半监督特征选择综述 *

张东方 [a]，陈海燕 [a, b†]，王建东 [a]

(南京航空航天大学 a. 计算机科学与技术学院; b. 软件新技术与产业化协同创新中心, 南京 211100)

摘 要: 随着信息技术的迅猛发展，在实验实践中获取完整的有标记数据集变得更加困难。如何针对半监督数据集，利用不完整的监督信息完成特征选择，已经成为模式识别与机器学习领域的研究热点。为方便研究者系统地了解半监督特征选择领域的研究现状和发展趋势，对半监督特征选择方法进行综述。首先探讨了半监督特征选择方法的分类，将其按理论基础的不同分为基于图的方法、基于伪标签的方法、基于支持向量机的方法以及其他方法；然后详细介绍并比较了各个类别的典型方法；之后整理了半监督特征选择的热点应用；最后展望了半监督特征选择方法未来的研究方向。

关键词: 机器学习；半监督学习；特征选择

中图分类号: TP391.4 doi: 10.19734/j.issn.1001-3695.2020.01.0001

Survey of semi-supervised feature selection methods

Zhang Dongfang[a], Chen Haiyan[a, b†], Wang Jiandong[a]

(*a. School of Computer Science & Technology*, *b. Collaborative Innovation Center of Novel Software Technology & Industrialization*, *Nanjing University of Aeronautics & Astronautics*, *Nanjing Jiangsu* 211100, *China*)

Abstract: With the rapid development of information technology, it is more difficult to obtain complete labeled data sets in the experiment and real-world applications. How to select features on semi-supervised data sets by incomplete supervisory information has become a research hotspot in the field of pattern recognition and machine learning. In order to facilitate researchers to systematically understand the research status and development trend of semi-supervised feature selection, this paper reviewed the semi-supervised feature selection methods. This paper first discussed the classification of semi supervised feature selection methods and divided them into graph-based methods, pseudo-label-based methods, SVM-based methods and other methods according to their theoretical basis, then introduced and compared typical methods for each category, and then sorted out hot applications of semi-supervised feature selection. Finally, this paper looked forward to the future research directions of semi-supervised feature selection.

Key words: machine learning; semi-supervised learning; feature selection

0 引言

特征选择是模式识别与机器学习的关键问题之一，它是削减假设空间大小、降低数据维度的重要方法。Liu 在文献[1]中指出，特征选择的动机是：降低特征空间的维度、加速学习算法的执行、提高学习算法的预测精度、改善特征的可视化和可理解性。特征选择是一个从原始特征集中选出特征子集的过程，其最主要任务是去除不相关特征及冗余特征，保留相关特征[2]。不相关特征是指与监督信息不相关，删除后不影响学习性能的特征；冗余特征通常是与监督信息相关，但同时又与其他特征高度相关，被删除后也不影响学习性能的特征。Dash 等人[3]指出了一个典型的特征选择方法应该包含的 4 个基本步骤：子集生成、子集评价、停止准则和结果验证。由于寻找一组最优的特征子集是 NP-难问题，只有问题维度极低时才能找到全局最优解，研究者们提出了很多启发式搜索策略。特征选择一般通过按一定的准则对特征进行排序，或者在不降低学习性能的前提下寻找一组最小的特征子集来完成[2]。

随着信息技术的迅猛发展，获取完整的有标记数据集变得更加困难，实践中常常需要处理部分标记信息缺失的数据集。在这种情况下，有监督特征选择方法无法利用全部的样本，而无监督特征选择方法无法利用部分的标签。故可以同时利用有标记数据和无标记数据进行特征选择的半监督特征选择技术更受青睐。它一方面可以挖掘全体数据样本的结构、分布信息，另一方面也可以利用少量的数据标签提供的类别信息。

本文对半监督特征选择方法进行了较为全面的研究和综述。许多研究者已经对特征选择方法进行过整理和总结[2,4~6]，如文献[4]依据特征选择与学习器的结合方式分类介绍了诸多方法，文献[5]集中讨论了基于图的特征选择方法，Cai 等人[2]讨论了机器学习任务中常用的特征选择方法等。本文重点讨论了半监督特征选择方法的分类，对目前常用的半监督特征选择方法进行了系统的整理和比较，整理了半监督特征选择的热点应用，并展望了该领域面临的挑战和发展趋势，对半监督特征选择方法进一步的应用、研究有一定指导意义。

1 半监督特征选择方法分类

特征选择方法按照其使用的训练数据的标记情况(有标记、无标记或部分标记)，可以被分为有监督、无监督以及半监督方法；按照其与后续学习模型的关系，可以被分为过滤型、包装型以及嵌入型。这两种分类策略并不冲突，是最被接受和广泛使用的分类角度。此外还有依据特征选择方法的

收稿日期: 2020-01-08；**修回日期:** 2020-03-25 **基金项目:** 中央高校基本科研业务费专项资金资助项目(NS2019054)

作者简介: 张东方(1996-)，男，山东泰安人，硕士研究生，主要研究方向为数据挖掘；陈海燕(1979-)，女(通信作者)，江苏常州人，讲师，硕导，博士，主要研究方向为机器学习、数据挖掘(chenhaiyan@nuaa.edu.com)；王建东(1945-)，男，教授，博导，博士，主要研究方向为人工智能、数据挖掘及信息安全.

理论基础、评价准则、搜索策略以及输出类型进行分类的多种分类策略，Cai 等人[2]对此进行了总结。

基于 Sheikhpour 等人[6]的研究，考虑到半监督特征选择方法与半监督机器学习之间的区别与联系，进一步对现有的方法及研究方向进行整理和归纳，本文将半监督特征选择方法分为:基于图的半监督特征选择方法、基于伪标签的半监督特征选择方法、基于支持向量机的半监督特征选择方法以及其他半监督特征选择方法。图 1 描述了本文对半监督特征选择方法进行分类后的层次结构。基于图的方法为数据集建立近邻图，并且根据特征对图几何结构的保持能力来识别相关特征。基于伪标签的半监督特征选择方法使用数据集的有标签数据训练用于标记无标签数据的分类器，通过生成伪标签的方式将半监督数据集转换为有监督数据集，进而使用有监督特征选择方法或其他学习器相关方法完成特征选择。基于支持向量机的半监督特征选择方法借助半监督支持向量机，在模型训练过程中完成特征选择。其他半监督特征选择方法主要包括基于稀疏模型、粗糙集以及信息论的方法，此类方法大多由某种有监督特征选择方法扩展而来。表 1 概括了基于图、基于伪标签以及基于支持向量机的三类半监督特征选择方法的优缺点。

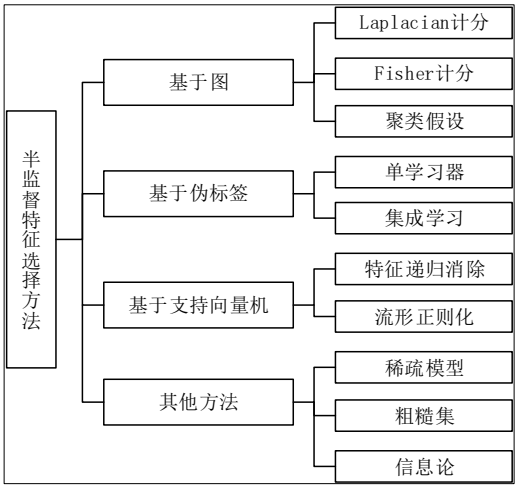

图 1　半监督特征选择方法的分类

Fig. 1　Categories of semi-supervised feature selection method

表 1　基于图、基于伪标签、基于 SVM 方法的优缺点

Tab. 1　General advantage and disadvantage of graph-based, pseudo-label-based and SVM-based methods

类型	优点	缺点
基于图	计算效率高	通常单独评价特征
	泛化能力强	忽视特征间相关性
	与分类器交换信息	计算成本高
基于伪标签	通常联合评价特征	低泛化性能
	考虑特征间依赖性	结果与分类器相关
	与分类器交换信息	
	联合评价特征	
基于支持向量机	考虑特征间依赖性	结果与分类器相关
	过拟合风险低	
	计算复杂度低	

此外，根据半监督特征选择方法的评价策略与学习算法的交互方式，半监督特征选择方法也可以分为过滤型(filter)、封装型(wrapper)和嵌入型(embedded)三种。过滤型半监督特征选择方法使用标记与未标记数据的内在特性，在不训练特定学习模型的情况下评价特征。封装型半监督特征选择方法往往利用单个学习器或集成学习模型预测未标记数据的标签，将半监督问题转换为有监督问题加以解决；或者直接基于学习器在不同特征集上的表现评价特征。嵌入型半监督特征选择方法在使用标记及未标记数据进行模型训练的过程中同时进行特征选择。表 2 概括了过滤型、封装型、嵌入型三类半监督特征选择方法的优缺点。

表 2　过滤型、封装型、嵌入型方法的一般优缺点

Tab. 2　General advantages and disadvantages of filter, wrapper and embedded methods

类型	优点	缺点
过滤型	计算效率高	忽视特征间相关性
	泛化能力强	通常单独评价特征
	易于实施和应用	
封装型	与分类器交换信息	计算成本高
	通常联合评价特征	低泛化性能
	考虑特征间依赖性	结果与分类器相关
	与分类器交换信息	
	联合评价特征	
嵌入型	考虑特征间依赖性	结果与分类器相关
	较低过拟合风险	
	较低计算复杂度	

在本文中，以 $X=[x_1,\ldots,x_l,x_{l+1},\ldots,x_n]\in R^{d\times n}$ 表示全体训练样本构成的矩阵，其中 l 为有标记样本的数量，n 为全体训练样本的数量，d 是每个样本的维度即全体特征的数量。$x_i\in R^d(1\le i\le n)$ 为第 i 个训练样本，f_r 为第 r 个特征且 f_{ri} 表示第 i 个样本的第 r 个特征值。令 $X_L=[x_1,\ldots,x_l]\in R^{d\times l}$ 表示 X 中的有标记样本集，$Y_L=[y_1,\ldots,y_l]\in\{1,\ldots,c\}$ 表示有标记样本集的标记向量，c 则是样本标签的类别总数。

2　基于图的半监督特征选择方法

图是一种重要的基础数据结构，很多真实数据集本身就具有图的结构。将每个数据样本视作带权图的一个顶点，再使用某种度量为带权图的边赋值，便可以建立原始数据集的图数据结构。基于图的半监督特征选择方法借助数据集的图数据结构来挖掘数据集的结构分布信息及类别分布信息，进而评价特征。该类方法都包含一个在半监督数据集上构造图的步骤，以及一个在构造的图上挖掘信息、评价特征的步骤。基于图的半监督特征选择方法不涉及具体学习器的训练，皆为过滤型方法，且借助谱图理论简化了推导与计算，拥有计算效率高、泛化能力强的优点。缺点是该类方法通常使用某种标准逐一独立评价特征，忽视特征间相关性。按照其理论基础与设计思路，主要可分为：基于 Laplacian 计分的方法、基于 Fisher 计分的方法、基于聚类假设的方法。

2.1　基于 Laplacian 计分

拉普拉斯计分(Laplacian score，LS)[7]是一种通过衡量特征的保持数据集局部结构的能力来评价特征、进行无监督特征选择的方法。其基本思想是：如果在原特征空间中距离相近的样本在某一特征的单维度上仍然保持近邻关系，那么该特征就被认为是保持了数据集局部结构，是应当被选择的特征。也就是说，被选择的特征保持了原始数据集中样本的分布结构信息。计算 Laplacian 计分包括三个主要步骤：1)构建一个包含 n 个节点的 p-最近邻图 G，并将其相似度 $S_{i,j}=\exp(-\|x_i-x_j\|/t)$ 作为图的边权值，其中 t 为一个常量。2)建立最近邻图的邻接矩阵 S，进而建立对角度矩阵 D，再建立拉普拉斯矩阵 $L=D-S$。3)用如下公式计算特征 f_i 的 Laplacian 计分：

$$\text{laplacian\_score}(f_i)=\frac{\tilde{f}_i^{\prime}L\tilde{f}_i}{\tilde{f}_i^{\prime}D\tilde{f}_i} \tag{1}$$

其中 $\tilde{f}_i$ 可以视作对特征向量 f_i 的标准化处理。特征的 Laplacian 计分越小，表明其对原始数据集结构的保持能力越强。

在半监督数据集中，全体样本定义了数据集的整体分布结构；而不完整的监督信息定义了数据集的类别分布结构。基于这一特点，不同研究者提出了不同的扩展方案，将 Laplacian 计分扩展并应用到半监督特征选择领域。

Zhao 等人[8]提出了一种基于 Laplacian 计分[7]和局部敏感判别分析的方法(locality sensitive discriminant feature，LSDF)。该方法使用半监督数据集建立了类内图 G^w 和类间图 G^b。类内图 G^w 连接了类别信息相同或者拥有近邻关系的样本，拥有近邻关系的样本之间的边权重为某一预设常数 γ，类别信息相同的样本之间边权重为 1；类间图 G^b 连接了类别信息不同的样本，边的权重都为 1。方法通过度量特征对数据集的结构保持能力来评价特征，由于类内图评估了样本间的相似性而类间图评估了样本间的不相似性，故被选择的特征应当拥有以下特性：在类内图连接的样本上特征值之差尽量小，在类间图连接的样本上特征值之差尽量大。借助谱图理论的推导及简化，对于第 r 个特征 f_r，LSDF 方法建立的特征评价准则为

$$I_r = \frac{f_r^T L^w f_r}{f_r^T L^b f_r} \tag{2}$$

其中 L_w、L_b 分别指类内图 G^w、类间图 G^b 对应的 Laplacian 矩阵。LSDF 方法最终输出全体特征按评价准则式(2)的排序。

Cheng 等人[9]定义了另一种半监督 Laplacian 评分(s-Laplacian score)以及分类信息增益(classification information gain，CIG)，并基于二者提出了一种半监督特征选择方法(graph-based semi-supervised feature selection，GSFS)。取参数 $\gamma=1$，GSFS 方法定义了与 LSDF 方法[8]相同的类内图 G^w 和类间图 G^b。其半监督 Laplacian 计分定义为

$$L_r = \frac{\tilde{f}_r^T L^w \tilde{f}_r}{\tilde{f}_r^T D^b \tilde{f}_r} \tag{3}$$

GSFS 方法同样得到全体特征按评价准则(3)的排序。在基于特征排序结果选出候选特征子集后，GSFS 方法设计了一个基于信息度量的冗余特征移除步骤。方法基于互信息定义了特征 f_i 关于特征 f_s 的分类信息增益：

$$\mathrm{CIG}(f_i, f_s) = \frac{I(Y; f_i \mid f_s)}{I(Y; f_i)} \tag{4}$$

GSFS 方法通过计算候选特征子集中两两特征间的分类信息增益，再结合二者半监督 Laplacian 计分大小关系，将二者中计分高且分类信息增益低的特征作为冗余特征加以去除，形成最终的特征子集。

除去常见的类别标签形式，成对约束也是监督信息的另一种表现形式。Benabdeslem 与 Hindawi[10]结合 Laplacian 计分[7]与约束得分[11]，定义了用于半监督特征选择的约束 Laplacian 计分，有效地度量了特征的局部结构保持能力和约束保持能力，提出了基于成对约束的半监督特征选择方法(constrained Laplacian score，CLS)。文章指出 Laplacian 计分法是评价特征局部结构保持能力的有效方法，但是它假设了同类样本的位置相近，这有时与实际情况是不相符的。另一方面，约束得分考量的是特征对成对约束信息的保持能力。在半监督的情境中，仅凭约束得分是不足以完成特征选择的，不能忽视无标记样本携带的信息。因此，Benabdeslem 与 Hindawi 提出使用全体训练样本及成对约束信息建立分别与 LSDF[8]中类内图 G^w、类间图 G^b 结构相同的近邻图 G^{kn}、负约束图 G^C (图的权值略有不同)，用于计算半监督约束 Laplacian 计分 C_r。其形式如下：

$$C_r = \frac{f_r^T L^{kn} f_r}{f_r^T L^C D^{kn} f_r} \tag{5}$$

基于 CLS[10]，Benabdeslem 等人[12]结合集成学习技术，还提出了更具效率与鲁棒性的集成约束 Laplacian 计分(ensemble constrained Laplacian score，EnsCLS)。Kalakech 等人[13]对基于成对约束监督信息的半监督特征选择方法进行了详细的探讨、实验和总结。

Sheikhpour 等人[14]通过将连续型标签信息转换为成对约束信息，设计了用于半监督回归问题特征选择的约束得分(semi-supervised constraint score，SSCS)。该方法处理有标记样本集，根据连续性标签值的大小判断有标记样本间的近邻关系，进而在近邻样本间添加正约束，在最远样本间添加负约束。在获得约束信息后，步骤与 CLS[10]方法基本相同。最终半监督约束得分形式为

$$SSCS_r = \frac{f_r^T L^{kn} f_r}{f_r^T L^C f_r} \tag{6}$$

同样针对回归问题，Sheikhpour 等人[15]还提出了基于图 Laplacian 的稀疏特征选择框架。

2.2 基于 Fisher 计分

Fisher 计分[16]是一种通过衡量特征类别判别能力来进行有监督特征选择的方法。在 Fisher 计分的基础上挖掘无标签样本携带的信息，也可以解决半监督特征选择问题。

Yang 等人[17]在保留 Fisher 计分原有特性的基础上，加入对数据集全局分布信息和局部结构信息的挖掘与利用，进而提出了用于半监督特征选择的半监督 Fisher 计分(semi-Fisher score)方法。半监督 Fisher 计分的形式如下：

$$SF_r = \frac{\sum_{i=1}^{c} n_i \left(\mu_r^i - \mu_r\right)^2 + \delta * V_r}{\sum_{i=1}^{c} n_i \left(\sigma_r^i\right)^2 + \lambda * J\left(f_r\right)} \tag{7}$$

其中 c 是类别序号，n_i 指第 i 个类中样本的数量，μ_r 定义了所有样本在第 r 个特征上的均值，μ_r^i、$(\sigma_r^i)^2$ 是第 i 类样本对应第 r 个特征的均值和方差。$\delta(\delta \geq 0)$ 与 $\lambda(\lambda \geq 0)$ 是两个控制参数。V_r 是第 r 个特征的方差计分[16]，度量了特征保存全局分布信息的能力；$J(f_r)$ 的定义如下：

$$J(f_r) = \sum_{i,j} \left(f_{ri} - f_{rj}\right)^2 S_{ij} = 2 f_r^T L f_r \tag{8}$$

$J(f_r)$ 度量了特征保存局部结构信息的能力。可以看出，Fisher 计分是半监督 Fisher 计分在参数 δ 与 λ 取值为 0 时的特殊情况。半监督 Fisher 计分可以有效的利用隐含在所有标记和未标记样本中的局部结构和全局分布信息，进而在半监督情境中评价特征。半监督 Fisher 计分法同样输出全体特征按评价准则式(7)的排序。

Hasanloei 等人[18]针对回归问题，提出了半监督 Fisher 与 Laplacian 计分(Semi-supervised Fisher and Laplacian score，SSFLS)。该方法处理有标记样本集，根据连续性标签值的大小判断有标记样本间的近邻关系；再结合全体样本在特征空间的近邻关系，为回归问题建立近邻图 G^{knn} 与远邻图 G^{kfn}。这与 LSDF[8]中的类内图 G^w 与类间图 G^b 相对应，计算细则也相同。SSFLS 结合了 Fisher 准则以及基于样本最近邻与最远邻计算的 Laplacian 图来评估特征。与 LSDF[8]类似，SSFLS 方法制定的的特征半监督 Fisher 与 Laplacian 计分为

$$SSFLS_r = \frac{f_r^T L^{knn} f_r}{f_r^T L^{kfn} f_r} \tag{9}$$

2.3 基于聚类假设

基于图的半监督特征选择方法也可以不借助已有的有监督或无监督评分方法，直接在半监督数据集的图数据结构上建立特征评价准则。

Zhao 和 Liu[19]提出了一种基于谱图理论和聚类假设的半监督特征选择方法(semi-supervised feature selection based on

spectral analysis，sSelect)。该方法通过为每个特征计算聚类指标，以评价聚类指标指代的聚类结果的方式评估所有特征。方法首先建立了 n 个节点的 k 最近邻图 G，通过 RBF 核函数为无向图 G 的边赋予权重，生成了图的邻接权重矩阵 S，进而得出度矩阵 D 与拉普拉斯矩阵 L。然后使用文献[20]中的方法，通过求解最小割问题的松弛形式为每个特征 f 生成对应的聚类指标 g。最后通过度量聚类指标 g 指代的聚类结果的簇间可分离性和与监督信息一致性来评估特征 f。其中 g 的可分离性使用最小割问题的目标函数值进行度量，而一致性则使用 $\hat{g}=\text{sign}(g)$ 与有标记信息标签值的互信息进行度量，最终的评估标准如下：

$$S_f=\lambda\frac{g^TLg}{g^TDg}+(1-\lambda)(1-\text{NMI}(\hat{g},y)) \tag{10}$$

其中 λ 是正则化参数，$\text{NMI}(\hat{g},y)$ 是 $\hat{g}$ 与 y 的标准化互信息。该方法独立地按照制定的评价准则评估每个特征，将特征按评估结果排序后再选出所需数量的 S_f 较小的特征组成特征集。

2.4 小结

基于图的半监督特征选择方法都使用了原始数据集的图数据结构以完成特征选择。基于 Laplacian 计分的方法从无监督特征选择 Laplacian 计分法入手，加入对有标签样本监督信息(类别标签或成对约束)的挖掘与利用，建立代表样本间相似性与不相似性的图数据结构，通过评价特征对原始数据集的结构保持能力完成特征选择。基于 Fisher 计分的半监督特征选择方法以有监督特征选择 Fisher 计分法为基础，加入对无标签样本特征信息的挖掘与利用，建立样本最近邻图，通过在 Fisher 计分的基础上加入对特征结构保持能力的评价完成特征选择。基于聚类假设的方法代表了不借助已有有监督或无监督特征评价标准，直接借助半监督数据集的图数据结构建立特征评价准则的方法。总的来说，该类方法皆属于过滤型方法，借助谱图理论简化了计算，拥有易于实施、计算效率高、泛化能力强的优点。但该类方法从输出类型[2]的角度看属于特征排序型，即输出的是特征依评价准则的排序，并不能完全解决特征选择问题得到最优解，通常需要用户指定最终特征子集的大小以贪婪地形成特征子集。且由于评价准则在制定时并未完全考虑特征间的相关关系，特征间评估完全独立，最终特征子集拥有一定的冗余性。表 3 对基于图的半监督特征选择方法的各典型方法从理论基础、特点、优缺点、以及时间复杂度方面进行了进一步概括与对比，其中 d 为原始特征集含有的特征数量，n 为全体训练样本的数量，c 为样本的类别总数。

表 3　基于图的半监督特征选择方法

Tab. 3　Graph-based semi-supervised feature selection method

算法简称	理论基础	特点	优缺点	时间复杂度
LSDF	Laplacian 计分	建立类内、类间图 根据特征结构保持能力评估特征	独立评价特征 忽视特征间相关性	$O(d\max(n^2,\log d))$
GSFS	Laplacian 计分 信息论	建立类内、类间图 根据特征结构保持能力评估特征 根据分类信息增益判别冗余特征	独立评价特征 移除了冗余特征	$O(d\max(n^2,\log d))$
CLS	Laplacian 计分 约束计分	建立近邻图、负约束图 使用成对约束形式监督信息 根据约束和结构保持能力评估特征	独立评价特征 忽视特征间相关性 对监督信息质量要求高	$O(d\max(n^2,\log d))$
SSCS	Laplacian 计分 约束计分	建立近邻图、负约束图 根据特征结构保持能力评估特征	独立评价特征 忽视特征间相关性 可以应对回归问题	$O(d\max(n^2,\log d))$
Semi-Fisher Score	Laplacian 计分 方差计分 Fisher 准则	建立近邻图 根据结构保持及判别能力评估特征	独立评价特征 忽视特征间相关性	$O(d\max(c,n^2,\log d))$
SSFLS	Laplacian 计分 Fisher 准则	建立近邻图、远邻图 根据结构保持及判别能力评估特征	独立评价特征 忽视特征间相关性 可以应对回归问题	$O(d\max(n^2,\log d))$
sSelect	聚类假设	建立近邻图 建立聚类指标	独立评价特征 忽视特征间相关性	$O(d\max(n^2,\log d))$

3　基于伪标签的半监督特征选择方法

基于伪标签的半监督特征选择方法又称基于自训练或协同训练的半监督特征选择方法[6]。该类方法从半监督数据集的样本入手，通过训练单个学习器或集成学习模型以标记所有或绝大多数无标签样本，之后再使用有监督特征选择或其他学习器有关的方法完成特征选择。该类方法都包含一个训练模型并为无标记样本生成伪标签的步骤。基于伪标签的半监督特征选择方法通常需要迭代地训练模型数次，不断采纳模型预测结果扩大标记训练集规模，属于封装型特征选择方法；在特征选择过程中考虑了特征间相关性，特征选择效果好，但其计算复杂度高、泛化性能差且容易出现过拟合的现象。

3.1 基于单学习器

基于单学习器的封装型半监督特征选择方法采用自训练的方式训练单个学习器，进而完成特征选择。

在文献[21]中，Ren 等人提出 FW-SemiFS(Forward semi-supervised feature selection)方法。该方法是该类方法的典型。FW-SemiFS 方法首先在有标记数据集上执行一个有监督特征选择方法 SFFS(Supervised sequential forward feature selection)[22]，然后使用所选特征训练一个简单的分类器以标记无标记样本，之后随机采纳预测结果以扩大有标记样本集的规模。方法迭代的执行上述步骤并记录 SFFS 的结果，将 SFFS 结果中出现频率最高的特征作为最终选择的特征，进而构建特征子集。FW-SemiFS 方法作为一种封装型方法丢失了可以联合评估特征的优势，计算复杂度高。

文献[23]介绍了一种基于 Biased-SVM 以及遗传算法的半监督特征选择方法 BSGA(Semi-supervised feature selection based on biased-SVM and genetic algorithm)。该方法首先通过一个考虑了类别不平衡问题的支持向量机预测无标记样本，形成有标记数据集。再以考虑了类别不平衡的分类准确率为

适应度函数，使用遗传算法在有标记数据集上完成特征选择。BSGA 方法可以较好的完成不平衡数据集的半监督特征选择。

3.2 基于集成学习

基于集成学习的封装型半监督特征选择方法通常以自训练或协同训练的方式训练多个分类器，进而组成集成学习器为无标记数据生成伪标签。与基于单学习器的方法不同，基于集成学习的方法在采纳无标记样本的预测结果时，通常要考虑预测结果的置信度。此类方法通常会采用 bagging 或 RSM[24]的集成学习策略为每个学习器生成不同的训练集。

使用集成学习技术，Hand 等人针对 FW-SemiFS 方法的不足之处加以改进,提出了 EN-SemiFS(ensemble-based semi-supervised feature selection) [25]方法。EN-SemiFS 方法使用 bagging 的集成学习策略，采用自训练方式训练单个分类器，进而形成集成分类器预测未标记样本的标签，并根据置信度采纳预测结果。迭代执行上述样本标记步骤数次后，最后使用有监督特征选择方法 SFFS [22]得到最终的特征子集。FW-SemiFS 方法旨在在循环中完善特征选择，故其需要执行 SFFS 多次来反复评估特征；而 EN-SemiFS 方法旨在循环中谨慎的增大有标记数据集的规模，故其只需要在数据集达到一定程度，充分挖掘整个数据集信息标记无标签数据后，执行一次 SFFS 选出相关特征即可。这一特性优化了 EN-SemiFS 方法计算资源的配比。

Barkia 等人[26]提出了一种基于集成学习的半监督封装式特征选择方法(semi-supervised feature improtance evaluation method，SSFI)，方法使用协同训练的策略训练了一个集成分类器，并将标准随机森林模型中的特征重要性评估方法扩展后应用到半监督场景中。SSFI 方法为每个小分类器生成不同的训练集以及不参与模型训练的 OOB(out of bags)样本。然后以协同训练的方式训练集成分类器，进而依置信度采纳无标签样本的预测结果。迭代地执行上述步骤直至集成分类器稳定后，最后执行一个基于重排列的特征评估方法评价并排序特征。与同种类的其他方法相比[21,25]，SSFI 方法的优势在于：1)使用协同训练得到了更好的集成学习器；2)提出了一种新的基于集成机制的特征评估方案，该评价方案较好的使用了 OOB 样本中的信息；3)将置信度测量应用到预测结果采纳、特征评估中。

Bellal 等人在文献[27]提出的 SEFR(semi-supervised ensemble learning guided feature ranking method)方法与上述的 SSFI[26]方法十分类似。二者皆采用了 bagging 的集成学习策略，使用了相同的基于重排列的特征评估方案以排序特征。不同之处在于，两种方法对集成学习框架的使用以及分类器训练策略的选择。SEFR 方法在使用 bagging 的集成学习策略，在为每个分类器生成对应训练集以及 OOB 样本后，为每个单学习器分别执行迭代自训练策略，在迭代中将标记后的无标签样本依置信度逐渐加入有标记样本集中；然后每个单学习器独立使用其对应的 OOB 样本评估所有特征。SEFR 方法的单个分类器在置信度计算、迭代训练、特征评估步骤中皆未与其他分类器及数据集产生信息交换。故 SSFI 使用协同训练策略训练分类器，形成集成分类器进而评估特征；而 SEFR 使用自训练策略训练分类器，集成的是每个分类器独立进行地对特征的最终评价。

Xiao 等人[28]提出了一种基于 GMDH(group method of data handling)神经网络的半监督特征选择方法(GMDH-based semi-supervised feature selection，GMDH-SSFS)，该方法可以良好地适应标记信息偏少的情形。在执行与同类其他方法[25~27]相似地样本标记步骤之后，GMDH-SSFS 方法通过训练一个 GMDH 神经网络来完成最终的特征选择。GMDH 神经网络是自组织数据挖掘领域的核心技术[29]，它可以自组织地选择所需的特征、建立模型的结构、设定参数的数值。GMDH 神经网络以全体特征为初始输入层，以固定的方式进行组合，采用最小二乘法估计参数，生成下一层候选模型；再通过外准则对候选模型进行筛选，保留一部分作为新的输入层；方法迭代执行，模型的复杂度不断增加，直至得到目标模型停止。最后，检查与目标模型有关的初始输入特征既是特征选择结果。GMDH-SSFS 方法的另一个特点是采用过采样技术处理了数据集中的样本不平衡问题。

3.3 小结

基于伪标签的半监督特征选择方法都存在一个无标签样本标记步骤。样本标记步骤通常是一个迭代的"预测-选择"过程，即先训练模型对样本进行预测，然后再根据置信度选择一定的无标签样本，采纳其预测结果，将其作为有标签样本参与下次迭代。置信度的获取难度与分类器类型有关，基于单学习器的方法受到分类器类型的影响，可能难以获取预测结果的置信度[25]。基于集成学习的方法通常采用众分类器预测结果的一致性作为其置信度度量。采用集成学习不仅提高了预测结果的准确率，而且自然地支持了本文对结果置信度的需求。在样本标记步骤完成之后，半监督特征选择就可以通过有监督特征选择方法或者学习器有关方法完成了。表 4 对基于伪标签的半监督特征选择方法的典型方法从所用分类器类型、训练方式以及特征评估方案方面进行了进一步概括与对比。

表 4 基于伪标签的半监督特征选择方法

Tab. 4 Pseudo-label-based semi-supervised feature selection method

算法简称	分类器类型	训练方式	特征评估方案
FW-SemiFS	单学习器	自训练	SFFS 结果频度
BSGA	单学习器	自训练	分类器精度
EN-SemiFS	集成分类器	自训练	SFFS 结果
SSFI	集成分类器	协同训练	基于重排列
SEFR	集成分类器	自训练	基于重排列
GMDH-SSFS	集成分类器	协同训练	GMDH 训练

4 基于支持向量机的半监督特征选择方法

半监督支持向量机(S^3VM)起初被称作直推式支持向量机(transductive support vector machine，TSVM)[30]。半监督支持向量机的目标是结合有标记及无标记样本信息，寻找拥有最大分类裕度的决策边界。基于支持向量机的半监督特征选择方法在训练半监督支持向量机的过程中完成特征选择，属于嵌入型特征选择方法。该类方法拥有计算效率高、过拟合风险低的优点，缺点是选择结果与分类器类型相关；主要分为基于特征递归消除的方法以及基于流形正则化的方法。

4.1 基于特征递归消除

基于 SVM 的递归特征消除(SVM-based recursive feature elimination，SVM-RFE)[31]是一种著名的有监督嵌入式特征选择方法。SVM-RFE 通过训练一个 SVM，将特征按对应的权重排序，使用后向消除的贪心策略逐步消除不重要特征，最终形成特征子集。Ang[32]提出了 SVM-RFE 方法在半监督情境中的两种拓展：基于半监督支持向量机的递归特征消除(S^3VM-based recursive feature elimination,，S^3VM-RFE)以及基于半监督支持向量机的特征选择(S^3VM-based feature selection，S^3VM-FS)。

S^3VM-RFE 方法首先将有标记数据集进行采样，形成有标记训练集。未被选中的样本被加入无标记数据集中，共同视作无标记训练集。使用有标记训练集训练一个有监督 SVM[31]，使用无标记训练集训练一个无监督单类 SVM[33]。特征权重向量是由有监督 SVM 和无监督 SVM 一起给出的，作为最终的评价标准。方法每次迭代将权重最小的特征剔除。

S³VM-FS 方法与 S³VM-RFE 方法步骤大致相同,但没有采用迭代执行的机制。S³VM-RFE 方法每次迭代剔除当前特征集中权重最小的特征，特征排序为多次迭代中特征被剔除顺序；而 S³VM-FS 仅执行一次训练 SVM 的过程，特征排序的依据为单次训练结果中特征权重的大小。

4.2 基于流形正则化

Xu 等人[34]提出了一个基于流形正则化的 SVM 嵌入型半监督特征选择方法(discriminative semi-supervised feature selection method based on manifold regularization，FS-manifold)。文章提出了一个结合分类裕度和流形正则化的评价准则，最优特征子集由此选出。在特征选择中考虑流形正则化以确保决策函数在所选特征的数据所形成的流形上是光滑的。文章从 Belkin 等人[35]提出的流形正则化框架开始，经过半监督支持向量机的对偶问题形式，将半监督特征选择等价变换为一个最优解在鞍点处的 min-max 优化问题(同时是组合优化问题)。然后在讨论了半监督特征选择和多核学习[36]的关系之后，使用多核学习的一种 Level method[37]求解了该 min-max 优化问题，从而完成了特征选择。方法总时间复杂度为 $O(n^{2.5}/\varepsilon^2)$ 。

4.3 小结

基于支持向量机的半监督特征选择方法综合挖掘有标签样本与无标签样本携带的信息，在训练半监督支持向量机的过程中完成特征选择。基于特征递归消除的方法是有监督特征选择方法(SVM-RFE)在半监督问题上的扩展，拥有不易过拟合、计算效率高的优点；但其结果与采用的半监督支持向量机有关。基于流形正则化的方法更注重最大化类间裕度，考虑了全体样本的局部结构；但其目标函数复杂，求解时通常需要解决一个复杂的优化问题。表 5 对基于支持向量机的半监督特征选择方法的典型方法从理论基础以及优缺点方面进行了进一步概括与对比。

表 5　基于支持向量机的半监督特征选择方法

Tab. 5　SVM-based semi-supervised feature selection method

算法简称	理论基础	优缺点
S3VM-RFE	半监督 SVM	仅适用二分类问题
	递归特征消除	准确度较高
S3VM-FS	半监督 SVM	仅适用二分类问题
	递归特征消除	执行速度快
FS-Manifold	半监督 SVM	较好的鲁棒性
	流形正则化框架	目标函数求解困难

5　其他半监督特征选择方法

5.1 基于稀疏模型

基于稀疏模型的半监督特征选择方法使用多种稀疏模型以完成特征选择，同时挖掘并利用有标记数据集与无标记数据集的信息。在半监督特征选择领域，目标是找到一个可以最佳地保存判别信息和训练数据分布信息的转换矩阵 $W\in R^{k\times c}(k<d)$ ，进而使用 W 的行向量范数来计算对应特征的权重。如果矩阵 W 的某些行收缩到零向量，则 W 可被用于特征选择，即零向量对应的特征被从特征集中移除[38]。考虑 $W\in R^{d\times c}(c<d)$ 为用于稀疏特征选择的转换矩阵，则半监督稀疏特征选择方法的目标函数通常为

$$\min_{W} \text{loss}(W)+\lambda\| W\|_{2,p}^{p} \tag{11}$$

其中 $\text{loss}(\cdot)$ 为损失函数，λ 为正则化参数。 $\|W\|_{2,p}$ 为矩阵的 $l_{2,p}$ 范数。求解该优化问题后,使用矩阵 W 的行向量范数来计算对应特征的权重，进而完成特征选择。

Han 等人[39]提出了一个基于稀疏模型的过滤式半监督特征选择方法(semi-supervised feature selection via spline regression，S^2FS^2R)，该方法引入了一种组合的半监督散度矩阵。半监督散度矩阵使用 Fisher 判别分析的类内散度矩阵来保存有标记样本的判别信息，使用样条散度矩阵来保存有标记及无标记数据的局部几何结构信息。基于文献[40]所述的降维框架，半监督稀疏特征选择的目标函数可以被定义为

$$\min_{W^TW=1} \text{Tr}\left(W^TMW\right)+\lambda\| W\|_{2,1} \tag{12}$$

其中正则化项 $\| W\|_{2,1}$ 用于保证 W 行向量的稀疏性，使得 W 可以用于特征选择，参数 $\lambda>0$ 控制正则化效果，矩阵 $M\in R^{d\times d}$ 是一个编码了标签信息及数据局部结构的半监督散度矩阵。正交约束 $W^TW=I$ 被用来避免解的任意缩放以及全零的平凡解。半监督散度矩阵 M 被定义为

$$M=A+\mu D \tag{13}$$

其中矩阵 $A\in R^{d\times d}$ 是一个储存了有标记数据标签信息的散度矩阵，矩阵 $D\in R^{d\times d}$ 是一个储存了有标记及无标记数据局部结构信息的散度矩阵，参数 $\mu(0\leq\mu\leq1)$ 控制了矩阵 D 的权重。算法的总时间复杂度为 $O(d^3)$ 。

Chen 等人[41]提出了一种基于重标度线性回归的嵌入型半监督特征选择方法(rescaled linear square regression，RLSR)。方法在训练稀疏线性回归模型的同时，将无标签样本的标签表示矩阵 Y_U 也作为优化目标变量，该模型的优化目标函数为

$$\begin{aligned}&\min\left(\|X^TW+\mathbf{1}b^T-Y\|_F^2+\gamma\| W\|_{2,1}\right)\\&\text{st. } W,b,Y_U\geq0, Y_U\mathbf{1}=\mathbf{1}, M\geq0\end{aligned} \tag{14}$$

Yuan 等人[42]在 Chen 等人[41]的基础上提出了一种名为判别半监督特征选择(discriminative semi-supervised feature selection，DSSFS)的嵌入型半监督特征选择方法。DSSFS 方法将 *ϵ–dragging* 技术引入 RLSR 中的稀疏线性回归模型中，以起到增大类间裕度的作用。此时优化目标函数的形式为

$$\begin{aligned}&\min\left(\left\|X^TW+\mathbf{1}b^T-Y-(2Y-\mathbf{1})\circ M\right\|_F^2+\gamma\| W\|_{2,1}^2\right)\\&\text{st. } W,b,Y_U\geq0, Y_U\mathbf{1}=\mathbf{1}, M\geq0\end{aligned} \tag{15}$$

其中矩阵 M 为待求解的 *ϵ–dragging* 矩阵。

基于稀疏模型的半监督特征选择方法是近年来研究热点之一。许多研究者尝试将不同的技术或准则融合到半监督稀疏特征选择框架中。文献[43]引入 L_1 范数图来获取流形结构信息，在进行稀疏特征选择的过程中同时尽可能保持数据集流形结构。 文献[44]通过建立 Laplacian 矩阵进而在稀疏特征选择框架中引入 Laplacian 正则化项，在稀疏特征选择的过程中保存了样本空间分布信息。

5.2 基于粗糙集

知识约简是粗糙集理论的核心内容之一，特征选择的过程既是知识库属性约简的过程。在粗糙集理论中，所谓属性约简，就是在保持知识库分类能力不变的条件下，删除冗余、不相关属性的过程。一般来说，基于经典粗糙集的方法仅可以处理离散型(分类型)特征。一些基于广义粗糙集的方法可以克服上述缺点，例如基于邻域粗糙集的 NFARNRS 方法[45]。

Dai 等人[46]提出了两种基于粗糙集的过滤型半监督特征选择方法(semi-supervised attribute selection algorithm based on rough set throry， semirough-P / semirough-D)。基于粗糙集理论，借助可辨对、正域等概念，将半监督数据集视作有标记样本集与无标记样本集两部分，semirough-P 方法定义了半监督环境下的特征子集评估函数 IMP。评估函数 IMP 的一个优势是其不仅可以评估单个特征，也可以评估特征子集。对于某一特征子集 $F'\subseteq F$ ，其评估函数形式如下：

$$\text{Imp}_{F'}(D)=\frac{\gamma_{F'}(D)}{\gamma_F(D)}+\frac{|\text{Dis}^S(F')|}{|\text{Dis}^S(F)|} \tag{16}$$

其中集合 D 为标记信息的取值范围，式(14)右侧首项由有标记样本计算，其中 $\gamma_F(d)=|\text{POS}_F(d)|/|S|$, 为特征集 F 对应的正域

样本个数与全体样本个数的比值，称为标签对特征集 F 的依赖度；式(16)右侧第二项由无标记样本计算，其中 $|\mathrm{Dis}^S(F)|$ 为特征集 F 对应的可辨对数量。方法分为两个步骤，首先从空集 T 开始，逐一向集合 T 中添加特征直至集合 T 的评估结果与全体特征组成的集合相同，这对应了保持知识库分类能力不变的条件；然后逐一检查集合 T 中已有的特征，如果去掉某个特征不会影响集合 T 的评估结果，那么就从集合 T 中去除该特征，这对应了删除冗余、不相关属性的过程。

作者提出的两种方法(semirough-P，semirough-D)步骤相同，仅在特征集的评估函数上有所不同。semirough-D 的特征集评估函数为

$$\mathrm{Imp}_{F'}(D)=\frac{|\mathrm{Dis}_B^L(F')|}{|\mathrm{Dis}_B^L(F)|}+\frac{|\mathrm{Dis}^S(F')|}{|\mathrm{Dis}^S(F)|} \tag{17}$$

两种方法的总时间复杂度皆为 $O(dn^2)$。

Liu 等人[47]基于粗糙集理论，采用集成机制评价特征集，提出了基于粗糙集和集成选择器的半监督封装型特征选择方法(semi-supervised feature selection via ensemble selector，RSES)。文章讨论了 semirough-P[46]方法的两个不足并针对性的提出了解决方案。semirough-P 方法没有预测未标记样本的标签，可能会丢失一些有价值的信息，而 RSES 方法使用一种基于图的标签传播算法(LPA)[48]计算未标记样本的标签。semirough-P 方法仅使用了一种特征集评估函数，而 RSES 方法使用了集成评估方式。从邻域决策错误率[49]中得到启发，文章提出了局部邻域决策错误率来表征邻域分类器[49]在不同决策类上的表现，进而建立了集成评价机制。与 semirough-P[46]方法不同，RSES 方法不含从已选特征集中消除特征的过程。方法的总时间复杂度为 $O(dn^2)$。

5.3　基于信息论

基于信息论的特征选择方法是近年来的一个研究热点，其通过降低特征子集内部相关性、提高特征子集与类别相关性完成特征选择。信息熵理论不要求假定数据分布是已知的，能够以量化的形式度量特征间线性或非线性的相关关系。信息增益(IG)定义为先验不确定性与期望的后验不确定性之间的差异，互信息(MI)描述的是两个随机变量之间相互依存关系的强弱。二者是基于信息论特征选择方法的常用信息度量。

Wang 等人[50]基于马尔可夫毯和信息论，提出了一种基于信息论的过滤型半监督特征选择方法(semi-supervised representatives feature selection algorithm based on information theory，SRFS)。SRFS 方法在去除不相关特征和识别冗余特征组任务上表现十分出色。方法分为三个步骤：首先，识别并去除不相关特征。在有标记数据集上，特征的相关性可以使用互信息度量，与标签互信息大的特征被认为是相关特征；在无标记数据集上，由于缺乏标签信息，故应用最大熵原则，认为信息熵较大的特征更与标签相关。然后，基于 SRFS 方法定义的特征相关性与冗余性，判别任意两个特征之间的冗余关系。以特征为顶点冗余关系为边，为特征集建立有向非循环图；进而借助图的划分操作，将特征聚类成簇。最后，从每个簇中选出具有代表性的特征，由于之前冗余特征被划分为若干互不相关的特征簇，簇内的特征相似而不同簇的特征不相似；故该方法可以极大减轻特征维度的影响，高效的找到最佳特征子集。方法的时间复杂度为 $O(d^2n)$。

文献[51]提出了一种基于信息熵的半监督粗糙特征选择算法 SFSE(entro-based rough semi-supervised feature selection)。SFSE 方法基于互补信息熵的概念，给出了半监督数据集有标记数据、无标记数据以及通过有标记数据标注的数据的不确定性度量；并基于上述三部分数据熵的组合，定义了半监督数据集上的互补信息熵；进而重新定义了特征重要度，设计了特征选择方法。SFSE 采用了一种基于 K-近邻的无标记数据标记算法，标注了部分无标签数据。另外，SFSE 中的互补信息熵是基于粗糙集理论定义的。

5.4　小结

基于稀疏模型的方法通过寻找一个行稀疏的转换矩阵来完成特征选择。基于粗糙集及基于信息论的方法借助粗糙集或信息论的基本概念，通过定义半监督情境下的新特征度量评估特征。这些方法大多属于过滤型特征选择方法，可以视作有监督特征选择方法在半监督情境下的扩展。此外还有一些基于有监督特征选择方法设计的半监督特征选择方法，例如 Cheng 等人基于逻辑迭代 Relief[52]方法设计的半监督逻辑迭代 Relief[53]方法、Xu 等人基于 mRMR 准则(minimal-redundancy maximal-relevance criterion)[54]设计的 RRPC(max-relevance and min-redundancy criterion based on pearson's correlation)[55]方法等。表 6 对本章提到的各典型方法从理论基础、特点、优缺点以及时间复杂度方面进行了进一步概括和对比。

表 6　其他半监督特征选择方法

Tab. 6　Other semi-supervised feature selection methods

算法简称	理论基础	特点	优缺点	时间复杂度
S^2FS^2R	稀疏模型 嵌入式降维框架	根据判别及结构保持能力评估特征 使用 $l_{2,1}$ 范数获取稀疏性	联合评价特征 目标函数难以求解	$O(d^3)$
RLSR	稀疏线性回归模型	引入标量因子	结果与模型相关 结果可解释性强	—
DSSFS	稀疏线性回归模型 ϵ–*dragging* 技术	引入 ϵ–*dragging* 矩阵	结果与模型相关 结果精度高	—
semirough-P semirough-D	粗糙集理论	基于粗糙集的特征评估准则 在有/无标签数据集上分别评价特征	联合评价特征	$O(dn^2)$
RSES	粗糙集理论	基于粗糙集的特征评估准则 特征集成评价机制	联合评价特征 结果精度高	$O(dn^2)$
SRFS	信息论 马尔可夫毯	基于信息论定义了相关/冗余特征 将特征聚类成簇	有效去除不相关特征 彻底去除冗余特征	$O(d^2n)$
SFSE	信息论 粗糙集理论	定义了半监督意义下的信息熵	独立评价特征	—

6　半监督特征选择应用动态

特征选择是模式识别与机器学习的关键问题之一，它是削减假设空间大小、降低数据维度的重要方法。半监督特征选择被广泛的应用于半监督数据集的降维处理。本章节对半监督特征选择的热点应用进行了整理。

6.1 计算机视觉与图像处理

计算机科学与社交媒体的迅速发展生成了越来越多的高维多媒体数据集。随之而来的维度灾难问题给现有的多媒体检索与理解技术带来了挑战。针对这一问题，Wang 等人[56]分析了基于 Laplacian 计分的图方法的不足，基于稀疏特征选择框架，结合自适应局部流形学习技术，设计了一种嵌入式半监督特征选择方法。设计的方法成功应对了不同的多媒体分析任务，包括自然场景标记、目标检测、行为识别以及 Web 图像标记任务[56]。Zeng 等人[57]指出为大规模数据建立 Laplacian 矩阵过于耗时且不现实，设计了一种无须构造图的嵌入式半监督特征选择方法，在网页与图像注释任务上取得了良好的表现。针对 Web 图像分类任务，Song 等人[58]将特征选择过程融入保持样本流形结构的半监督学习中，设计了一种嵌入式半监督特征选择方法，在样本预测的过程中完成了特征选择。Shi 等人[59]将稀疏半监督特征选择应用于多视角数据的多媒体数据分析任务，在图像标注、视频概念检测以及 3-D 运动数据分析任务中取得了良好结果。针对网络图像标注问题，文献[60]提出了一种基于 Hessian 正则化的半监督特征选择方法，在实验数据集上取得了较高准确率。

超高分辨率遥感图像在例如水质检测与灾害预警的许多实际应用中得到了广泛应用。研究人员需要在大量可用特征中筛选具有类别判别力的特征。另一方面，相对低廉的采集成本与昂贵的标注成本在这一领域形成了大量半监督数据集。Chen 等人应用稀疏半监督特征选择，有效的改善了基于对象的图像分析方法在超高分辨率遥感图像上的性能[61]。更进一步，Chen 等人[62]提出将超高分辨率遥感图像看做多视图数据以更好的分析数据内在结构，并基于稀疏半监督特征选择框架设计了一种新的多视图半监督特征选择方法。实验证明，针对超高分辨率遥感图像，这种多视图方案与视图内特征是合理有效的，为分析超高分辨率遥感图像的数据结构提供了一种新方法[62]。

6.2 计算机辅助诊断

计算机辅助诊断常涉及使用算法进行决策，如通过机器学习的分类模型进行判别，为医生提供辅助意见。在诸多诊断工具数据中识别相关特征是计算机辅助诊断的关键[63]。针对阿兹海默症诊断问题，An 等人[63]借助半监督 Laplacian 矩阵来挖掘半监督样本集的结构信息，同时进行样本特征的评估，选出具有判别力的样本与特征；并将上述步骤迭代执行，以选出最终样本集与特征集。实验证明 An 等人设计的半监督特征选择方法在核磁共振成像(MRI)与单核苷酸多态性(SNP)数据集上取得了更好的性能，改善了辅助诊断结果[63]。Jiang 等人[64]将基于标签传播聚类的半监督特征选择方法应用于识别与不同临床表型相对应的疾病相关基因，进而了解疾病进展过程中的基因功能和受影响途径。

6.3 音频语料分析

在涉及音频数据分析的机器学习任务中，半监督特征选择也是一个关键问题。在音频分类任务中，存在着大量未被挖掘和利用的特征或特征组合信息。Yang 等人[65]基于 mRMR 特征选择框架[54]，以约束补偿 Laplacian 计分与一种半监督巴氏距离作为半监督相关性与冗余性度量，设计了一种应用于音频分类任务的半监督特征选择方法。实验证明 Yang 等人的方法可以有效移除不相关特征，显著提高分类精度[65]。

情感识别指从音频或语料中自动判断说话人情绪的任务，是改善人机交互体验的关键问题之一。针对语音情感识别问题，Sun 等人[66]在应用半监督特征选择方法的基础上结合了一种说话人归一化方法，实验证明了该方法对语音情感识别的有效性，可以在说话人样本数量较少的情况下取得良好的归一化效果。

6.4 药物设计

结构活性定量关系(quantitative structure activity relationship，QSAR)是药物设计领域一种有效的计算技术，它通过目标的化学结构信息来预测其生物反映活性。在 QSAR 研究中，很多抑制剂的抑制活性没有被经验确定，已知的抑制活性是通过耗时且昂贵的实验获得的，故半监督特征选择技术适于处理 QSAR 数据集。Sheikhpour 等人[15]设计了从回归问题数据集中提取成对约束信息的方案，进而应用半监督约束得分解决了 QSAR 问题的特征选择任务。实验结果说明，在 QSAR 模型上，基于全体样本局部结构以及成对约束信息进行特征选择的半监督方法显著优于仅使用有标记样本进行特征选择的有监督方法[15]。Hasanloei 等人[18]则设计了从回归问题数据集中定义样本间相邻关系的方案，进而应用半监督 Laplacian 得分解决了 QSAR 问题的特征选择任务。实验同样证明了相较有监督方法，Hasanloei 等人[18]方法在解决 QSAR 问题特征选择任务上的优越性。

7 结束语

处理实际应用中常见的监督信息不完整的数据集，有监督特征选择方法和无监督特征选择方法各有其局限性。半监督特征选择方法一方面可以挖掘全体数据样本的结构、分布信息，另一方面可以充分利用不完整的监督信息，从而更好地完成特征选择任务，起到削减假设空间大小、降低数据维度的作用。本文首先讨论了半监督特征选择方法的分类，探讨了各类方法的优缺点；然后对常用的半监督特征选择方法进行了综述，整理了其热点应用。

今后有待进一步研究的工作包括以下几点：

a)近年来，半监督特征选择在机器学习和计算机视觉等领域得到了广泛的研究与应用。大部分半监督特征选择方法是基于图的，并且它们的表现与其使用的 Laplacian 图密切相关。传统方法中使用的 Laplacian 图结构与性质彼此相似，建立后不能修改，限制了图方法的性能[67]。在特征选择过程中动态地更新 Laplacian 图信息是提高图方法性能的可行思路。

b)回顾半监督特征选择方法，大部分半监督特征选择方法是针对分类问题提出的，少有针对回归问题提出的半监督特征选择方法[68]。针对回归问题的半监督特征选择方法是当下急需探索的研究领域。除去本文讨论的图方法[14,18]，基于稀疏模型的半监督特征选择方法被认为是可行方案之一[6]。

c)可扩展性与稳定性是特征选择技术在准确率之后的更深层次追求[2]。可扩展性良好的方法应当拥有多项式时间的复杂度，能够在尺寸不一的数据集上统一取得良好结果。稳定性良好的方法应当不受样本集采样偏差的影响，在增删少量样本时结果保持一致。提高特征选择方法的可扩展性与稳定性，也值得进一步研究与探索。

参考文献：

[1] Liu Huan, Motoda H. Feature selection for knowledge discovery and data mining [M]. Springer Science & Business Media, 2012.

[2] Cai Jie, Luo Jiawei, Wang Shulin, *et al.* Feature selection in machine learning: A new perspective [J]. Neurocomputing, 2018, 300: 70-79.

[3] Dash M, Liu Huan. Feature selection for classification [J]. Intelligent data analysis, 1997, 1 (3): 131-156.

[4] 李邵琴, 杜建强, 聂斌, 等. 特征选择方法综述 [J]. 计算机工程与应用, 2019, 55 (24): 10-19. (Li Zhiqin, Du Jianqiang, Nie Bin, *et al.* Summary of feature selection methods [J]. Computer Engineering and Applications, 2017, 55 (24): 10-19.)

[5] 张文静, 王备战, 张志宏. 基于图的特征选择算法综述 [J]. 安徽大学学报 (自然科学版) , 2017, 041 (001): 10-20. (Zhang Wenjing, Wang

Beizhan, Zhang Zhihong. A survey of feature selection algorithms based on graph. Journal of Anhui University (Natural Science Edition) , 2017, 041 (001): 10-20.)

[6] Sheikhpour R, Sarram M A, Gharaghani S, *et al.* A survey on semi-supervised feature selection methods [J]. Pattern Recognition, 2017, 64: 141-158.

[7] He Xiaofei, Cai Deng, Niyogi P. Laplacian score for feature selection [C]// Advances in neural information processing systems. 2006: 507-514.

[8] Zhao Jidong, Lu Ke, He Xiaofei. Locality sensitive semi-supervised feature selection [J]. Neurocomputing, 2008, 71 (10-12): 1842-1849.

[9] Cheng Hongrong, Deng Wei, Fu Chong, *et al.* Graph-based semi-supervised feature selection with application to automatic spam image identification [C]// International Workshop on Computer Science for Environmental Engineering and EcoInformatics. Springer, Berlin, Heidelberg, 2011: 259-264.

[10] Benabdeslem K, Hindawi M. Constrained laplacian score for semi-supervised feature selection [C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2011: 204-218.

[11] Zhang Daoqiang, Chen Songcan, Zhou Zhihua. Constraint score: A new filter method for feature selection with pairwise constraints [J]. Pattern Recognition, 2008, 41 (5): 1440-1451.

[12] Benabdeslem K, Elghazel H, Hindawi M. Ensemble constrained Laplacian score for efficient and robust semi-supervised feature selection [J]. Knowledge and Information Systems, 2016, 49 (3): 1161-1185.

[13] Sheikhpour R, Sarram M A, Sheikhpour E. Semi-supervised sparse feature selection via graph Laplacian based scatter matrix for regression problems [J]. Information Sciences, 2018, 468: 14-28.

[14] Kalakech M, Biela P, Macaire L, *et al.* Constraint scores for semi-supervised feature selection: A comparative study [J]. Pattern Recognition Letters, 2011, 32 (5): 656-665.

[15] Sheikhpour R, Sarram M A, Gharaghani S. Constraint score for semi-supervised feature selection in ligand-and receptor-based QSAR on serine/threonine-protein kinase PLK3 inhibitors [J]. Chemometrics and Intelligent Laboratory Systems, 2017, 163: 31-40.

[16] Bishop C M. Neural networks for pattern recognition [M]. Oxford university press, 1995.

[17] Yang Ming, Chen Yinjuan, Ji Genlin. Semi_Fisher Score: A semi-supervised method for feature selection [C]// 2010 International Conference on Machine Learning and Cybernetics. IEEE, 2010, 1: 527-532.

[18] Hasanloei M A V, Sheikhpour R, Sarram M A, *et al.* A combined Fisher and Laplacian score for feature selection in QSAR based drug design using compounds with known and unknown activities [J]. Journal of computer-aided molecular design, 2018, 32 (2): 375-384.

[19] Zhao Zheng, Liu Huan. Semi-supervised feature selection via spectral analysis [C]// Proceedings of the 2007 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2007: 641-646.

[20] Shi J, Malik J. Normalized cuts and image segmentation [J]. IEEE Transactions on pattern analysis and machine intelligence, 2000, 22 (8): 888-905.

[21] Ren Jiangtao, Qiu Zhengyuan, Fan Wei, *et al.* Forward semi-supervised feature selection [C]// Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, Heidelberg, 2008: 970-976.

[22] Whitney A W. A direct method of nonparametric measurement selection [J]. IEEE Transactions on Computers, 1971, 100 (9): 1100-1103.

[23] 杜利敏，徐扬．一种面向不平衡数据的半监督特征选择算法 [J]. 河南理工大学学报：自然科学版, 2017, 36 (5): 095-099. (Du Limin, Xu Yang. A semi-supervised feature selection algorithm for imbalanced data [J]. Journal of Henan Polytechnic University (Natural Science) , 2017, 36 (5): 095-099.)

[24] Hady M F A, Schwenker F. Combining committee-based semi-supervised learning and active learning [J]. Journal of Computer Science and Technology, 2010, 25 (4): 681-698.

[25] Han Y, Park K, Lee Y K. Confident wrapper-type semi-supervised feature selection using an ensemble classifier [C]// 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC) . IEEE, 2011: 4581-4586.

[26] Barkia H, Elghazel H, Aussem A. Semi-supervised feature importance evaluation with ensemble learning [C]// 2011 IEEE 11th International Conference on Data Mining. IEEE, 2011: 31-40.

[27] Bellal F, Elghazel H, Aussem A. A semi-supervised feature ranking method with ensemble learning [J]. Pattern Recognition Letters, 2012, 33 (10): 1426-1433.

[28] Xiao Jin, Cao Hanwen, Jiang Xiaoyi, *et al.* GMDH-based semi-supervised feature selection for customer classification [J]. Knowledge-Based Systems, 2017, 132: 236-248.

[29] Lemke F, Müller J A. Self-organising data mining [J]. Systems analysis modelling simulation, 2003, 43 (2): 231-240.

[30] Wu Zhili, Li C. Feature Selection with Transductive Support Vector Machines [M]// Feature Extraction. Springer, Berlin, Heidelberg, 2006: 325-341.

[31] Guyon I, Weston J, Barnhill S, *et al.* Gene selection for cancer classification using support vector machines [J]. Machine learning, 2002, 46 (1-3): 389-422.

[32] Ang J C, Haron H, Hamed H N A. Semi-supervised SVM-based feature selection for cancer classification using microarray gene expression data [C]// International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2015: 468-477.

[33] Schölkopf B, Platt J C, Shawe-Taylor J, *et al.* Estimating the support of a high-dimensional distribution [J]. Neural computation, 2001, 13 (7): 1443-1471.

[34] Xu Zenglin, King I, Lyu M R T, *et al.* Discriminative semi-supervised feature selection via manifold regularization [J]. IEEE Transactions on Neural networks, 2010, 21 (7): 1033-1047.

[35] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples [J]. Journal of machine learning research, 2006, 7 (Nov): 2399-2434.

[36] Lanckriet G R G, Cristianini N, Bartlett P, *et al.* Learning the kernel matrix with semidefinite programming [J]. Journal of Machine learning research, 2004, 5 (Jan): 27-72.

[37] Xu Zenglin, Jin Rong, King I, *et al.* An extended level method for efficient multiple kernel learning [C]// Advances in neural information processing systems. 2009: 1825-1832.

[38] Yang Liming, Wang Laisheng. Simultaneous feature selection and classification via semi-supervised models [C]// Third International Conference on Natural Computation (ICNC 2007) . IEEE, 2007, 1: 646-650.

[39] Han Yahong, Yang Yi, Yan Yan, *et al.* Semisupervised feature selection via spline regression for video semantic recognition [J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 26 (2): 252-264.

[40] Yan Shuicheng, Xu Dong, Zhang Benyu, *et al.* Graph embedding and extensions: A general framework for dimensionality reduction [J]. IEEE

transactions on pattern analysis and machine intelligence, 2006, 29 (1): 40-51.

[41] Chen Xiaojun, Yuan Guowen, Nie Feiping, *et al.* Semi-supervised feature selection via sparse rescaled linear square regression [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 32 (1): 165-176.

[42] Yuan Guowen, Chen Xiaojun, Wang Chen, *et al.* Discriminative semi-supervised feature selection via rescaled least squares regression-supplement [C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[43] 严菲，王晓栋. 鲁棒的半监督多标签特征选择方法 [J]. 智能系统学报, 2019, 14 (4): 812-819. (Yan Fei, Wang Xiaodong. A robust, semi-supervised, and multi-label feature selection method [J]. CAAI transactions on intelligent systems, 2019, 14 (4): 812-819.)

[44] 吴锦华，万家山，伍祥, *et al.* 一种基于 Laplacian 的半监督特征选择模型 [J]. 重庆科技学院学报 (自然科学版) , 2019, 021 (001): 85-89. (Wu Jinhua, Wan Jiashan, Wu Xiang, *et al*. A Semi-Supervised Feature Selection Framework Based on Laplacian [J]. Journal of Chongqing University of Science and Technology (Natural Science) , 2019, 021 (001): 85-89.)

[45] Hu Qinghua, Yu Daren, Liu Jinfu, *et al.* Neighborhood rough set based heterogeneous feature subset selection [J]. Information sciences, 2008, 178 (18): 3577-3594.

[46] Dai Jianhua, Hu Qinghua, Zhang Jinghong, *et al.* Attribute selection for partially labeled categorical data by rough set approach [J]. IEEE transactions on cybernetics, 2016, 47 (9): 2460-2471.

[47] Liu Keyu, Yang Xibei, Yu Hualong, *et al.* Rough set based semi-supervised feature selection via ensemble selector [J]. Knowledge-Based Systems, 2019, 165: 282-296.

[48] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks [J]. Physical review E, 2007, 76 (3): 036106.

[49] Hu Qinghua, Yu Daren, Xie Zongxia. Neighborhood classifiers [J]. Expert systems with applications, 2008, 34 (2): 866-876.

[50] Wang Yintong, Wang Jiandong, Liao Hao, *et al.* An efficient semi-supervised representatives feature selection algorithm based on information theory [J]. Pattern Recognition, 2017, 61: 511-523.

[51] 王锋，刘吉超，魏巍. 基于信息熵的半监督特征选择算法 [J]. 计算机科学, 2018, 45 (S2): 437-440. (Wang Feng, Liu Jichao, Wei Wei. Semi-supervised feature selection algorithm based on information entropy [J]. Computer Science, 2018, 45 (S2): 437-440.)

[52] Sun Yijun, Todorovic S, Goodison S. A feature selection algorithm capable of handling extremely large data dimensionality [C]// Proceedings of the 2008 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2008: 530-540.

[53] Cheng Yubo, Cai Yunpeng, Sun Yijun, *et al.* Semi-supervised feature selection under logistic i-relief framework [C]// 2008 19th international conference on pattern recognition. IEEE, 2008: 1-4.

[54] Peng Hanchuan, Long Fuhui, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE Transactions on pattern analysis and machine intelligence, 2005, 27 (8): 1226-1238.

[55] Xu Jin, Tang Bo, He Haibo, *et al.* Semisupervised feature selection based on relevance and redundancy criteria [J]. IEEE transactions on neural networks and learning systems, 2016, 28 (9): 1974-1984.

[56] Wang Xiaodong, Chen R C, Yan Fei, *et al.* Semi-supervised adaptive feature analysis and its application for multimedia understanding [J]. Multimedia Tools and Applications, 2018, 77 (3): 3083-3104.

[57] Zeng Zhiqiang, Wang Xiaodong, Chen Yuming. Multimedia annotation via semi-supervised shared-subspace feature selection [J]. Journal of Visual Communication and Image Representation, 2017, 48: 386-395.

[58] Song Xiaonan, Zhang Jianguang, Han Yahong, *et al.* Semi-supervised feature selection via hierarchical regression for web image classification [J]. Multimedia Systems, 2016, 22 (1): 41-49.

[59] Shi Caijuan, An Gaoyun, Zhao Ruizhen, *et al.* Multiview Hessian semisupervised sparse feature selection for multimedia analysis [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 27 (9): 1947-1961.

[60] 史彩娟，阮秋琦，刘健, *et al.* 基于 Hessian 半监督特征选择的网络图像标注 [J]. 计算机应用研究, 2015 (2): 606-608. (Shi Caijuan, Ruan Qiuqi, Liu Jian, *et al.* Web image annotation based on Hessian semi-supervised feature selection [J]. Application Research of Computers, 2015 (2): 606-608.)

[61] Chen Xi, Song Lin, Hou Yuguan, *et al.* Efficient semi-supervised feature selection for VHR remote sensing images [C]// 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) . IEEE, 2016: 1500-1503.

[62] Chen Xi, Liu Wei, Su Fulin, *et al.* Semisupervised multiview feature selection for VHR remote sensing images with label learning and automatic view generation [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10 (6): 2876-2888.

[63] An L, Adeli E, Liu Mingxia, *et al.* Semi-supervised hierarchical multimodal feature and sample selection for Alzheimer's disease diagnosis [C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2016: 79-87.

[64] Jiang Xue, Chen Miao, Wang Weidi, *et al.* Label Propagation Based Semi-supervised Feature Selection to Decode Clinical Phenotype of Huntington's Disease [C]// International Conference on Intelligent Computing. Springer, Cham, 2019: 529-542.

[65] Yang Xukui, He Liang, Qu Dan, *et al.* Semi-supervised minimum redundancy maximum relevance feature selection for audio classification [J]. Multimedia Tools and Applications, 2018, 77 (1): 713-739.

[66] Sun Yaxin, Wen Guihua. Emotion recognition using semi-supervised feature selection with speaker normalization [J]. International Journal of Speech Technology, 2015, 18 (3): 317-331.

[67] Shi Caijuan, Gu Zhibin, Duan Changyu, *et al.* Multi-view adaptive semi-supervised feature selection with the self-paced learning [J]. Signal Processing, 2020, 168: 107332.

[68] Doquire G, Verleysen M. A graph Laplacian based approach to semi-supervised feature selection for regression problems [J]. Neurocomputing, 2013, 121: 5-13.