个矩形块长度代表其颜色以及其编号对应的特征选择方法对特征的评分大小. 堆叠柱状图上方带颜色的字母缩写表示颜色以及编号和特征选择方法之间的对应关系. 比如橙色字母缩写”0-VM” 表示橙色和编号”0” 对应特征选择方法 Variance Maximization (VM). 使用堆叠柱状图展示特征在不同特征选择方法上的评分, 既能可视化特征总评分, 也可展示出各个特征选择方法对总评分的贡献.

柱状图 (图 3D) 用来展示特征在不同类上出现频率的分布. 每个柱状图对应一个类. 比如图 3 (b) 中从左至右共有 6 个柱状图, 分别对应分类问题中的 6 个类. 图 3D 对应第 6 个类, 展示特征在第 6 个类上出现频率的分布. 处于同一水平位置的柱 (包括堆叠柱状图中的柱) 对应同一个特征. 使用多个柱状图展示特征在不同类上出现频率分布情况, 可以清楚地展现特征在不同类上出现频率分布的不均匀性. 而这种不均匀性, 是评判一个特征区分不同类能力大小的重要依据. 比如一个特征仅在某一个类上出现, 那么该特征就能很好地区分这个类与其他类.

在堆叠柱状图和柱状图上方是特征在不同特征选择方法上的评分或者是在不同类上特征出现频

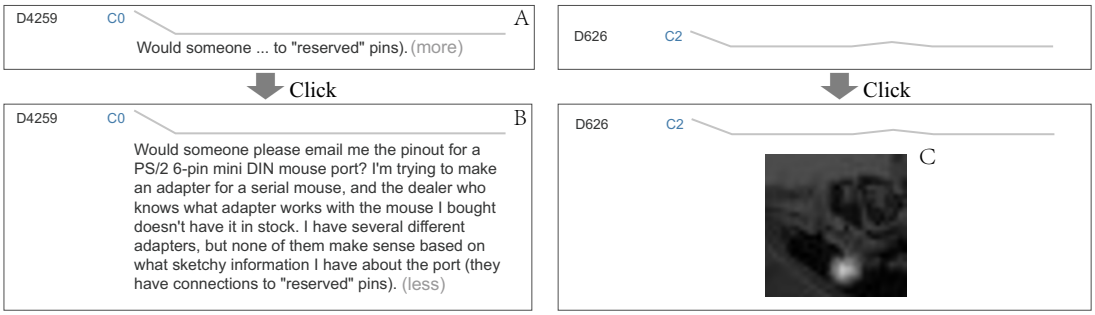

图 4　样本可视化在文本和图像数据上的不同展现形式: (a) 文本数据集; (b) 图像数据集

Figure 4　Two different forms of instance visualization: (a) text; (b) image

率的统计分布直方图. 比如图 3E 是在第 6 个类上特征出现频率的统计分布直方图. 统计分布直方图可以帮助专家查看特征在某个属性上 (即特征选择方法的评分或者在类上的频率) 的分布情况, 以及筛选所需要的特征 (见本小节的**可扩展性**部分).

排序方式. 特征可视化提供多个与特征选择任务相关的排序方式:

(1) 按照 Crowd-EFS 算法输出的特征置信度排序. 由于交互式特征选择是在 Crowd-EFS 算法结果基础上进行, 所以展示众包学习模型 (M^3V 模型) 的结果有助于特征选择任务. M^3V 模型的输出包括正确类标 $\boldsymbol{y}$ 以及每个特征类标等于 1 (即被选出) 的置信度 $\boldsymbol{P}$. 特征 f_1 比特征 f_2 排序更高当且仅当 $\boldsymbol{P}_1$ 大于 $\boldsymbol{P}_2$.

(2) 按照特征在不同类上出现频率分布的熵排序. 熵衡量了一个分布的均匀性. 一个特征在不同类上出现频率分布越不均匀, 其区分不同类的能力越强. 因此, 将特征按照熵从小到大排序可以帮助专家选择对分类任务有效的特征.

(3) 按照特征选择方法评分之和排序. 不同特征选择方法对特征的评分之和是衡量特征有效性的重要指标. 堆叠柱状图能直观地展现评分之和.

可扩展性 (Scalability). 当特征数量增多时, 特征可视化会产生可扩展性问题. 为缓解该问题, 本文使用鱼眼 (fish-eye) 技术以及多种基于过滤的交互方法. 鱼眼技术详细展示专家感兴趣特征的信息, 而粗略展示其他特征的信息 (图 3B). 这样可使得可展示的特征数量从数十个增加到数百个. 考虑到专家可能对按照某些条件筛选出来的特征比较感兴趣, 特征可视化支持专家对特征进行过滤. 比如, 专家可以通过控制面板 (图 3F) 选择仅展示当前已选择的特征或者是未选择的特征, 也可以设定显示特征的数目, 以过滤掉其余的特征. 此外, 为了方便查看特定特征选择方法评分值或者是在特定类上的分布频率处于某个范围内的特征, 专家也可以点击频率分布直方图 (图 3E) 中的柱形来筛选出该柱形对应的特征.

5.2　交互迭代分析环境

本小节将介绍 FeatureExplorer 的交互迭代分析环境. 该环境包括性能可视化、样本可视化、它们与特征可视化之间的协调关联以及一些交互功能.

5.2.1 性能可视化

了解在当前所选特征上训练的分类器的性能对特征选择任务来说非常重要 [21]. 最常见的衡量分类器性能的指标是准确率. 然而, 对多分类问题, 准确率提供的信息有限 [27], 缺少单个类与其他类混淆程度的信息. 为此, 将分类器区分某个类与其他类的问题转化为多个二分类问题, 并在此基础上用柱状图 (图 3 (a)) 展示多个二分类问题的精确率 (precision)、召回率 (recall)、假正例 (false positive) 和假反例 (false negative). 前两者是常用的度量二分类器性能的指标. 后两者展示了被错分样本的数量, 在分类任务的背景下可以辅助专家有效地进行特征选择 [21].

5.2.2 样本可视化

专家通常需要结合特征在样本中的具体语义以对其在分类任务中的作用进行分析 [37]. 例如, 单词作为文本数据的特征, 其上下文可以直观地帮助专家理解该单词在分类任务中的作用. 比如, 对于单词"law", 专家需要结合上下文才能知道其是指法律类文档中的" 法律" 还是指科技类文档中的" 定律". 为此, 样本可视化展示特征在样本内容中的具体语义.

在性能可视化中点击假正例或假反例对应的柱形 (bar) 后, 样本可视化会展示该柱形对应的样本; 在特征可视化中点击某个特征后, 样本可视化会展示包含该特征的样本. 样本可视化如图 3 (c) 所示, 其由多个单元竖直排列而成, 每个单元代表一个样本. 每个单元 (图 3G) 包含该样本的编号、所属类别、特征取值统计分布情况以及特征在样本内容中的具体语义. 其中特征取值统计分布情况用一条灰色折线图表示. 折线上点的横坐标代表特征在该样本上的取值, 纵坐标代表特征的个数. 特征在样本中具体语义的展现形式与数据集的类型相关. FeatureExplorer 支持最常见的文本数据以及图像数据. 其他类型的数据也可以方便地集成进入系统.

对于文本数据, 样本可视化初始展示缩略文本. 如果在特征可视化中选择了某个特征, 则缩略文本 (图 4A) 会展示特征 (单词) 在文本中出现的位置以及前后文 (前后各两个单词), 其余位置的文本会用省略号表示。该展现形式可以给出某个特征在文本中位置的概览. 点击样本后可看到完整文本 (图 4B).

对于图像数据, 本文使用深度神经网络 (deep neural network, 简称为 DNN) 提取的特征 [27]. 由于 DNN 特征可解释性低于文本特征, 因此样本可视化展示图像对于某个特征的响应区域 (图 4C) 来帮助专家理解该特征在分类任务中所起的作用. 如图 4C 所示, 车辆图像对某个特征的响应区域集中在车轮区域 (亮度高的区域为响应区域), 说明该特征主要检测的是图像中车辆的车轮. 如果一张图像在该特征上的取值大, 那么该图像有较大概率属于车辆类. 本文使用 Zhou 等人最近提出的计算图像对特征的响应区域的方法 [28] 来求解图像对特征的响应区域.

5.2.3 交互

本小节将介绍专家的反馈对 Crowd-EFS 算法进行更新的流程、专家进行特征构造以及使用 FeatureExplorer 交互迭代分析环境改善所选特征的典型交互迭代流程.

算法更新. 专家的反馈分为两种: 删除某个特征选择方法或者选择/删除某个特征. 当专家决定删除某个特征选择方法时, 该方法对应的评分将被从算法输入中删除, 然后算法会重新运行得到新的特征选择结果. 如果专家选择 (或删除) 某个特征, 那么该特征在所有特征选择方法上的评分会被置为对应特征选择方法所能给出的最高分 (或最低分). 此时如果直接重新运行算法, 由于吉布斯采样 (Gibbs sampling) 随机性较大, 特征选择结果的变化可能会较大, 导致专家无法理解和信任新的结果. 为保持稳定性, 算法在更新时, 所有工人的可靠性向量 $\boldsymbol{\eta}$ 和混淆矩阵 $\boldsymbol{\Phi}$ 使用上一轮的结果进行初始化, 这样

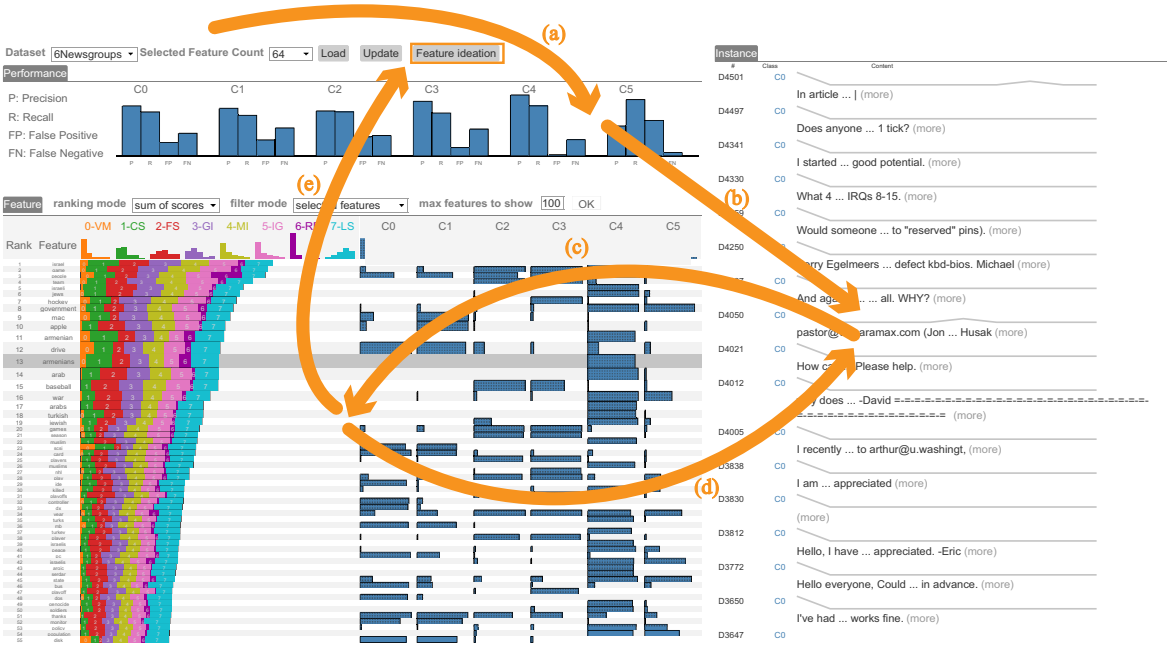

图 5　FeatureExplorer 的典型交互迭代分析流程

Figure 5　A typical interactive and iterative analysis workflow of FeatureExplorer

可让特征选择结果尽快收敛到上一轮结果的附近, 避免过多随机采样带来的不稳定性. 专家在交互式更新算法时, 可以右键点击特征可视化中特征选择方法名称或者特征对应的柱形 (bar), 然后在弹出的菜单中进行选择或删除操作.

特征构造. FeatureExplore 支持通过合并多个特征来完成特征构造. 将多个对分类任务作用相似的特征合并为一个新特征可以减少数据特征维度和提高分类器性能 [21]. 比如 20NewsGroups[29] 数据集中的"armenians" 和"arab" 两个特征, 都是指中东地区的城市或国家, 他们对于区分中东话题类文本的作用相似, 所以可以把他们合并成一个新的特征"middle-east". 点击系统控制面板中的"feature ideation" 按钮, 一个新的窗口会弹出. 专家可以在该窗口输入新特征的名称以及其所包括的特征. 新的特征在不同特征选择方法中的评分将会被重新计算, 并在特征可视化中展示.

典型交互迭代流程. 专家可使用交互迭代分析环境迭代式地改善所选特征. 典型的交互迭代流程如图 5 所示. 整个流程从性能可视化发起. 专家首先从性能可视化查看当前分类器在每个类上的精确率和召回率等性能度量 (图 5 (a)), 定位当前被错误分类的样本. 随后使用样本可视化查看被错误分类样本的特征值统计分布情况 (图 5 (b)), 以找到这些被错误分类样本的共同点. 之后专家使用特征可视化查看特征选择方法的结果以及特征在分类任务中所起的作用 (图 5 (c)). 查看特征选择方法的结果可以帮助专家筛选合适的特征选择方法用于集成. 查看特征在分类任务中所起的作用可以帮助专家探究当前所选特征造成样本被错误分类的原因, 从而决定选择或删除某些特征. 在此过程中, 专家还需要结合样本可视化了解特征在样本内容中的具体语义 (图 5 (d)). 此外, 根据特征的语义, 专家可以交互式地构造新的特征. Crowd-EFS 算法以及分类器会接受专家的反馈并进行更新 (图 5 (e)). 更新后的特征选择结果和分类器分类结果会反馈到三个可视化中, 专家根据更新的结果进行下一轮的迭代.

6 实验与分析

本文的实验包括两个部分. 第一部分是数值实验, 用以说明本文提出的 Crowd-EFS 算法的有效性. 第二部分是两个在真实数据集上进行的案例分析, 用以说明本文所开发的 FeatureExplorer 能够在 Crowd-EFS 算法结果的基础上, 进一步提高分类准确率.

6.1 数值实验

本文选取两种方法与 Crowd-EFS 算法进行对比: 基于评分和的集成特征选择算法 (baseline-1) 和单个最好特征选择方法 (baseline-2). 选取基于评分和的集成特征选择算法作为对比方法的原因是该方法为目前最新的集成特征选择算法之一. 另外一类最新的集成特征选择算法是包裹式的集成特征选择算法 [7]. 本文并未选用该类算法与 Crowd-EFS 算法进行对比, 原因是该类算法需要多次训练分类器, 耗时长, 难以适用于实际应用.

数据集与实验设置. 为了说明 Crowd-EFS 算法的普适性, 本文选取了 4 个不同类型的数据集: 文本、图像、基因以及 2003 年神经信息处理系统进展大会 (NIPS) 特征选择比赛数据 [2)]. 其中 Rueters-r0[30] 是文本数据集 Reuters[31] 的一个子集, 使用 TF-IDF 作为其特征, 包括 11 个类别, 1504 个样本和 2886 个特征. 图像数据集 COIL-20[32] 使用 ResNet-56 的全局平均池化层的输出作为特征, 包括 20 个类别, 1440 个样本和 1024 个特征. LUNG[3)] 是基因数据集, 包括 5 个类别, 203 个样本和 3312 个特征. Gisette 是比赛数据集, 包括 2 个类别, 7000 个样本和 5000 个特征. 本文选取 8 种常见的特征选择方法参与集成特征选择: Variance Maximization (VM), Chi-square (CS), F-score (FS), Gini Index (GI), Mutual Information Maximization (MI), Information Gain Maximization (IG), ReliefF (RF), Laplacian Score (LS). 其中包括有监督的特征选择方法 (CS、FS、GI、MI、IG、RF) 和无监督的特征选择方法 (VM、LS). 实验中, 本文分别选择特征总个数的 0.5% , 1% , 2% , 5% , 10% , 20% 进行实验.

实验中使用支持向量机 (SVM)[33] 作为分类模型. 在特征选择方法计算出结果后, SVM 在训练集 (training set) 上进行训练. SVM 在测试集 (testing set) 上的分类准确率作为评价特征选择方法结果好坏的指标. 为了行文方便, 后文中该准确率简称为某特征选择方法在测试集上的准确率 (比如 Crowd-EFS 算法在测试集上的准确率). 在实验中, 对于文本数据集, SVM 使用线性核函数; 对于其他数据集, SVM 使用高斯核函数. SVM 的惩罚参数统一设为 1. 高斯核函数的宽度参数设为特征值的倒数.

为得到 Crowd-EFS 算法与基于评分和的集成特征选择算法的准确率, 数据被划分为 80% 训练集和 20% 测试集. 为得到单个最好特征选择方法的准确率, 需要一个验证集 (validation set) 来帮助挑选出单个最好特征选择方法. 因此数据被划分为 60% 训练集, 20% 验证集和 20% 测试集. 在验证集上准确率最高的一个特征选择方法被选作为单个最好特征选择方法, 并使用该特征选择方法在测试集上的准确率作为单个最好特征选择方法的准确率.

结果与分析. 在 4 个数据集上, Crowd-EFS 算法带来的平均准确率提升分别为 1.77% , 0.63% , 2.85% , 0.69%. 这里平均准确率提升指的是在不同特征选择比例下准确率提升的平均值. 图 6 展示了 4 组数据在不同特征选择比例下的详细实验结果. 从图中可以看出, 本文所提出的 Crowd-EFS 算法在所有数据集上、不同特征选择比例情况下都不差于基于评分和的集成特征选择算法 (baseline-1) 以及单个最好特征选择方法 (baseline-2). 在 LUNG 数据集上选择取 1% 的特征时, Crowd-EFS 算法相比于基于评分和的集成特征选择算法的准确率提升幅度达到 7.31% (图 6A). Crowd-EFS 算法优于基

2) http://clopinet.com/isabelle/Projects/NIPS2003/
3) http://featureselection.asu.edu/datasets.php

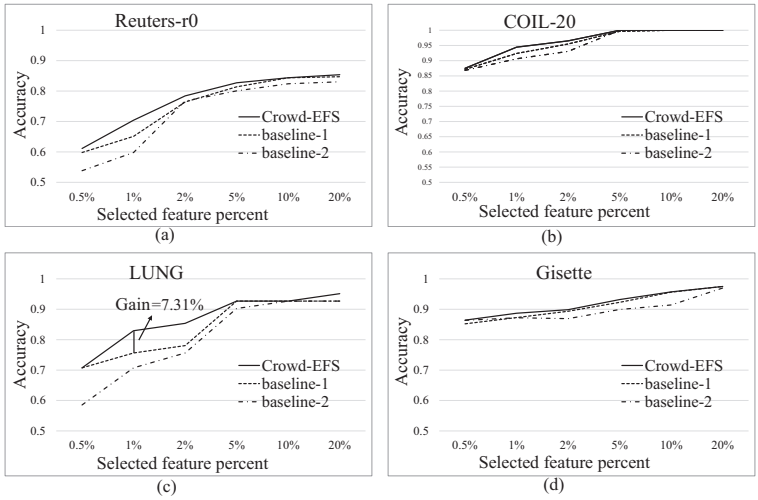

图 6 Crowd-EFS 算法与 baseline-1baseline-2 在 4 个数据集上的实验结果

Figure 6 Experiment results of Crowd-EFS, baseline-1 and baseline-2 on 4 datasets.

于评分和的集成特征选择算法的原因是其能够估计不同特征选择方法性能的混淆矩阵和可靠性向量, 并使用可靠性向量对不同特征选择方法进行加权集成. 此外, 在 4 个数据集上, 特征选择比例较少时, Crowd-EFS 算法相比于基于评分和的集成特征选择算法的准确率增益较少. 这可能是由于选择的特征较少时, M^3V 模型的输入数据较少, 导致模型对不同特征选择方法性能的估计不够准确, 从而使得增益减少.

6.2 案例分析

本节介绍两个分别在文本和图像数据集上进行的案例分析. 两个案例分析分别由专家 E_1 和 E_2 与作者合作完成. 第一个案例分析说明, 在选择不同特征比例情况下, 使用 FeatureExplorer 均能在 Crowd-EFS 算法结果的基础上进一步提高准确率. 第二个案例分析说明, 使用 FeatureExplorer 不仅能进一步提高准确率, 而且能在仅使用部分特征的情况下, 准确率超过使用所有特征的准确率.

6.2.1 文本数据

此案例分析由专家 E_1 与作者合作完成. 该案例分析中使用的文本数据是 20NewsGroups-6. 该数据集包含 20NewsGroups[29] 数据集中 6 个话题的文本. 6 个话题分别为: 个人电脑硬件 (pc.hardware)、苹果电脑硬件 (mac.hardware)、棒球 (baseball)、冰球 (hockey)、中东话题 (mideast)、政治话题混合 (politics.misc). 本文将这 6 个话题分别记为 C_i (i=0, 1, ⋯, 5). 该数据中训练集和测试集的大小分别为 4522 和 1131. 实验中使用 TF-IDF 作为该数据集的特征. 特征总数为 12279. FeatureExplorer 初始装载选择 0.5% 特征时 Crowd-EFS 算法的结果. 通过可视化结果可以发现, Crowd-EFS 算法将某些仅被少数特征选择方法选出的有效特征选出。如图 3I, "israelis" 仅被特征选择方法 CS 选出, 但其语义为中东地区的民族, 与分类任务相关, 最终被 Crowd-EFS 算法选出。专家在此结果基础上进行改善 (**R1**). 在该所选特征上训练的 SVM 的测试集准确率为 74.95%. 在下文中, 如果没有特别说明, 准确率均指在测试集上的准确率.

查看分类器性能 (**R2**, **R4**). 专家 E_1 首先查看性能可视化, 发现大部分类 (C_0, C_1, C_2, C_3, C_4) 的精确率和召回率都相差不大 (差异在 0.088 至 0.170 之间), 并且精确率略高于召回率 (图 3A). 然而 C5

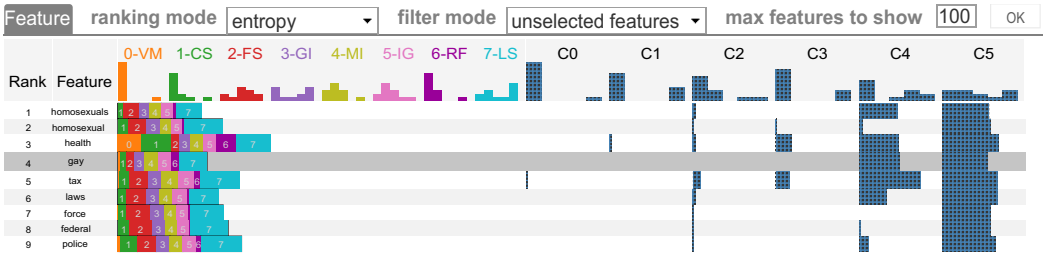

图 7　在特征可视化中，专家 E_1 通过排序和过滤功能筛选所需特征

Figure 7　The expert gets features needed through ranking and filtering functions in feature visualization

的精确率要小于召回率, 并且两者差异较大 (差异为 0.387). 这表明 C_5 的假正例偏多, 即其他类 (C_0, C_1, C_2, C_3, C_4) 的样本容易被 SVM 误判为 C_5. 专家 E_1 决定结合特征可视化与样本可视化探究其原因.

专家 E_1 点击 C_5 假正例对应的柱形 (bar) 后, 样本可视化展示出对应的样本, 即被误判为 C_5 的样本 (图 3 (c)). 可以看到, 被误判为 C_5 的样本的特征值统计分布折线图都非常相似: 大多数特征的取值都为 0. 专家 E_1 随后在样本可视化中展示所有 C_5 的样本 (点击性能可视化中 C_5 的名称). 发现 C_5 样本的特征值统计分布折线图和 C_5 假正例的非常相似. 专家 E_1 表示这可能是造成 C_5 假正例较多的直接原因.

分析特征 (**R3**). 为进一步探究 C_5 假正例较多的原因, 专家 E_1 选择在特征可视化中查看 Crowd-EFS 算法所选特征. E_1 通过特征在不同类上出现频率的分布 (图 3D) 以及对应的统计分布直方图 (图 3E) 立刻发现, Crowd-EFS 算法所选特征在 C_5 的样本上出现频率较少. 这导致 C_5 样本的特征向量非常稀疏 (0 值较多). 考虑到实验中使用的 SVM 的核函数为线性核函数, 非常稀疏的向量之间在此相似度度量方法下非常相似, 所以其他类中特征向量稀疏的样本容易被误判为 C_5. 为此, 专家 E_1 决定选择一些在 C_5 样本中出现频率较大, 并且在其他类样本上出现频率较少的特征. 他首先使用特征可视化的过滤功能, 让特征可视化仅显示当前未被选择的特征. 然后按照特征在不同类上出现频率分布的熵排序, 并点击特征可视化中 C_5 对应的频率分布直方图中值最大的柱形来选择在 C_5 样本中出现频率较大的特征. 此时特征可视化中从上到下排列的即为专家所需特征 (图 7) . 专家 E_1 依次选取了”homosexuals”, ”homosexual”, ”health”, ”gay”, ”tax”, ”laws”, ”force”, ”federal” 等 8 个特征. Crowd-EFS 算法接受该反馈并对特征选择结果进行更新. 在新的特征选择结果上, C_5 的假正例减少 36 个, SVM 的准确率从 74.95% 提升至 78.60%.

分析单个特征选择方法 (**R1**, **R3**). 专家 E_1 选择在特征可视化中展示当前已选择的特征 (图 3B). 发现排序靠前的特征中, 大多数特征选择方法的评分均较高, 然而 ReliefF 特征选择方法的评分较低. 这说明该特征选择方法的结果与 Crowd-EFS 算法的结果差异较大, 所选特征并不是最有效的特征. 因此, 专家 E_1 决定删除该特征选择方法, 并重新更新 Crowd-EFS 算法以及 SVM. 更新后, 准确率从 78.60% 提升至 79.00%.

构造特征 (**R4**) . 在特征可视化中, 专家 E_1 发现某些特征在不同类上出现频率的分布非常相似. 比如”armenians” 和”arab” 两个特征, 都仅仅出现在 C_4 (中东话题) 中 (图 3H) . 结合他们的语义 (都是指中东地区的城市或国家) , 专家 E_1 决定把它们合并成一个新的特征”middle-east”. 类似地, 专家还发现”armenian”, ”turkey”, ”israels”, ”serdar” 也是指中东地区的城市或国家, 均可合并入新特征”middle-east” 中. 按照同样的方法, 专家 E_1 构造了其他 4 个新特征, 准确率从 79.00% 提升到

图 8　在 CIFAR10 数据集上, Crowd-EFS 算法的结果

Figure 8　Feature selection results of the Crowd-EFS algorithm on CIFAR10 dataset

了 80.19%.

至此, 新得到的特征总共有 64 个, 相比 Crowd-EFS 算法的特征选择结果来说, 特征数目并未增加, 但是准确率从未借助可视化时的 74.95% 提升至了 80.19%.

专家 E_1 在特征选择比例为 2% , 10% , 20% 时, 分别使用 FeatureExplorer 进行了案例分析. 结果表明, 在不同的特征选择比例下, FeatureExplorer 均能在不增加特征数的情况下将准确率分别从 87.09% , 91.34% , 92.48% 提高至了 89.57%, 92.48%, 93.37%. 这说明 FeatureExplorer 能够帮助专家选择出 Crowd-EFS 算法未能选出的有效特征, 以及构造出对分类任务有效的新特征.

6.2.2　图像数据

本案例分析由专家 E_2 与作者合作完成. 在此案例分析中所使用的图像数据集 CIFAR10[34] 包含 10 个类别: 飞机 (airplane) 、汽车 (automobile), 鸟 (bird), 猫 (cat), 鹿 (deer), 狗 (dog), 蛙 (frog), 马 (horse), 船 (ship), 卡车 (truck). 这 10 类图像分别记为 C_i(i=0, 1, ⋯, 9). 该数据的训练集和测试集大小分别为 50000 和 10000. 实验中使用 ResNet-56 提取该数据集的特征. 特征总数为 256. 要特别指出的是, 使用所有特征能达到的准确率为 92.72%. FeatureExplorer 初始装载了在选择 50% 特征时 Crowd-EFS 算法的结果 (R1) . 在该所选特征上训练的 SVM 的测试集准确率为 92.48%.

查看分类器性能 (**R2**, **R4**) . 专家 E_2 首先查看性能可视化, 发现所有类的精确率和召回率都很高, 但是假正例和假反例的数量有较大差别. 其中 C_1 和 C_9 的假正例和假反例的数量都较少, 而 C_3 和 C_5 的假正例和假反例的数量都偏多 (图 $8A_1$ 和图 $8A_2$) . 专家 E_2 决定先探究 C_3 分类性能较差的原因, 于是点击 C_3 假正例对应的柱形. 样本可视化展示被误分为 C_3 的样本 (图 8C) . 专家浏览之后发现被误分为 C_3 的样本中, 最多的是来自 C_5 的样本.

分析特征 (**R3**) . 考虑到 C_3 (猫) 和 C_5 (狗) 两个类的图像存在着较多相似之处, 专家 E_2 猜测可能是因为当前所选特征在两个类上的出现频率分布比较相似, 导致 C_5 的样本被大量错分为 C_3. 于

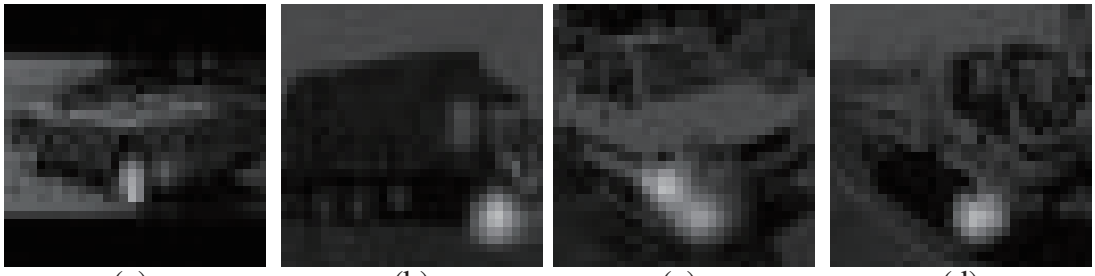

图 9 C_1 **(汽车) 和** C_9 **(卡车) 的样本对特征 f-29 和 f-169 的响应区域: (a)** C_1 **(汽车) 的样本对特征 f-29 的响应区域; (b)** C_9 **(卡车) 的样本对特征 f-29 的响应区域; (c)** C_1 **(汽车) 的样本对特征 f-169 的响应区域; (d)** C_9 **(卡车) 的样本对特征 f-169 的响应区域**

Figure 9 Response areas of (a) C_1 on f-29; (b) C_9 on f-29; (c) C_1 on f-169; (d) C_9 on f-169;

是专家 E_2 在特征可视化中选择只显示当前 Crowd-EFS 算法所选特征. 专家发现, 在 Crowd-EFS 算法所选特征中, 有很多特征在两个类上出现频率的分布非常相似, 如图 $8B_1$ 和图 $8B_2$. 为区分两个类的图像, 专家 E_2 决定选择一些仅在 C_3 上出现频率较大, 而在 C_5 上出现频率较小的特征; 或者是仅在 C_5 上出现频率较大, 而在 C_3 上出现频率较小的特征. 专家 E_2 在特征可视化中选择只显示当前未被选择的特征, 然后按照特征在不同类上出现频率分布的熵排序, 并仅显示在 C_3 上出现频率大的特征. 专家 E_2 在所得排序结果中筛选出了 f-11, f-189 等 5 个在 C_3 上出现频率大, 在 C_5 上出现频率小的特征。专家用同样的方法筛选出了 f-109, f-252 等 4 个在 C_5 上出现频率大, C_3 上出现频率小的特征。这一步一共选择出了 9 个特征, 准确率从 92.48% 提升至了 92.71%.

构造特征 (**R4**) . 专家 E_2 同时还发现, 有些特征在图像上的响应区域非常相似. 如图 9 所示, 特征 f-29 和 f-169 都是检测汽车或卡车的轮子. 对于该分类任务来说, 这两者作用是类似的, 因此可以合并成一个新的特征" 轮子". 按照这个方法, 专家 E_2 总共构造了 4 个新特征, 准确率从 92.71% 提升到了 92.76%.

至此, 新得到的特征总共为 128 个, 相比 Crowd-EFS 算法的特征选择结果来说, 特征数目并未增加, 但是准确率从 92.48% 提升至了 92.76%. 并且该特征选择结果仅包含了 50% (128) 的特征, 所得准确率 (92.76%) 已经超过了使用所有 256 个特征时的准确率 (92.72%) .

7 结论

本文提出一种可以有效结合多种特征选择方法结果, 并利用专家知识逐步改善所选特征的交互式特征选择方法. 该方法包括一个基于众包学习的集成特征选择算法 (Crowd-EFS 算法) 以及一个可视分析系统 (FeatureExplorer) .

为了有效地结合多种特征选择方法的结果, Crowd-EFS 算法将不同特征选择方法的性能用混淆矩阵和可靠性向量刻画, 并通过 M^3V 模型对混淆矩阵, 可靠性向量和集成特征选择结果进行联合估计. 在该过程中, Crowd-EFS 算法根据可靠性向量调整各个特征选择方法的权值, 从而使得一些只被少数可靠的特征选择方法选出的有效特征最终被选出. 为结合专家的知识逐步提高 Crowd-EFS 算法所得特征选择结果的有效性, FeatureExplorer 提供了多种特征排序方式, 从不同角度展现单个特征选择方法的特征选择结果, 以及一个交互迭代分析环境来帮助专家交互迭代式地更新算法以及逐步改善所选特征. 最后, 在真实数据集上, 本文通过一个数值实验以及两个案例分析验证了所提出的交互式特征

选择方法的有效性和有用性.

本文提出的交互式方法目前只在文本和图像数据上验证了可行性和有效性, 未来工作之一是将本文的方法拓展到其他类型的数据, 比如语音数据. 目前系统假设用户的反馈都是正确的, 但在实际应用中, 专家也可能提供错误的信息。因此, 本文计划将用户发起的改进和系统发起的改进统一到一个可视分析框架下, 有效地解决专家反馈中可能出现的错误并处理专家反馈和模型分析结果之间的冲突.

参考文献

1 Guyon I, and Andre E. An introduction to variable and feature selection. Journal of Machine Learning Research, 2009, 3(Mar): 1157-1182

2 Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Antwerp, Belgium, 2008: 313-325

3 Wang H, Khoshgoftaar T M, Napolitano A. A comparative study of ensemble feature selection techniques for software defect prediction. In: Proceedings of the International Conference on Machine Learning and Applicatioin. Hyatt Regency Bethesda, USA, 2010: 135-140

4 Li X, Zhang T W, Guo Z. A Novel Ensemble Method of Feature Gene Selection Based on Recursive Partition-Tree (in Chinese). Chinese Journal of Computers, 2004, 27(5): 675-682 [李霞, 张田文, 郭政. 一种基于递归分类树的集成特征基因选择方法. 计算机学报, 2004, 27(5): 675-682]

5 Bol ó n-Canedo V, S á nchez-Marono N, Alonso-Betanzos A. Distributed feature selection: An application to microarray data classification. Applied soft computing, 2015, 30: 136-150

6 Netzer M, Millonig G, Osl M, et al. A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry. Bioinformatics, 2009, 25(7): 941-947

7 Yang F, Mao K Z. Robust feature selection for microarray data based on multicriterion fusion. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(4): 1080-1092

8 Tian T, Zhu J. Max-Margin Majority Voting for learning from crowds. In: Proceedings of Advances in Neural Information Processing Systems. Palais des Congr è s de Montr é al, Canada, 2015: 1621-1629

9 Liu M, Jiang L, Liu J, et al. Improving learning-from-crowds through expert validation. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 2329-2336

10 Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238

11 Guo D. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. Information Visualization, 2003, 2(4): 232-246

12 MacEachren A, Xiping D, Hardisty F, et al. Exploring high-D spaces with multiform matrices and small multiples. In: Proceedings of the IEEE Symposium on Information Visualization. Seattle, USA, 2003: 31-38

13 Ingram S, Munzner T, Irvine V, et al. Dimstiller: Workflows for dimensional analysis and reduction. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology. Salt Lake City, USA, 2010: 3-10

14 Yang J, Patro A, Huang S, et al. Value and relation display for interactive exploration of high dimensional datasets. In: Proceedings of the IEEE Symposium on Information Visualization. Austin, USA, 2004: 73-80

15 Lin H, Gao S, Gotz D, et al. RCLens: Interactive rare category exploration and identification[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(7), 2223-2237

16 Seo J, Shneiderman B. A rank-by-feature framework for interactive exploration of multidimensional data. Information visualization, 2005, 4(2): 96-113

17 Piringer H, Berger W, Hauser H. Quantifying and comparing features in high-dimensional datasets. In: Proceedings of the International Conference Information Visualisation. London, UK, 2008: 240-245

18 May T, Bannach A, Davey J, et al. Guiding feature subset selection with an interactive visualization. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology. Providence, USA, 2011: 111-120

19 Johansson S, Johansson J. Interactive dimensionality reduction through user-defined combinations of quality metrics. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 993-1000

20 Krause J, Perer A, Bertini E. INFUSE: Interactive feature selection for predictive modeling of high dimensional data. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1614-1623

21 Brooks M, Amershi S, Lee B, et al. FeatureInsight: Visual support for error-driven feature ideation in text classification. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology. Chicago, USA, 2015: 105-112

22 Liu S, Xiao J, Liu J, et al. Visual diagnosis of tree boosting methods. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 123-132

23 Zhu J, Ning C, Eric P Xing. Bayesian inference with posterior regularization and applications to infinite latent SVMs. Journal of Machine Learning Research, 2014, 15(1): 1799-1847

24 Zhou D, Basu S, Mao Y, et al. Learning from the wisdom of crowds by minimax entropy. In: Proceedings of Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012: 2195-2203

25 Zhou D, Liu Q, Platt J, et al. Aggregating ordinal labels from crowds by minimax conditional entropy. In: Proceedings of the International Conference on Machine Learning. Beijing, China, 2014: 262-270

26 Ren D, Amershi S, Lee B, et al. Squares: Supporting interactive performance analysis for multiclass classifiers. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 61-70

27 Donahue J, Jia Y, Vinyals O, et al. DeCAF: A deep convolutional activation feature for generic visual recognition. In: Proceedings of the International Conference on Machine Learning. Beijing, China, 2014: 647-655

28 Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2921-2929

29 Lang K. Newsweeder: Learning to filter netnews. In: Proceedings of the International Conference on Machine Learning. Tahoe City, USA, 1995: 331-339

30 Han E H S, Karypis G. Centroid-based document classification: Analysis and experimental results. In: Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery. Lyon, France, 2000: 424-431

31 Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of the European Conference on Machine Learning. Chemnitz, Germany, 1998: 137-142

32 Nene S A, Nayar S K, Murase H. Columbia object image library. Technical Report: CUCS-005-96, 1996

33 Vapnic V. The nature of statistical learning theory. Berlin, Germany: Springer Science&Business Media, 1995

34 Krizhevsky A. Learning multiple layers of features from tiny images. Technical report, 2009

35 Xiao J, Liu M, Liu S. A Visual Analysis System for News Data (in Chinese). Journal of Computer-Aided Design & Computer Graphics, 2016, 28(11): 1863-1871 [肖剑楠, 刘梦尘, 刘世霞. 新闻数据可视分析系统. 计算机辅助设计与图形学学报. 2016, 28(11):1863-1871]

36 Wu Y, Cui W, Song Y, et al. A Survey on Topic-Based Visual Text Analytics (in Chinese). Journal of Computer-Aided Design & Computer Graphics, 2012, 24(10): 1266-1272 [巫英才, 崔为炜, 宋阳秋, 等. 基于主题的文本可视分析研究. 计算机辅助设计与图形学学报. 2012, 24(10):1266-1272]

37 Jiang L, Liu S, Chen C. Recent research advances on interactive machine learning. Journal of Visualization. 2018, Nov(12):1-17

38 Zhou Z. Abductive learning: towards bridging machine learning and logical reasoning. Science China - Information Sciences. 2019, 62(7): 076101:1-076101:1-3

An interactive feature selection method based on learning-from-crowds

Changjian CHEN[1], Liu JIANG[1], Na LEI[2] & Shixia LIU[1*]

1. *Tsinghua University, Beijing* 100084*, China*;
2. *Dalian University of Technology, Dalian* 116024*, China*
* Corresponding author. E-mail: shixia@tsinghua.edu.cn

Abstract Ensemble feature selection algorithms aggregate the results of multiple feature selection methods for more effectively selecting a subset of features. However, these ensemble algorithms usually treat each feature selection method equally, without considering their performance differences, which may fail to include the features selected by a relatively smaller number of methods. To tackle this problem, we propose an interactive feature selection method, which can more effectively aggregate the results of multiple feature selection methods and iteratively improve the selected features by integrating expert knowledge. The proposed method includes a learning-from-crowds-based ensemble feature selection algorithm and a visual analysis system. The algorithm models the performance of multiple feature selection methods, calculates their reliabilities, and aggregates their results. To integrate expert knowledge, the visual analysis system provides a set of ranking schemes to assist experts in understanding the result of an individual feature selection method and the roles played by the features in classification tasks. A numerical experiment conducted on four real-world datasets shows that the proposed algorithm can improve the classification accuracy by 0.63-2.85% compared with the state-of-the-art ensemble feature selection algorithms. We also conducted two case studies on text and image data to demonstrate that the proposed visual analysis system can further improve the classification accuracy by 0.28-5.24%.

Keywords Ensemble feature selection, learning-from-crowds, visual analysis, interactive visualization, ranking visualization

Changjian CHEN was born in 1994. He received a BS degree in electronic engineering from University of Science and Technology of China. Currently, he is a Ph.D. candidate at Tsinghua University. His research interests include learning from crowds and interactive machine learning.

Liu JIANG was born in 1996. He received the bachelor degree in electronic engineering from University of Science and Technology of China, Hefei, in 2016. He is a Ph.D. candidate. His research interests include learning from crowds, interactive machine learning and high-dimensional data visualization.

Na LEI was born in 1977. She is currently a professor at the DUT-RU International School of Information Sciences and Engineering at Dalian University of Technology. She was a visiting professor at the University of Texas at Austin from 2007 to 2008, at the State University of New York at Stony Brook from 2014 to 2015 and at Tsinghua University from 2015 to 2016. Her research interests include computational geometry, computer graphics, and computer vision.

Shixia LIU was born in 1974. She is an associate professor at Tsinghua University. Her research interests include visual text analytics, visual social analytics, interactive machine learning, and text mining. She worked as a research staff member at IBM China Research Lab and a lead researcher at Microsoft Research Asia. She received a B.S. and M.S. from Harbin Institute of Technology, a Ph.D. from Tsinghua University. She is an associate editor-in-chief of IEEE Trans. Vis. Comput. Graph.