面向开放世界的半监督特征选择算法

王 锋 武文强 梁吉业
（山西大学计算机与信息技术学院 太原 030006）

摘 要 现有的半监督学习方法大多遵循封闭世界假设，即在模型训练过程中类别信息保持不变，标记数据可以覆盖所有类别。然而，在实际应用中，这一假设往往难以满足，未标记数据中通常会包含大量的未知类数据样本。为此，近年来研究人员提出了一个极具挑战性的研究方向：将半监督学习推广到不仅能够有效识别已知类的未标记数据样本，还能对未知的新类样本进行学习，从而构建面向开放世界的半监督学习机制。为应对这一挑战，本文基于符号型数据，提出了一种面向开放世界的半监督特征选择算法（OpenSSFS）。该算法将耦合学习引入到了符号型样本相似性度量以及类别关联性分析中，构建了新的样本相似性和类别相关性度量，并据此依次构建了三个核心模块：面向未标记已知类数据的自适应伪标签生成算法，面向未标记未知类数据的粒化和新类发现算法，以及基于类别相关性的特征选择算法。对给定的开放世界数据集，首先计算已知类数据样本的特征选择结果，并通过伪标签生成算法为未标记的已知类样本分配伪标签，进而基于所有已知类样本更新特征选择结果；其次，识别未知类未标记样本中的新类，并计算新类上的特征选择结果；最后，融合已知类样本和未知类样本的有效特征子集，确定最终的特征选择结果。为了有效验证所提新算法的有效性，本文在模拟的开放世界数据环境中进行了实验分析，分别测试了该算法在不同比例的已知类和未知类，以及不同比例的标记样本和未标记样本上的性能。实验结果表明，OpenSSFS算法在多种场景下均展现了较好的分类性能：首先，在包含50%已知类和50%未知类，且拥有50%标记样本的数据集上，新算法的分类精度最高提升了近70%，显著优于其他对比算法；其次，随着标记样本比例从90%降至10%，新算法的性能依然优于其他算法，且未出现明显下降，显示出较强的鲁棒性；最后，即使在已知类比例较低的情况下，OpenSSFS算法仍能保持良好的性能，适用于开放性更高的任务场景。此外，实验分析中还对算法中的参数阈值进行了详细分析和讨论。

关键词 半监督学习；开放世界学习；特征选择；耦合学习；成对相似性
中图法分类号 TP182 **DOI号** 10.11897/SP.J.1016.2025.01273

Semi-Supervised Feature Selection Algorithm for Open-World

WANG Feng WU Wen-Qiang LIANG Ji-Ye
()School of Computer and Information Technology, Shanxi University, Taiyuan 030006)

Abstract Existing semi-supervised learning methodologies typically operate under the closed-world assumption, wherein category information remains static throughout the learning process; that is, the labeled data utilized for model training encompasses all categories. However, this assumption frequently proves challenging to satisfy in practical applications. The unlabeled data often contain a substantial number of samples that belong to unknown classes. Consequently, researchers have identified a highly demanding research avenue in recent years: extending semi-supervised learning to enable not only the accurate identification of unlabeled data samples from known classes but also the discovery and learning of new, previously unknown classes, thereby establishing a semi-supervised learning framework for open-world scenarios. To tackle this

收稿日期：2024-12-11；在线发布日期：2025-04-01。本课题得到国家自然科学基金面上项目（62276158，62376141）资助。**王 锋**（通信作者），博士，副教授，中国计算机学会（CCF）会员，主要研究领域为数据挖掘、机器学习、粒计算。E-mail：sxuwangfeng@126.com。**武文强**，硕士研究生，中国计算机学会（CCF）会员，主要研究领域为动态数据挖掘与开放集学习。**梁吉业**，博士，教授，中国计算机学会（CCF）会士、IEEE Fellow，主要研究领域为粒计算、数据挖掘、机器学习。

challenge, this paper introduces a semi-supervised feature selection algorithm tailored for open-world scenarios based on categorical data (OpenSSFS). This algorithm integrates coupled learning into the similarity measurement of categorical samples and the relevance analysis of classes relationships, thereby establishing a novel similarity metric and a new class correlation metric. Based on these metrics, the new algorithm systematically constructs three core modules in sequence. The first one is an adaptive pseudo-label generation algorithm for unlabeled known-class data. The second one focuses on granulation and the discovery of novel classes within unlabeled data of unknown categories. And the final one presented is a feature selection algorithm based on classes relevance analysis. For a given open-world dataset, the first step involves computing the feature selection results for the known-class data samples, assigning pseudo-labels to the unlabeled known-class samples using the pseudo-label generation algorithm, and updating the feature selection results by incorporating all the known-class samples. In the second step, new classes within the unlabeled samples of the unknown class are identified, and the feature selection results based on new classes are computed. Finally, by integrating the effective feature subsets from both the known-class and unknown-class samples, the final feature selection outcome is determined. To further validate the effectiveness of the new algorithm proposed in this paper, an open world data environment is simulated in the experimental analysis. The new algorithm is tested and evaluated on the same dataset with varying proportions of known and unknown classes as well as different ratios of labeled and unlabeled samples. The experimental results indicate that the OpenSSFS algorithm has demonstrated excellent classification performance in various scenarios. Firstly, on a dataset comprising 50% known classes and 50% unknown classes, with 50% labeled samples, the new algorithm achieves a classification accuracy improvement of up to nearly 70%, demonstrating significantly superior performance compared to other contrast algorithms. Secondly, as the proportion of labeled samples is reduced from 90% to 10%, the performance of the new algorithm not only surpasses that of other algorithms but also maintains stability without any significant deterioration, thereby demonstrating its considerable robustness. Finally, the experimental findings regarding different proportions of known and unknown classes reveal that, even when there are only a few known classes, the new algorithm can still demonstrate excellent performance and handle more open task scenarios effectively. Furthermore, the experimental analysis carried out an analysis and discussion on the parameter threshold values set within the new algorithm.

Keywords semi-supervised learning; open-world learning; feature selection; coupled learning; pairwise similarity

1 引 言

半监督学习能够有效处理少量标注数据集的建模和学习问题，其目标是通过同时利用少量标记样本和大量未标记样本进行训练，使模型能够充分利用大量未标记样本，从而有效提升泛化性能[1-4]。现有的半监督学习机制大多假设接近封闭环境的学习任务，即在模型训练阶段所有可能的类别均为已知类。然而，随着数据获取工具的不断更新与发展，现实世界中的未标记数据中包含大量无法分属到任何已知类的未知类样本。这些数据样本通常会降低学习模型的泛化能力，甚至有可能被当作噪声数据。因此，随着超越封闭世界假设的学习任务日益增多，迫切需要探索新一代的半监督学习机制，以更有效地应对学习过程中类别信息的变化。对此，研究人员在近年来提出了一个极具挑战性的研究方向，即将半监督学习推广到不仅能够有效识别已知类的未标记数据样本，还能对未知的新类样本进行学习，从而构建面向开放世界的半监督学习机制。

面向开放世界的学习任务中承认未知类的存在，并尝试有效处理这些情况，赋予学习模型发现新事物和新知识的能力，从而提升其灵活性和鲁棒性。近年来，关于开放世界假设下各类学习模型的研究成果在国内外重要学术期刊和会议上均有报道，并呈现出上升趋势。现有的开放世界假设下的学习方法可以分为以下几类：新类检测、已知类分类与新类检测、面向无标记数据的新类发现以及零样本学习。其中，新类检测主要是确认测试样本是否属于已知类，其核心步骤在于度量已知类样本与未知类样本之间的相似性[5-8]。该类方法中一个常用的处理机制是将所有的已知类构建成一个类，将所有的未知类构建成一个类，从而学习任务抽象成一个二分类问题。优点是当已知类彼此之间的差异以及未知类彼此之间的差异可以被忽略时，上述方法能够有效应对新类检测的任务；然而，其缺点在于当类别间的差异较为明显时，该类处理机制的性能会明显受限。针对已知类分类与新类检测的相关探索中，一类有效的方法是开放集识别，其目标是对已知类别样本进行正确分类的同时，还要对出现的未知类别样本进行准确地判别[9]。文献[10]中首次阐述了开放集识别的相关概念和定义，并基于SVM构造了1-vs-set模型，以有效应对未知类别的识别问题。在此基础上，基于SVM以及其他多种机器学习模型的开放集识别方法均引起了研究人员的广泛关注[11-14]。此外，文献[15]还构建了OpenMax算法，通过对经典的分类模型进行扩展，实现了对未知类的有效识别。面向无标记数据的新类发现算法通常假设无标记样本仅来自未知类且未知类的个数是已知的。然而，在实际应用中，无标记数据既可能来源于已知类也可能来源于未知类，且未知类的个数通常无法预先确定[16-20]。在零样本学习中，整个类别空间是已知的，理论上不存在未知类，但是测试数据与训练数据的类别交集为空集[21-26]。综上所述，开放世界学习与上述四类算法存在一定的相似，但在训练过程中，不仅需要实现对已知类样本的识别，还要完成对新类的检测和发现。

为应对现实开放世界中新类的检测与发现，研究人员对传统封闭环境下的半监督学习机制进行了扩展，提出了开放世界半监督学习的概念[27-32]。文献[28]首次提出了开放世界半监督学习方法ORCA。该方法通过利用不确定性自适应边界来减小训练过程中已知类与新类之间类内方差的差距，从而有效避免了因过快学习已知类而导致对新类的忽视。文献[29]提出了广义类别发现方法GCD，其核心思想是利用对比表示学习和聚类来估计未标记数据样本的类别类标签以及类别数。OpenLDN引入了成对相似性损失来对未标记样本进行聚类，并为新类样本生成伪标签，从而将开放世界半监督学习问题转换为封闭世界半监督学习问题[30]。文献[31]中提出了一种TRSSL算法，其核心思想是结合样本的不确定性以及类别分布的先验知识为未标记样本生成可靠的伪标签。然而，该算法中的不确定性估计过程比较耗时。此外，文献[32]利用Sinkhorn-Knopp算法，克服了已知类的伪标签过度自信问题，并据此提出了一个可以识别所有样本（包含标记样本和未标记样本）并能区分所有类别（包含已知类和未知类）的统一框架。

特征选择作为一种常用的数据预处理技术，是众多机器学习模型中的一个关键步骤，其主要目标是去除冗余和无关特征，提高模型的学习效率，并有效避免过拟合[33-40]。为有效应对部分标记数据集的特征选择问题，许多研究者将半监督学习引入特征选择过程中，构建了半监督特征选择机制[41-43]。然而，现有的半监督特征选择机制同样仅适用于封闭环境下的数据降维问题，无法有效处理测试数据中包含新类或者未知类的问题，即在现实开放环境下进行特征选择时存在局限性。为此，基于符号型数据，本文构建了一种面向现实开放世界的半监督特征选择机制。该方法的核心思想包括以下几个方面：

（1）引入了一种基于耦合学习的未标记数据样本相似性度量，并将其应用于求解未标记样本的成对相似性样本以及聚类分析中；

（2）在对未标记样本的识别中，引入了一种两级判别指示函数，通过分析未标记样本的成对相似样本与标记样本的相似度来判别未标记样本是否属于已知类数据；

（3）发展了面向未标记未知类数据样本的粒化方法和新类发现算法；

（4）通过综合分析类内凝聚性和类间关联性，设计了一种类别相关性的特征选择算法，并将其应用于开放世界半监督特征选择问题的求解中。

2　基本概念

为了有效处理开放世界中数据集的有效特征子集选取问题，本节引入如下的数据表示方法及其相

关概念。

在开放世界半监督学习中，数据集$\mathcal{D}$由标记数据样本集$\mathcal{D}_l$和未标记样本集$\mathcal{D}_u$组成，即$\mathcal{D}=\mathcal{D}_l\cup\mathcal{D}_u$。标记样本集$\mathcal{D}_l=\{(x_i^l,y_i)\}_{i=1}^m$，相应的标记类别(已知类)集合为$\mathcal{Y}_l=\{\mathcal{Y}_1,\mathcal{Y}_2,\cdots,\mathcal{Y}_{known}\}$，其中标记样本集的类标签$y_l$来自标记类别集合$\mathcal{Y}_l$，$y_l\in\mathcal{Y}_l$。未标记样本集$\mathcal{D}_u=\{(x_i^u)\}_{i=1}^n$，并假设未标记样本对应的类别集合为$\mathcal{Y}_u=\{\mathcal{Y}_1,\mathcal{Y}_2,\cdots,\mathcal{Y}_{known},\cdots,\mathcal{Y}_{unknown}\}$，其中未标记样本的类标签$y_u$来自未标记类别集合$\mathcal{Y}_u$，$y_u\in\mathcal{Y}_u$，需要指出的是在学习模型训练过程中并不使用未标记样本的类标签。在开放世界中，未标记类别集合$\mathcal{Y}_u$通常会包含已知类和一些新类，$\mathcal{Y}_l\cap\mathcal{Y}_u\neq\varnothing$且$\mathcal{Y}_l\neq\mathcal{Y}_u$，即$\mathcal{Y}_l\subset\mathcal{Y}_u$。因此，新类(未知类)的集合表示为$\mathcal{Y}_{novel}=\mathcal{Y}_u\backslash\mathcal{Y}_l$。综上，开放世界学习中包含了新类的发现以及传统(封闭世界)的半监督学习。

开放世界中，$\mathcal{D}=\{x_i\}_{i=1}^{m+n}$为数据样本集，$\mathbb{R}^{dim}$表示样本特征维度，$V_a$表示特征$a\in\mathbb{R}^{dim}$的值域，并且有$V=\bigcup_{a\in\mathbb{R}^{dim}}V_a$；对于任意的$a\in\mathbb{R}^{dim}$和样本$x\in\mathcal{D}$，$f:\mathcal{D}\times\mathbb{R}^{dim}\to V$是一个信息函数，并且有$f(x,a)\in V_a$。基于特征子集$B\subseteq\mathbb{R}^{dim}$可以诱导一个等价关系$R_B=\{(x,y)\in\mathcal{D}\times\mathcal{D}|f(x,a)=f(y,a),\forall a\in B\}$。根据$R_B$可将样本集$\mathcal{D}$划分为一组等价类，表示为$U/R_B=\{[x]_B|x\in\mathcal{D}\}$，其中$[x]_B=\{y\in\mathcal{D}|(x,y)\in R_B\}$。基于上述相关概念，对任意样本子集$X\subseteq\mathcal{D}$，其基于特征子集$B$上下近似算子分别为：$\overline{B}(X)=\{x\in\mathcal{D}|[x]_B\cap X\neq\varnothing\}$和$\underline{B}(X)=\{x\in\mathcal{D}|[x]_B\subseteq X\}$。

定义1. 设给定样本集为$\mathcal{D}=\mathcal{D}_l\cup\mathcal{D}_u$，其对应的特征集和类信息分别为$\mathbb{R}^{dim}$和$\mathcal{Y}$，则可得$\mathcal{D}/B=\{X_1,X_2,\cdots,X_m\}$和$\mathcal{D}/\mathcal{Y}=\{Y_1,Y_2,\cdots,Y_n\}$。在此基础上，特征子集$B\subseteq\mathbb{R}^{dim}$的互补信息熵定义为

$$E(B)=\sum_{i=1}^{m}\frac{|X_i|}{|\mathcal{D}|}\left(1-\frac{|X_i|}{|\mathcal{D}|}\right)\tag{1}$$

B相对于$\mathcal{Y}$的条件互补信息熵定义为

$$E(\mathcal{Y}|B)=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|\mathcal{D}|}\cdot\frac{|X_i-Y_j|}{|\mathcal{D}|}\tag{2}$$

3 基于耦合学习的相似性度量

为有效分析符号型数据样本的相似度，文献[40]中提出了一种基于互补信息熵的耦合信息分析策略，并据此定义了一种基于耦合学习的数据样本相似性度量。耦合学习主要关注于挖掘数据本身所包含的耦合信息，即数据的取值、特征、样本以及类别之间都存在一定的相互作用和影响[40]。尤其在符号型数据中，对这种耦合性的挖掘和探索对于揭示隐藏在数据样本中的复杂信息具有至关重要的意义。为度量开放世界中符号型数据样本的相似性，本文针对标记类数据样本的相似性度量，采用了文献[40]提出的相似性度量来求解；针对未标记类数据样本，本节构建了基于耦合学习的无监督符号型数据相似性度量。该度量综合考虑了符号型数据取值的出现频率和分布比例，并据此分别讨论了数据取值的内部相似度和外部相似度，详见定义2-5。

定义2. 令$\mathcal{D}_u$是一个无标记数据样本集，其特征集为$\mathbb{R}^{dim}$，且$A_k\subseteq\mathbb{R}^{dim}$，则$\forall x_1,x_2\in\mathcal{D}$和$a_\xi\in A_k$，$x_1$和$x_2$在特征$a_\xi$上的内部相似度定义为

$$S_{unseen}^{in}(\xi,x_1,x_2)=\begin{cases}\dfrac{\left|\mathcal{D}-[x_1]_{a_\xi}\right|}{|\mathcal{D}|}, & [x_1]_{a_\xi}=[x_2]_{a_\xi}\\ \kappa^2\cdot\dfrac{\left|\mathcal{D}-[x_1]_{a_\xi}\right|\cdot\left|\mathcal{D}-[x_2]_{a_\xi}\right|}{|\mathcal{D}|^2}, & [x_1]_{a_\xi}\neq[x_2]_{a_\xi}\end{cases}\tag{3}$$

其中$\kappa=\dfrac{\min(\left|[x_1]_{a_\xi}\right|,\left|[x_2]_{a_\xi}\right|)}{\max(\left|[x_1]_{a_\xi}\right|,\left|[x_2]_{a_\xi}\right|)}$。

针对未标记类样本，定义2中主要给出了两个样本在同一个特征a_ξ上的内部相似度。该度量综合考虑了不同取值出现频率的接近程度以及数据取值在特征a_ξ的所有取值中的分布比例。在此基础上，定义3中讨论了样本的外部相似度，即a_ξ以外的特征$a_\gamma(\gamma\neq\xi)$对样本在特征a_ξ上的相似性的贡献。本文借助粗糙集理论中的上近似算子概念，该算子表示包含目标概念的等价类。这一概念与本节中分析的不同特征取值之间的耦合信息高度契合，因此引入基于上近似算子的样本外部相似度度量。定义3中主要综合考虑了不同取值对应的上近似算子的接近程度及其在所有取值中的分布比例，具体定义如下。

定义3. 令$\mathcal{D}_u$是一个无标记数据样本集，其特征集为$\mathbb{R}^{dim}$，且$A_k\subseteq\mathbb{R}^{dim}$，则$\forall x_1,x_2\in\mathcal{D}$和$a_\xi\in A_k$，$x_1$和$x_2$在特征$a_\gamma\in A_k(\xi\neq\gamma)$上的外部相似度定义为

$$S_{unseen}^{out}(\xi,x_1,x_2)=\begin{cases}\dfrac{\left|\mathcal{D}-\bar{a}_\gamma([x_1]_{a_\xi})\right|}{|\mathcal{D}|}, & \bar{a}_\gamma([x_1]_{a_\xi})=\bar{a}_\gamma([x_2]_{a_\xi})\\ \sigma^2\cdot\dfrac{\left|\mathcal{D}-\bar{a}_\gamma([x_1]_{a_\xi})\right|\cdot\left|\mathcal{D}-\bar{a}_\gamma([x_2]_{a_\xi})\right|}{|\mathcal{D}|^2}, & \bar{a}_\gamma([x_1]_{a_\xi})\neq\bar{a}_\gamma([x_2]_{a_\xi})\end{cases}\tag{4}$$

其中$\sigma=\dfrac{\min(\left|\bar{a}_\gamma([x_1]_{a_\xi})\right|,\left|\bar{a}_\gamma([x_2]_{a_\xi})\right|)}{\max(\left|\bar{a}_\gamma([x_1]_{a_\xi})\right|,\left|\bar{a}_\gamma([x_2]_{a_\xi})\right|)}$。

基于定义2-3中的定义，通过组合不同样本在特征a_ξ上的内部相似度和外部相似度，定义4-5中分别给出了两个未标记样本在单个特征和特征子集上的相似性度量，具体定义如下。

定义4.　令$\mathcal{D}_u$是一个无标记数据样本集，其特征集为$\mathbb{R}^{dim}$，且$A_k\subseteq\mathbb{R}^{dim}$，则$\forall x_1,x_2\in\mathcal{D}$和$a_\xi,a_\gamma\in A_k$，$x_1$和$x_2$在特征$a_\xi$上的相似度定义为

$$S'_{unseen}(\xi,x_1,x_2)=\alpha S_{unseen}^{in}(\xi,x_1,x_2)+\beta\frac{\sum_{\gamma=1,\gamma\neq\xi}^{|A_k|}S_{unseen}^{out}(\xi,x_1,x_2)}{|A_k|-1}\tag{5}$$

其中$\alpha+\beta=1,\alpha=\beta=0.5$。

定义5.　令$\mathcal{D}_u$是一个无标记数据样本集，其特征集为$\mathbb{R}^{dim}$，且$A_k\subseteq\mathbb{R}^{dim}$，则$\forall x_1,x_2\in\mathcal{D}$和$a_\xi,a_\gamma\in A_k$，$x_1$和$x_2$在特征子集$A_k$上的相似度定义为

$$S_{unseen}(x_1,x_2)=\frac{1}{|A_k|}\sum_{a_\xi\in A_k}S'_{unseen}(\xi,x_1,x_2)\tag{6}$$

4　面向开放世界的伪标签生成算法

针对开放世界的半监督学习，未标记样本通常同时包含已知类样本和未知类样本。为此，本节分别给出了针对未标记已知类样本和未标记未知类样本的标签生成算法。在4.1节中，基于样本成对相似性，引入了一种面向开放世界中未标记已知类样本的自适应伪标签生成算法；在4.2节中，基于聚类分析，提出了一种面向开放世界中未标记未知类样本的新类发现算法。

4.1　面向未标记已知类样本的自适应伪标签生成算法

针对开放世界的半监督学习，现有研究大多从成对相似性样本的角度展开探索，主要采用余弦相似度来度量样本间的相似度，并通过设定相似度阈值来判别和生成伪标签。然而，伪标签的生成质量会明显受到设定阈值的影响，从而影响模型的整体性能；另外，余弦相似度并不适合处理符号型数据。为此，本节引入了一种新的面向开放世界的自适应伪标签生成算法，可有效识别未标记样本中的已知类样本，并为其生成相应类别的伪标签。该算法的核心思想是引入了一种新的相似度判别机制，以替代传统的阈值判别方式。具体而言，通过分析成对相似未标记样本和标记样本的相似度比值来自适应生成相应的伪标签，从而为该类数据生成高质量的伪标签。

该算法具体包含以下两个核心步骤：首先，分别构建了面向未标记样本的相似度矩阵和面向已知类样本的类中心，并据此依次求解了未标记样本的成对相似性样本、成对相似性样本与标记样本的相似度矩阵以及标记样本与对应类心的相似度值；其次，构建了两级判别指示函数，通过分析未标记样本的成对相似样本与标记样本的相似度来判别未标记样本是否属于已知类别，并为其生成相应类别的伪标签。上述步骤中的两级判别策略的具体内容是：一级判别指示函数主要通过分析未标记样本与其成对相似样本的相似度$\mathbb{D}$和成对相似样本与标记数据的相似度d，初步判别未标记样本是否属于已知类数据，如果未标记样本与成对相似样本相似度小于成对相似样本与标记数据的最小相似度或者大于最大相似度，说明该未标记样本是未知类数据，反之如果未标记样本与成对相似样本相似度在成对相似样本与标记数据相似度区间内，则选取成对相似样本与标记数据相似度大于未标记样本与成对相似样本相似度对应的标记样本集Θ；在此基础上，二级判别指示函数基于标记样本集Θ来确认未标记样本的伪标签，具体策略是通过计算成对相似样本与Θ中标记样本的相似度d和Θ中标记样本与其对应类心的相似度d^*的比值；将比值大于1的样本按照比值降序排列，然后选择比值最大的标记样本所对应的标签作为未标记样本的伪标签。为更清晰地展示上述两级判别机制，图1中给出了判别过程示意图。基于上述分析，算法1中详细介绍了基于成对相似性的伪标签生成算法。由于在算法中引入了类心的计算，定义6中通过借鉴面向符号型数据的聚类分析中类簇中心的定义[44]，引入一种符号数据类心的定义。该定义是将一类中所有样本各特征的众数设定为该类的中心点对应特征的取值，其中众数是指每个特征在该类中的所有取值中出现频率最高的值。

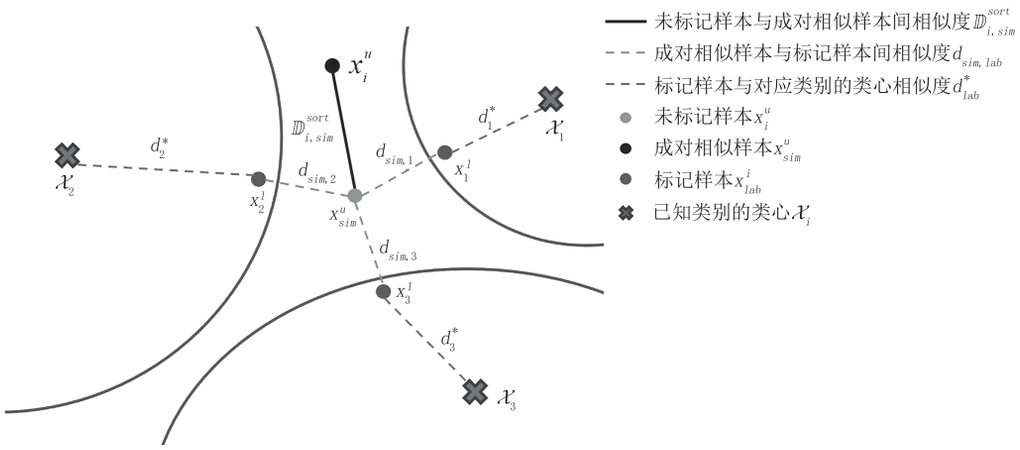

图1 自适应伪标签生成判别机制

定义6. 令$\mathcal{D}_l$是一个有标记数据样本集，其特征集和类信息分别为$\mathbb{R}^{dim}$和$\mathcal{Y}$，且有$\mathcal{D}_l/\mathcal{Y}=\{Y_1, Y_2, \cdots, Y_L\}$和$A_k \subseteq \mathbb{R}^{dim}$，类别$Y_i$的类心定义为

$$\mathcal{X}_i=\{\mathrm{mode}(a_j), a_j \in A_k\} \tag{7}$$

其中，$\mathrm{mode}(a_j)$=表示特征a_j在Y_i上的众数，即特征a_j在样本集Y_i上的所有取值中出现频率最高的值。

算法1. 基于成对相似性的自适应伪标签生成算法.

输入：数据集$\mathcal{D}=\mathcal{D}_l \cup \mathcal{D}_u$

输出：标记数据集$\mathcal{D}_l'=\mathcal{D}_l \cup \mathcal{D}_{PL}$

1. 根据定义6，求解标记数据$\mathcal{D}_l$中每一个类别的类心$\mathcal{X}_i$，其中$i=1,2,\cdots,label$；

2. 令$n \leftarrow \mathcal{D}_u$，求解未标记数据相似度矩阵$\mathbb{D}$：$\mathbb{D}_{ij}=S_{unseen}(x_i^u, x_j^u)$，其中$i,j=1,2,\cdots,n$；

3. 基于相似度矩阵$\mathbb{D}$，$i=1,2,\cdots,n$：

3.1 求解x_i^u的成对相似样本x_{sim}^u，即与x_i^u相似度最大的无标记样本，其中$\mathbb{D}_{i,sim}^{sort}$表示$x_i^u$与$x_{sim}^u$的相似度值；

3.2 分别计算x_{sim}^u与每个标记样本x_{lab}^l的相似度$d_{sim,lab}=S_{unseen}(x_{sim}^u, x_{lab}^l)$，以及每个标记样本与其对应类心的相似度$d_{lab}^*=S_{seen}(\mathcal{X}_{lab}, x_{lab}^l)$，其中$lab=1,2,\cdots,m$；

4. $\Theta \leftarrow \phi$，对于x_i^u，一级判别$\mathbb{I}(\mathbb{D}_{i,sim}^{sort} > d_{sim,lab})$：如果$x_i^u$对应的$\mathbb{D}_{i.sim}^{sort}$大于$d_{sim,lab}$，则$\Theta \leftarrow \Theta \cup x_i^u$；否则$x_i^u$为未知类数据；

5. 基于样本集Θ，二级判别$\mathbb{I}(\partial > 1 \,\&\, \arg\max\{\partial\})$，其中$\partial=d_{sim,lab}/d_{lab}^*$：如果$\partial>1$，则选择$\partial$值最大样本的标签生成$x_i^u$的伪标签；否则$x_i^u$是未知类数据；

6. 生成伪标签样本集$\mathcal{D}_{PL}=(x_i^u, y_i)$，其中$i=1,2,\cdots,pl, y_{pl} \in \mathcal{Y}_l$.；

7. 返回有标记样本集$\mathcal{D}_l'=\mathcal{D}_l \cup \mathcal{D}_{PL}$.

算法1中，步骤1中根据定义6求解了标记数据在已知类别下关于已知类数据特征子集的类心集合$\mathcal{X}=\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{label}\}$；步骤2构建了未标记样本$\{x_i^u\}_{i=1}^n$相似度矩阵$\mathbb{D}$；步骤3.1查找$x_i^u$的成对相似未标记样本$x_{sim}^u$；步骤3.2计算成对相似样本$x_{sim}^u$与所有标记数据$\{x_{lab}^l\}_{lab=1}^m$的相似度，以及标记数据$\{x_{lab}^l\}_{lab=1}^m$与对应类别的类心$\{\mathcal{X}_i\}_{i=1}^{label}$的相似度；步骤4-5实现了针对已知类未标记样本的两级判别，为未标记样本生成伪标签。

算法1的时间复杂度：步骤1的时间复杂度为$O(m)$。根据定义5计算步骤2的时间复杂度为$O(n^2\mathbb{R}^{dim})$。步骤3分为两步，步骤3.1的时间复杂度为$O(n\log n)+O(n)=O(n\log n)$，步骤3.2的时间复杂度为$O(nm\mathbb{R}^{dim})+O(m)=O(nm\mathbb{R}^{dim})$，步骤3的总体时间复杂度为$O(nm\mathbb{R}^{dim})$。步骤4的时间复杂度为$O(m\log m)+O(m^2)=O(m^2)$。步骤5中，假设$\Theta$的大小为$k$，该步骤的时间复杂度为$O(k)+O(k\log k)=O(k\log k)$。算法1的总时间复杂度为$\max(O(n^2\mathbb{R}^{dim}), O(nm\mathbb{R}^{dim}))=O(n^2\mathbb{R}^{dim})$。

4.2 面向未标记未知类样本的新类发现算法

在4.1节的基础上，本节进一步讨论面向未标记未知类样本的新类发现方法。针对未标记样本中的未知类样本，本节引入了经典的面向符号型数据的K-Modes算法对其进行聚类，以获取未标记数据样本中的新类。由于未知类样本的类别数无法预先确定，本节中引入了聚类分析中常用的评价指标误差平方和(SSE)，以评估K-Modes算法中的K值，并据此选择最优的K值及其相应的聚类结果。最后，将求解得到的类簇结果确定为未标记样本的新类。另外，针对K-Modes算法中样本相似度的计算，本节中仍然使用本文定义2-5中的基于耦合学习的相似性度量，具体步骤详见算法2。

算法2. 基于耦合学习的K-Modes聚类算法.

输入：未知类数据集$\mathcal{D}_u'$，聚类数K

输出：类簇$U_{cluster}=\{U_1, U_2, \cdots, U_K\}$

1. 从$\mathcal{D}'_u$中随机抽取K个样本作为初始聚类中心$\mathcal{X}=\{\mathcal{X}_1,\mathcal{X}_2,\cdots,\mathcal{X}_K\}$；

2. 基于定义6计算$\mathcal{D}'_u$中每个样本与所有类心的距离，并将每个样本分配到离它最近的类簇中，形成初始类簇$\{U_1,U_2,\cdots,U_K\}$；

3. 基于定义6重新求解每个类的类中心，更新$\mathcal{X}=\{\mathcal{X}_1,\mathcal{X}_2,\cdots,\mathcal{X}_K\}$；

4. 重复步骤2-3中过程，直到总距离（各类簇中的样本与其类心的距离总和）不再降低，返回最终的聚类结果$U_{cluster}=\{U_1,U_2,\cdots,U_K\}$。

对于给定的未知类样本集和不同的K值（K≥2），使用算法2可获得多组聚类结果。在此基础上，结合评估指标SSE，确定最优的K值及其对应的聚类结果，具体步骤详见算法3.

算法3. 基于K-Modes和SSE的未知类样本新类发现算法.

输入：未知类数据集$\mathcal{D}'_u$

输出：k个新类$\{U_1,U_2,\cdots,U_k\}$

1. FOR $k=2$; $k\leqslant|\mathcal{D}'_u|$; $k++$DO

聚类参数为k，采用算法2求解样本集$\mathcal{D}'_u$的聚类结果，并求解聚类结果对应的SSE值；

END FOR

2. 基于步骤1中的多个k值对应的多个SSE值，使用“肘部法”绘制SSE随k值变化的曲线，找到“肘部”作为最佳的k值，进而确定最佳k值，$k\leftarrow k$；

3. 将步骤2中选择到的最佳k值及其对应的聚类结果，作为最终的新类确定结果，并返回。

5 面向开放世界的半监督特征选择算法

基于第4节中介绍的面向已知类未标记样本的伪标签生成方法以及面向未知类未标记样本的新类发现算法，本节介绍面向开放世界的半监督特征选择算法。

为求解标记数据集的有效特征子集，本节首先提出了一种基于类别相关性的特征选择算法。针对符号型数据，文献[40]通过分析数据取值间的耦合信息，提出了一种基于耦合学习的特征选择算法。然而，该算法只适用于封闭集的特征选择，且未对数据类别间的耦合信息作分析和探索。为此，本节中通过分析数据类别间的耦合性，提出了一种基于类别相关性的特征选择算法。其中，类别间的耦合信息主要综合考虑了类内凝聚性和类间关联性两个方面，定义7和定义8分别阐述了类内凝聚性和类间关联性的定义，定义9则提出了一种新的类别相关性的度量，具体定义如下。

定义7. 令$T=(\mathcal{D},A_k\cup\mathcal{Y})$是一个数据表，且$\mathcal{D}/\mathcal{Y}=\{Y_1,Y_2,\cdots,Y_L\}$，则第$i$类的类内凝聚性$\varepsilon^i_{in}$定义为

$$\varepsilon^i_{in}=\frac{1}{|Y_i|}\sum_{j=1}^{|Y_i|}S_{seen}(\mathcal{X}_i,x_j)\tag{8}$$

其中，$x_j\in Y_i$且$\mathcal{X}_i$为第i类的类心。

定义8. 令$T=(\mathcal{D},A_k\cup\mathcal{Y})$是一个数据表，且$\mathcal{D}/\mathcal{Y}=\{Y_1,Y_2,\cdots,Y_L\}$，则$\forall Y_i,Y_j\in\mathcal{D}/\mathcal{Y}(j\neq i)$，其类间关联性$\varepsilon^i_{out}$定义为

$$\varepsilon^i_{out}(Y_i,Y_j)=\max\{\sup_{x_i\in Y_i}\inf_{x_j\in Y_j}S_{seen}(x_i,x_j),\sup_{x_j\in Y_j}\inf_{x_i\in Y_i}S_{seen}(x_j,x_i)\}\tag{9}$$

其中，$\sup(\cdot)$表示上确界，即最大值；$\inf(\cdot)$表示下确界，即最小值。

定义9. 令$T=(\mathcal{D},A_k\cup\mathcal{Y})$是一个数据表，且$\mathcal{D}/\mathcal{Y}=\{Y_1,Y_2,\cdots,Y_L\}$，则类别相关性$\varepsilon$定义为

$$\varepsilon=\sum_{i=1}^{L}\frac{\omega_i\varepsilon^i_{in}}{\sum_{j=1,j\neq i}^{L}\omega_j\varepsilon^i_{out}(Y_i,Y_j)}\tag{10}$$

其中，$\omega_i=|Y_i|/|\mathcal{D}|$且$\omega_j=|Y_j|/|\mathcal{D}|$。

基于上述类别相关性的定义，算法4中详细描述了基于类别耦合性的特征选择算法步骤。

算法4. 基于耦合性的特征选择算法.

输入：数据集$\mathcal{D}$，类别相关性集合$E\leftarrow\varnothing$

输出：目标特征子集$\mathcal{A}_{select}$

1. 根据基于耦合学习的特征权重求解算法[40]计算特征权重$w(a_i)(i=1,2,\cdots,|\mathbb{R}^{dim}|)$；

2. 根据特征权重进行降序排序，记$B=\{a'_1,a'_2,\cdots,a'_{|\mathbb{R}^{dim}|}\}$；

3. FOR $i=1$; $i\leqslant|\mathbb{R}^{dim}|$; $i++$DO

根据特征权重选取前i个特征，记A_k；

根据定义9计算类别相关性ε_i；

$E\leftarrow E\cup\varepsilon_i$；

END FOR

4. FOR $j=1$; $j\leqslant|E|$; $j++$DO

计算$M(j)=\varepsilon_{j+1}-\varepsilon_j$；

IF $M(j)<\sigma$ THEN

保留类别相关性ε_j，根据下标选取前j个特征，记$\mathcal{A}_{select}$；

END IF

END FOR

5. 返回目标特征子集$\mathcal{A}_{select}$；

基于上述分析，针对包含已知类未标记样本和未知类未标记样本的数据集，下面将详细阐述面向开放世界的半监督特征选择算法。该算法主要包含

以下几个核心步骤：首先，利用算法4根据原始标记数据样本计算初始已知类数据的有效特征子集；在此基础上，结合算法1中为已知类未标记样本生成的伪标签，在所有已知类样本上更新特征选择结果。其次，基于算法2和算法3确定的未知类未标记样本中的新类，使用算法4求解未知类未标记样本集的有效特征子集。最后，通过融合已知类样本和未知类样本的有效特征子集，确定最终的特征选择结果，详细的算法步骤见算法5。

算法5. 面向开放世界的半监督特征选择算法(OpenSSFS).

输入：数据集 $\mathcal{D}=\mathcal{D}_l\cup\mathcal{D}_u$

输出：特征子集 $\mathcal{A}=\mathcal{A}'_{label}\cup\mathcal{A}_{unlabel}$

1. 使用算法4在标记数据 $\mathcal{D}_l$ 上计算初始已知类数据特征子集 $\mathcal{A}_{label}$；

2. 使用算法1为未标记数据中已知类数据并生成相应伪标签，即 $\mathcal{D}_{PL}=\{(x_i^u,y_i)\}_{i=1}^{pl}, y_{pl}\in\mathcal{Y}_l$；将伪标签数据合并到原始的标记数据中，$\mathcal{D}'_l=\mathcal{D}_l\cup\mathcal{D}_{PL}$；剩余的未标记未知类数据表示为 $\mathcal{D}'_u$；

3. 使用算法4求解 $\mathcal{D}'_l$ 上的特征子集 $\mathcal{A}'_{label}$；

4. 对未知类数据 $\mathcal{D}'_u$ 使用算法2和算法3进行聚类，形成类簇 $U_{cluster}=\{U_1,U_2,\cdots,U_K\}$，并为每个类生成伪标签 $T=\{(\mathcal{D}'_u,\mathbb{R}^{dim}\cup U_{cluster})\}$；

5. 对 $T=\{(\mathcal{D}'_u,\mathbb{R}^{dim}\cup U_{cluster})\}$，使用算法4求解特征子集 $\mathcal{A}_{unlabel}$；

6. 融合已知类数据特征子集 $\mathcal{A}'_{label}$ 和未知类数据特征子集 $\mathcal{A}_{unlabel}$，返回最终的特征选择结果 $\mathcal{A}=\mathcal{A}'_{label}\cup\mathcal{A}_{unlabel}$。

在算法5中，步骤1-3是基于原始标记数据计算初始已知类数据的有效特征子集，并在此基础上，结合算法1生成的伪标签，在所有已知类样本上更新特征选择结果；步骤4-5是对未知类数据的特征选择；步骤6则融合已知类和未知类数据的有效特征子集，确定最终的有效特征子集。

算法5的时间复杂度：根据上述分析，步骤1的时间复杂度为 $\max(O(|\mathbb{R}^{dim}|^2), O(Nm\mathbb{R}^{dim}))$。步骤2的时间复杂度为 $O(n^2\mathbb{R}^{dim})$。步骤3的时间复杂度为 $\max(O(|\mathbb{R}^{dim}|^2), O(N(m+pl)\mathbb{R}^{dim}))$。在步骤4-5中，假设 $iter$ 是迭代次数，$k_{\max}$ 是最大 k 值，该步骤的时间复杂度 $O(k_{\max}(n-pl)|\mathbb{R}^{dim}|iter)+O((n-pl)^2\mathbb{R}^{dim})$。算法5的总时间复杂度为 $O(n^2\mathbb{R}^{dim})$。

6 实验分析

为验证本文提出的OpenSSFS算法的有效性，本节在多个公开数据集上进行了仿真实验，以评价和分析该算法在面向开放世界数据集中的性能。实验分析中使用的数据集是在UCI数据库中选取的12个符号型数据集，详细信息见表1。该表第二列中符号的意义分别是：S表示规模较小或者维度较低的数据集，D表示维度较高的数据集，L表示规模较大的数据集，C表示类别数比较多的数据集。为模拟开放世界数据集环境，实验设计中首先将所有数据集的70%作为训练集，30%作为测试集。在训练集中，将类别分为50%的已知类和50%的未知类。在此基础上，在已知类数据中按比例抽取一部分作为标记数据，剩余数据均为未标记数据。本文仿真实验的测试环境是Intel(R) Core(TM) i7-10700 CPU @ 2.90 GHz，内存16.0 GB，算法编程语言为Python，使用的开发工具是JetBrains PyCharm Professional Edition2023.3.5。另外，对表1中存在缺失值的符号型数据集，本实验中对其缺失值进行了填补，填补策略是删除缺失值数量大于样本三分之一的特征，对于缺失值小于样本数量三分之一的特征，用该列众数填充。在此基础上，具体的实验设计及结果将在6.1～6.5节中进行详细介绍。

表1 实验数据集

No	T	Datasets	Instances	Features	Class
1	S	Soybean-small	47	35	4
2	S	Zoo	101	16	7
3	S	Dermatology	366	34	6
4	D	Lung cancer	33	56	3
5	D	MPE	1080	80	8
6	D	MIC	1700	111	8
7	D	Lymphoma	96	4026	9
8	D	Nci9	60	9712	9
9	S/C	Soybean-large	307	35	19
10	L/C	Letter	20 000	16	26
11	L/C	Krkopt	28 056	6	18
12	L	19sat.trn	4420	35	6

6.1 特征选择性能比较

为有效验证OpenSSFS算法的性能，本节选择了四种特征选择算法作为对比算法，具体包括：Relief-F算法[45]、基于信息增益的特征选择算法(FSIG)[46]、基于互信息的特征选择算法(FSMI)[47-48]和FSCL算法[40]。其中Relief-F、FSIG和FSMI是三种经典的特征选择算法；而FSCL是一种基于耦合学习的特征选择算法。由于本文提出的OpenSSFS算法在求解有效特征子集的过程中采

用了基于类别耦合性的特征选择算法(算法4)，因此在本节的实验分析中，选择了一种基于耦合学习的特征选择算法(FSCL)作为对比算法。为了评估特征选择结果的分类性能，本节引入了四种常用的机器学习分类器来对上述四种算法得到的特征选择结果进行评价，分别是支持向量机(SVM)、朴素贝叶斯(NBC)、随机森林(RandomForest)和决策树(C4.5)。在此基础上，将每个数据集的类别分为50%的已知类和50%的未知类，详细的实验比较结果见表2～表9，其中表2～表5是基于在已知类数据样本中抽取50%的样本为标记样本集的实验结果；表6～表9的实验结果则是基于在已知类数据样本中抽取10%的样本为标记样本集的实验结果。表2～表9中Known表示在已知类数据样本上的分类精度，Novel表示在未知类数据样本上的分类精度，All表示在所有数据样本上的分类精度。

表2　基于SVM的分类精度比较结果(50%标记率)

No	OpenSSFS			FSCL			Relief-F			FSIG			FSMI		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
1	100	100	**100**	85.00	75.00	80.00	85.00	75.00	80.00	100.00	60.00	80.00	86.67	33.33	60.00
2	100	81.66	**90.83**	84.19	77.63	80.91	89.43	60.57	75.00	95.33	86.33	90.83	95.95	85.71	90.83
3	100	90.36	**95.18**	100	76.34	88.17	85.02	81.82	83.42	98.59	61.41	80.00	81.27	87.83	84.55
4	64.44	68.90	**66.67**	52.76	58.36	55.56	32.07	45.71	38.89	33.33	44.45	38.89	38.07	39.71	38.89
5	100	92.74	**96.37**	95.41	89.77	92.59	93.04	90.28	91.66	79.27	80.51	79.89	84.19	76.33	80.26
6	53.23	46.01	**49.62**	43.96	43.08	43.52	58.18	28.34	43.26	45.36	35.48	40.42	52.38	28.36	40.42
7	64.23	70.67	**67.45**	46.67	73.33	60.00	45.00	75.00	60.00	46.27	70.33	58.30	33.33	86.67	60.00
8	34.12	10.32	**22.22**	17.07	16.27	16.67	17.79	15.55	16.67	14.25	7.97	11.11	17.11	5.11	11.11
9	69.24	58.88	**64.06**	63.62	60.06	61.84	65.15	54.01	59.58	67.51	51.11	59.31	56.11	53.61	54.86
10	88.64	87.62	**88.13**	89.90	82.84	86.37	89.77	85.09	87.43	99.78	83.48	86.63	89.73	83.77	86.75
11	52.57	44.43	**48.50**	40.09	53.19	46.64	34.97	28.13	31.55	22.08	29.54	25.81	34.19	40.07	37.13
12	79.10	73.10	**76.10**	79.69	68.87	74.28	67.22	79.88	73.55	70.27	79.65	74.96	68.63	83.11	75.87

表3　基于Naive Bayes的分类精度比较结果(50%标记率)

No	OpenSSFS			FSCL			Relief-F			FSIG			FSMI		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
1	100	100	**100**	100	86.66	93.33	100	86.66	93.33	100	60.00	80.00	70.01	63.33	66.67
2	97.63	90.71	**94.17**	88.87	86.29	87.58	83.34	85.00	84.17	85.83	70.83	78.33	90.91	59.09	75.00
3	100	94.54	**97.27**	85.65	94.27	89.96	82.73	79.09	80.91	61.82	83.64	72.73	74.76	90.70	82.73
4	83.34	100	**91.67**	76.19	90.47	83.33	34.92	42.86	38.89	34.92	42.86	38.89	34.92	42.86	38.89
5	100	98.06	**99.03**	100	92.92	96.46	92.18	88.04	90.11	92.64	87.04	89.84	92.41	88.39	90.40
6	67.74	18.38	**43.06**	34.64	48.42	41.53	40.00	31.12	35.56	23.32	18.06	20.69	22.38	19.00	20.69
7	95.27	91.39	**93.33**	90.74	88.52	89.63	91.07	88.19	89.63	84.23	80.95	82.59	88.98	82.14	85.56
8	44.10	22.56	**33.33**	33.33	22.23	27.78	33.33	22.23	27.78	29.19	15.25	22.22	20.20	13.14	16.67
9	74.31	69.67	**71.99**	68.89	61.35	65.12	68.89	54.73	61.81	58.64	48.58	53.61	53.92	53.58	53.75
10	65.20	64.30	**64.75**	61.82	61.54	61.68	61.57	48.89	55.23	61.11	55.73	58.42	62.16	52.88	57.52
11	22.22	26.38	**24.30**	20.87	24.47	22.67	33.60	8.98	21.29	18.68	21.24	19.96	15.51	28.39	21.95
12	61.91	64.33	**63.12**	61.18	54.66	57.92	57.32	49.46	53.39	42.09	49.31	45.70	50.16	66.56	58.36

从表2～表5中的实验结果可以看出，在抽取50%的已知类样本作为标记数据的前提下，OpenSSFS算法的分类性能在所有数据集上均优于其余算法。具体而言，在小规模数据集上，OpenSSFS算法在数据集Lung cancer上的分类性能提升最为显著，提升了约70%；在高维数据集上，OpenSSFS算法同样展现了优良的分类性能，特别是在Nci9数据集上，分类精度提升了近一倍；在较大规模的数据集Krkopt上，OpenSSFS算法的分类性能提升也尤为显著；另外，实验结果还进一步验证了OpenSSFS算法对于类别数较多的数据集也能展现很好的分类性能。OpenSSFS算法能求解出性能更优的特征子集，主要归功于以下几方面的改进：首先，本文在借鉴经典Relief-F算法计算特征权重的

表4 基于RandomForest的分类精度比较结果(50%标记率)

No	OpenSSFS			FSCL			Relief-F			FSIG			FSMI		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
1	100	100	**100**	100	100	100	100	100	100	100	100	100	83.33	63.33	73.33
2	100	95.00	**97.50**	91.13	90.09	90.61	86.54	95.12	90.83	100	81.66	90.83	100	81.66	90.83
3	100	90.90	**95.45**	86.36	93.02	89.69	77.08	93.82	85.45	83.24	65.86	74.55	79.74	83.90	81.82
4	63.88	75.00	**69.44**	41.17	58.83	50.00	32.07	45.71	38.89	33.33	27.79	30.56	21.12	40.00	30.56
5	100	99.66	**99.83**	100	98.14	99.07	100	98.16	99.08	97.12	87.42	92.27	92.79	92.39	92.59
6	48.39	47.45	**47.92**	35.13	46.15	40.64	38.87	31.69	35.28	39.11	38.39	38.75	43.60	34.74	39.17
7	71.25	79.13	**75.19**	67.91	67.65	67.78	69.68	66.62	68.15	52.95	61.13	57.04	57.78	63.70	60.74
8	48.25	29.53	**38.89**	39.12	16.44	27.78	35.72	19.84	27.78	25.00	19.44	22.22	22.71	21.73	22.22
9	90.17	73.89	**82.03**	88.79	70.91	79.85	80.23	76.71	78.47	88.57	62.55	75.56	78.89	71.97	75.42
10	96.23	89.77	**93.00**	95.43	88.13	91.78	95.73	89.73	92.73	100	84.70	92.35	100	85.06	92.53
11	73.46	65.92	**69.69**	59.94	48.18	54.06	39.78	19.60	29.69	23.81	29.65	26.73	39.54	56.10	47.82
12	69.67	83.73	**76.70**	79.34	69.98	74.66	73.68	66.16	69.92	61.22	85.38	73.30	79.61	70.61	75.11

表5 基于C4.5的分类精度比较结果(50%标记率)

No	OpenSSFS			FSCL			Relief-F			FSIG			FSMI		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
1	100	100	**100**	97.77	88.89	93.33	95.00	78.34	86.67	73.34	100	86.67	86.67	33.33	60.00
2	100	95.00	**97.50**	85.71	91.57	93.64	75.00	100	87.50	95.65	72.69	84.17	97.22	77.78	87.50
3	100	96.36	**98.18**	92.32	87.66	89.99	75.46	86.36	80.91	98.59	59.59	79.09	70.35	87.83	79.09
4	63.88	75.00	**69.44**	49.44	45.00	47.22	35.77	42.01	38.89	42.78	35.00	38.89	31.43	40.79	36.11
5	100	96.42	**98.21**	97.32	93.44	95.38	98.04	93.32	95.68	86.32	83.94	85.13	85.73	84.63	85.18
6	33.87	25.29	29.58	24.93	29.23	27.08	33.64	18.02	25.83	29.45	36.67	**33.06**	32.43	18.95	25.69
7	69.78	73.92	**71.85**	55.83	59.73	57.78	54.69	60.13	57.41	49.96	44.12	47.04	55.24	52.16	53.70
8	51.11	37.77	**44.44**	44.87	32.91	38.89	42.63	35.15	38.89	40.17	26.49	33.33	38.83	16.73	27.78
9	88.11	64.69	76.40	75.56	68.58	72.07	88.23	53.71	70.97	78.57	71.99	75.28	78.89	75.83	**77.36**
10	87.78	76.16	81.97	85.66	73.84	79.75	89.73	74.01	81.87	88.14	76.72	**82.43**	99.83	63.27	81.55
11	79.53	58.97	**69.25**	69.75	56.43	63.09	37.24	17.22	27.23	33.17	17.01	25.09	36.16	48.08	42.12
12	66.20	75.58	70.89	70.34	65.70	68.02	76.67	55.53	66.10	76.35	69.19	**72.77**	60.33	76.77	68.55

表6 基于SVM的分类精度比较结果(10%标记率)

No	OpenSSFS			FSCL			Relief—F			FSIG			FSMI		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
1	93.33	66.67	**80.00**	85.00	75.00	80.00	85.00	75.00	80.00	100	60.00	80.00	56.45	63.55	60.00
2	85.70	77.64	81.67	78.33	76.83	77.58	75.00	61.66	68.33	91.30	83.70	**87.50**	100	75.00	87.50
3	97.17	90.65	**93.91**	86.07	76.67	81.37	78.39	71.61	75.00	75.05	69.59	72.32	89.09	69.09	79.09
4	57.14	65.08	**61.11**	39.27	55.17	47.22	23.33	37.79	30.56	17.38	38.18	27.78	21.11	40.01	30.56
5	96.12	93.38	**94.75**	93.54	90.42	91.98	90.07	88.33	89.20	51.04	69.38	60.21	54.61	68.19	61.40
6	44.71	43.27	**43.99**	36.95	43.89	40.42	37.51	43.33	40.42	37.23	41.11	39.17	41.90	38.94	40.42
7	51.55	59.31	**55.43**	46.85	62.79	54.82	50.81	59.33	55.07	40.00	60.00	50.00	33.33	66.67	50.00
8	26.27	7.07	**16.67**	14.96	7.26	11.11	12.97	9.25	11.11	15.00	7.22	11.11	18.43	3.79	11.11
9	56.08	71.88	**63.98**	60.12	63.18	61.65	65.07	53.23	59.17	58.57	44.21	51.39	55.48	50.00	52.92
10	88.24	87.82	**88.03**	84.23	88.41	86.32	89.50	85.30	87.40	62.08	56.18	59.13	88.07	71.09	79.58
11	52.18	44.58	**48.38**	51.18	42.06	46.62	22.46	24.16	23.31	24.44	23.90	24.17	27.14	28.04	27.59
12	73.11	78.77	**75.94**	81.80	64.62	73.21	73.38	70.06	71.72	71.05	76.01	73.53	87.17	63.81	75.49

基础上,引入了耦合学习的思想,深入挖掘特征以及类别中隐藏的有用信息;其次,通过引入新的样本相似性度量,在伪标签生成过程及基于聚类的新类发现过程中更精确地度量了数据样本间的相似性;最

表7　基于Naive Bayes的分类精度比较结果(10%标记率)

No	OpenSSFS			FSCL			Relief-F			FSIG			FSMI		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
1	100	86.66	**93.33**	100	86.66	93.33	85.00	75.00	80.00	100	60.00	80.00	56.45	63.55	60.00
2	100	81.66	**90.83**	93.48	75.00	84.24	89.16	67.50	78.33	58.26	51.74	55.00	78.57	71.43	75.00
3	100	94.54	**97.27**	90.02	86.36	88.19	81.21	73.33	77.27	71.17	56.11	63.64	79.84	56.52	68.18
4	63.88	75.00	**69.44**	47.66	69.00	58.33	23.33	37.79	30.56	25.71	35.41	30.56	30.00	25.56	27.78
5	98.76	94.66	**96.71**	88.09	94.65	91.37	86.23	92.89	89.56	78.17	79.87	79.02	87.59	71.05	79.32
6	43.11	39.35	**41.23**	58.04	23.08	40.56	31.95	33.33	32.64	13.51	20.11	16.81	22.38	19.00	20.69
7	94.12	85.88	**90.00**	87.43	85.09	86.26	87.73	84.13	85.93	80.17	76.27	78.22	80.98	78.48	79.73
8	37.26	18.30	**27.78**	31.40	13.04	22.22	30.53	13.91	22.22	29.19	15.25	22.22	20.20	13.14	16.67
9	72.45	71.39	**71.92**	69.97	55.71	62.84	58.59	51.13	54.86	52.86	32.42	42.64	50.27	41.67	45.97
10	66.17	62.99	**64.58**	52.52	66.82	59.67	50.18	60.08	55.13	55.08	20.82	37.95	58.14	48.66	53.40
11	17.14	31.42	**24.28**	20.74	24.48	22.61	18.09	22.97	20.53	31.48	6.86	19.17	24.17	16.03	20.10
12	67.98	58.12	**63.05**	61.61	48.79	55.20	41.79	56.47	49.13	34.68	35.16	34.92	61.05	53.59	57.32

表8　基于RandomForest的分类精度比较结果(10%标记率)

No	OpenSSFS			FSCL			Relief-F			FSIG			FSMI		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
1	100	100	**100**	100	100	100	100	100	100	100	100	100	54.33	79.01	66.67
2	100	81.66	**90.83**	81.22	100	90.61	81.66	100	90.83	95.65	79.35	87.50	85.22	89.78	87.50
3	96.36	87.28	**91.82**	73.24	88.44	80.84	85.31	60.15	72.73	65.94	57.70	61.82	77.32	64.50	70.91
4	48.42	57.14	**52.78**	33.33	50.01	41.67	26.37	51.41	38.89	25.71	35.41	30.56	30.17	30.95	30.56
5	100	98.42	**99.21**	98.14	96.52	97.33	99.17	97.75	98.46	93.33	89.97	91.65	90.48	90.34	90.41
6	45.15	40.97	**43.06**	40.00	36.06	38.03	27.23	43.33	35.28	29.73	33.61	31.67	41.00	34.56	37.78
7	67.44	73.44	**70.44**	65.14	56.42	60.78	64.07	60.11	62.09	50.69	55.97	53.33	54.96	60.44	57.70
8	42.76	23.90	**33.33**	36.67	14.91	25.79	34.50	18.18	26.34	24.44	20.00	22.22	22.71	21.73	22.22
9	100	61.84	**80.92**	81.49	71.53	76.51	78.15	72.41	75.28	78.89	71.51	75.28	73.62	50.00	61.81
10	100	85.86	**92.93**	100	83.00	91.50	100	85.20	92.60	68.35	63.91	66.13	89.74	87.26	88.50
11	69.97	68.87	**69.42**	47.93	39.93	43.93	14.51	25.13	19.82	37.35	13.11	25.23	28.66	38.92	33.79
12	71.60	81.80	**76.70**	72.37	68.35	70.36	52.47	82.35	67.41	76.46	68.18	72.32	66.23	77.21	71.72

表9　基于C4.5的分类精度比较结果(10%标记率)

No	OpenSSFS			FSCL			Relief-F			FSIG			FSMI		
	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
1	100	86.66	**93.33**	86.66	86.68	86.67	95.00	78.34	86.67	85.98	87.36	86.67	27.69	52.31	40.00
2	86.92	94.74	**90.83**	85.24	89.30	87.27	77.43	90.91	84.17	84.55	83.79	84.17	94.71	80.29	87.50
3	100	94.54	**97.27**	91.89	82.53	87.21	79.17	75.37	77.27	66.37	80.91	73.64	85.52	70.84	78.18
4	49.83	55.73	**52.78**	39.16	44.18	41.67	23.33	37.79	30.56	35.71	19.35	27.78	30.56	25.00	27.78
5	98.44	93.82	**96.13**	97.06	89.08	93.07	98.81	88.79	93.80	86.17	83.61	84.89	89.42	80.86	85.14
6	34.71	16.95	**25.83**	18.00	25.62	21.81	11.35	26.71	19.03	24.44	21.94	23.19	27.78	18.88	23.33
7	59.32	63.34	**61.48**	52.40	55.00	53.70	53.14	55.00	54.07	44.44	47.42	45.93	52.33	40.27	46.30
8	42.32	35.46	**38.89**	34.56	21.00	27.78	26.57	17.87	22.22	25.21	19.23	22.22	22.69	21.75	22.22
9	83.54	64.74	**74.14**	61.86	72.86	67.36	68.38	62.18	65.28	77.73	65.61	71.67	69.16	50.00	59.58
10	91.23	71.93	**81.58**	87.3	71.64	79.47	88.38	74.36	81.37	70.78	57.86	64.32	80.93	79.67	80.30
11	60.00	78.18	**69.09**	66.68	59.42	63.05	6.54	26.54	16.54	13.42	33.62	23.52	35.25	20.49	27.87
12	69.90	70.68	70.29	66.38	68.62	67.50	63.87	66.01	64.94	63.83	81.59	**72.71**	67.21	68.85	68.03

后，本文设计的基于成对相似性的自适应伪标签生成方法区别于传统的阈值判定方法，根据样本间的耦合关系弹性选择相似度更高的标记样本，能够鲁棒地生成高质量伪标签，更加适用于处理非平衡数

据集。综上所述，针对开放世界数据集，OpenSSFS算法能够找到性能更优的有效特征子集。

从表6～表9中的实验结果可以看出，在抽取10%样本作为标记数据的前提下，OpenSSFS算法依然表现出较好的性能，在保持较高的精度的情况下，受标记率影响程度最低。与表2～表5中所展示的算法在50%标记率的数据集上的分类性能相比，OpenSSFS算法的分类性能未出现明显的下降，表明OpenSSFS算法具有较强的鲁棒性和泛化性。另外，在表6和表9中的部分数据集上，OpenSSFS算法的分类性能略低于对比算法，这主要由于该算法在求解特征权重的过程中会受随机抽样以及抽样次数的影响。

6.2 特征选择结果比较

为进一步比较本文提出的OpenSSFS算法的性能，本节在表10列出了该算法和四种对比算法在所有数据集上50%标记率情况下的特征选择对比结果。根据表10中的实验比较结果可得，OpenSSFS算法在大部分数据集上均能找到一个特征个数与其余算法接近的特征子集；结合5.1节中的实验结果可得，OpenSSFS算法在特征个数接近的前提下，可求解出分类性能更优的特征子集。其中，除了数据集MPE和Krkopt，OpenSSFS算法求解出的特征子集长度与对比算法接近甚至更短。结合表2～表5中的实验结果可得，OpenSSFS算法在这些数据集上的分类性能明显优于其余算法，表明该算法能够在大部分数据集上求解出更短但性能更优的特征子集，进一步验证了本文提出的OpenSSFS算法的有效性。

6.3 特征选择高效性分析

为进一步验证算法OpenSSFS的高效性，本节中对比分析了该算法和FSCL算法以及Relief-F算法在所有数据集上基于50%已知类，标记率分别是10%和50%对应的计算时间。实验结果见表11，该表中的“提高率”一列表示OpenSSFS相比较Relief-F算法的计算时间的提高百分比。从表11中的实验结果可得，算法在标记率50%的数据集上的计算耗时低于标记率10%数据集对应的计算耗时，其原因是由于标记样本越少则需要更多的时间求解相似样本。在规模较大或者维度较高的数据集上（数据集5-12），算法OpenSSFS和FSCL的计算时间比较接近，但计算效率低于Relief-F算法。其主要原因是由于OpenSSFS和FSCL中采用了基于耦合学习的样本相似性度量。该度量的计算时间复杂度高于Relief-F算法中采用的相似性度量。

表10 在数据集由50%的已知类和50%的新类组成，只有50%的已知类被标记，特征选择数量比较

No	OpenSSFS	FSCL	Relief-F	FSIG	FSMI
1	11	11	12	12	12
2	9	11	11	11	11
3	13	16	15	16	16
4	10	13	14	17	17
5	66	59	65	61	69
6	55	54	55	54	58
7	1983	1998	2020	2020	2026
8	4616	5284	5939	6455	6480
9	23	26	26	25	26
10	12	14	13	13	15
11	6	6	4	4	4
12	17	17	17	18	18

此外，表11中的实验结果也进一步表明，针对高维数据集，OpenSSFS算法的计算耗时仍比较高，这也是我们后续研究内容的重点之一，探索更高效的面向高维数据集的耦合性相似度度量以及相应的开放世界特征选择算法。

6.4 已知类和新类判别分析

为了进一步评估本文提出的OpenSSFS算法在不同已知类比例下的性能表现，本节选取了表1中的四个数据集，并采用随机森林（RandomForest）分类器进行了实验，结果见图2。其中，由于数据集Dermatology和Zoo的类别数量较少，如果将已知类的比例起始值设置为0.1，仅有1个已知类别，对实验没有实际意义。因此，这两个数据集的已知类比例区间设定为0.2到0.9，剩余两个数据集的已知类比例区间设定为0.1到0.9。图中的纵轴表示相应的分类精度。根据图2的实验结果可得，随着已知类比例的增加，实验中选取的四种算法的整体性能均在上升，但是OpenSSFS算法的分类精度均高于其余算法，展现了良好的分类性能，进一步证明了OpenSSFS算法受已知类别数量的影响较少，体现了该算法具有较强的适用性。尤其是在较低的已知类比例下，相比较其余三种算法，OpenSSFS算法仍能获取到较高的分类精度，也再次验证了该算法的鲁棒性。综上分析，OpenSSFS算法可以更有效地处理开放性更高的任务场景。

6.5 标记数据比例分析

本节针对OpenSSFS算法在多组数据集上不同标记数据比例的测试与对比分析进行了实验，结果如图3所示。该图中横轴表示每组数据上的标记数据比例范围是从0.1到0.9，纵轴表示了在标记数据

表11　在数据集由50%的已知类和50%的新类组成，已知类被标记率为10%和50%情况下时间效率比较

No	10%				50%			
	OpenSSFS(s)	FSCL(s)	Relief−F(s)	提高率(%)	OpenSSFS(s)	FSCL(s)	Relief−F(s)	提高率(%)
1	0.6163	0.8206	3.7123	83.40	0.5595	0.7878	3.0247	81.50
2	0.9672	1.5226	6.0070	83.90	0.9476	1.4420	5.7308	83.46
3	8.1257	11.4027	49.2100	83.49	7.1553	10.6210	47.3560	84.89
4	0.6615	0.8226	3.9591	83.29	0.5078	0.6778	3.4734	85.38
5	694.6028	676.8090	653.2971	−6.32	641.9543	628.9071	608.9494	−5.42
6	266.7854	232.1065	209.4379	−27.38	218.1449	208.2466	166.5999	−30.94
7	3965.5274	3806.7460	3553.7714	−11.59	3413.0481	3221.5536	3178.2890	−7.39
8	4531.1249	4592.8871	4579.1140	1.05	4359.8578	4373.3618	4331.9427	−0.64
9	176.6213	172.8065	170.4198	−3.64	124.4675	120.8990	105.2582	−18.25
10	7613.3925	7490.5500	7034.5873	−8.23	6992.0485	6860.8901	6753.6149	−3.53
11	6482.5678	6442.3300	6427.0593	−0.86	6041.5335	6148.4294	5861.5044	−3.07
12	2328.1650	2275.8398	1967.0032	−18.36	1570.1729	1427.6901	1240.8986	−26.54
Ave	**2172.43**	**2142.054**	**2054.798**	−−	**1947.533**	**1916.959**	**1858.887**	−−

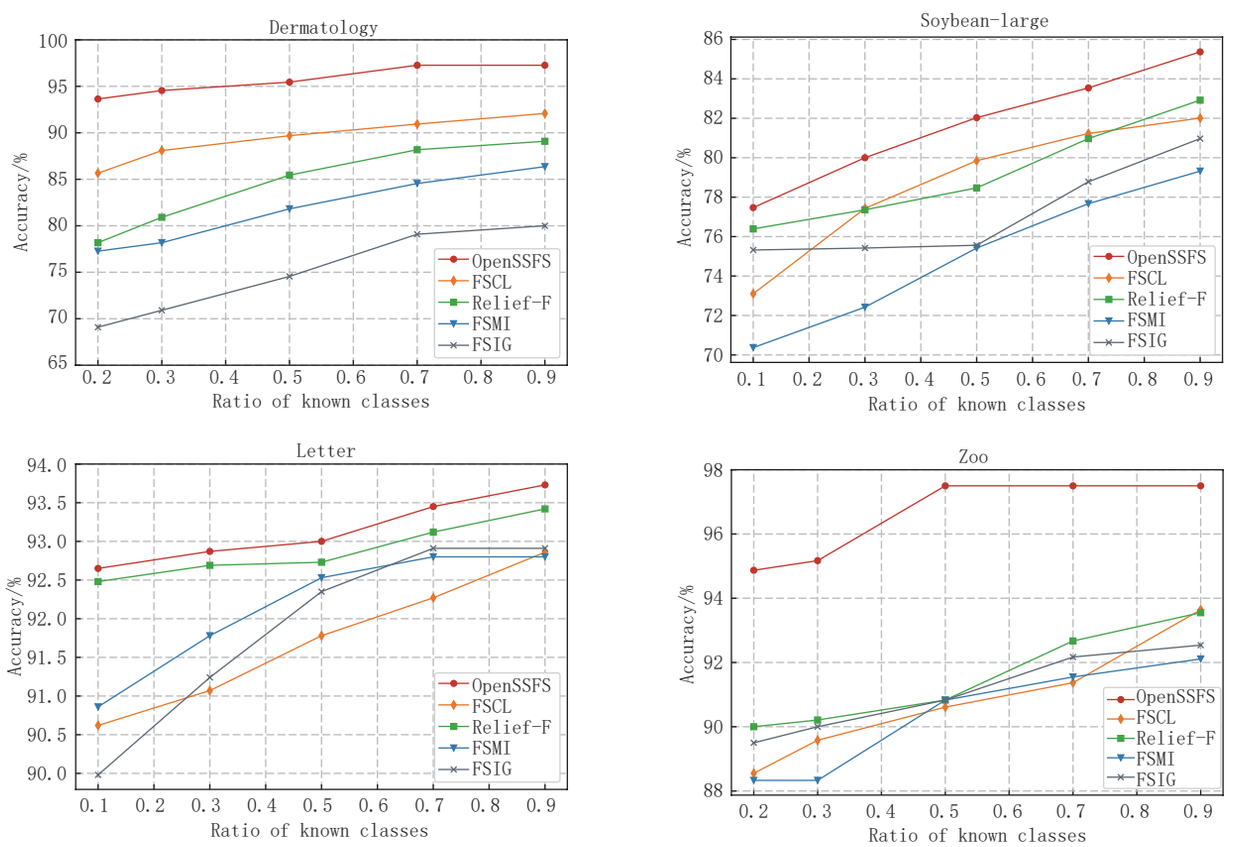

图2　五种算法基于四组数据集基于不同已知类比例的分类性能比较

不同比例值下对应的分类精度。图3中的实验结果表明，在不同标记数据比例下，OpenSSFS算法的整体性能均优于其余四种对比算法，且受标记数据比例变化的影响较小，即随着标记数据比例值的不断增加，OpenSSFS算法的性能呈现了较平稳的上升；而其余四种算法的性能随着标记数据比例的变化，均呈现了较为明显的上升。因此，本节中的实验结果进一步验证了OpenSSFS算法的稳定性和较强的适用性，尤其在标记数据占比比较低的情况下，仍能求解出有效的特征子集，进而呈现了更优的分类性能。结合6.3节中的实验结果可得，OpenSSFS算法不仅可以更好地处理开放性更高的学习任务，而且能更好地处理标记数据较少的现实场景，有效节省了训练成本。

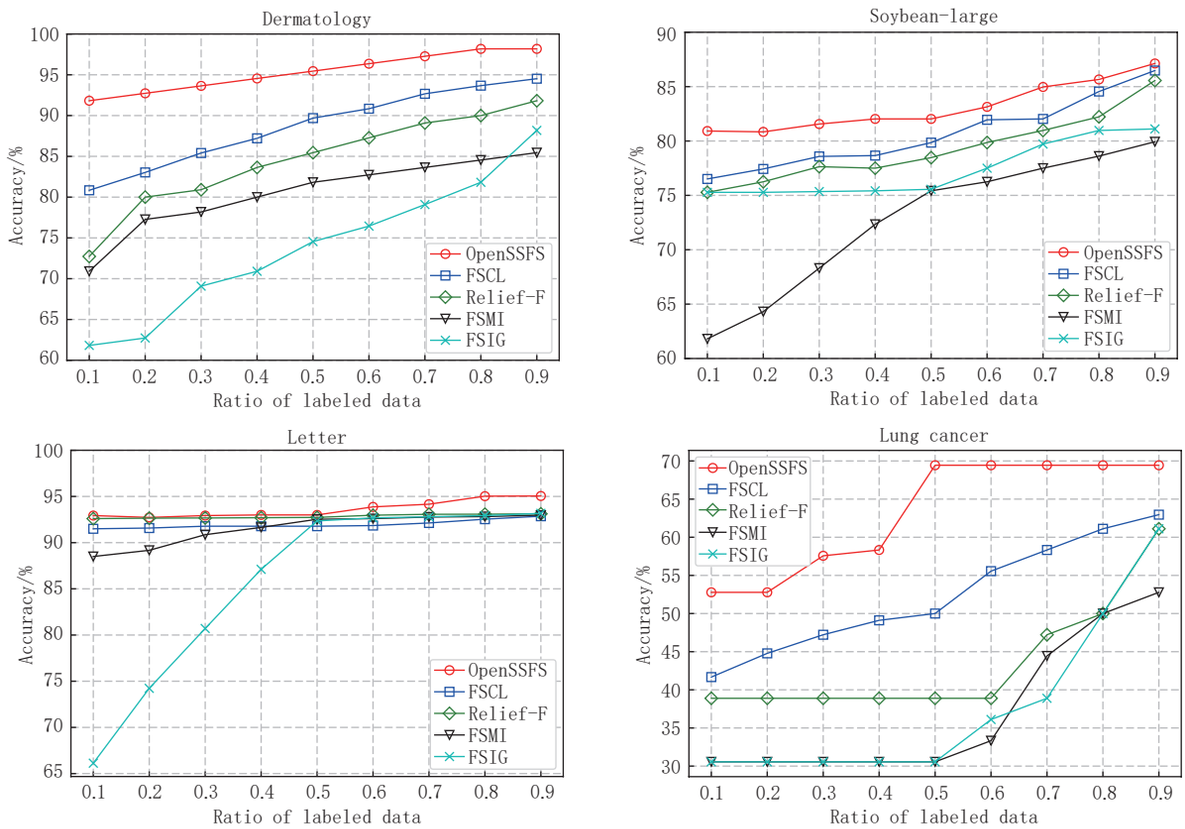

图3 五种算法基于多组数据集基于不同标记数据比例的性能比较

6.6 参数敏感度分析

本节实验对算法设计中的参数进行了分析和比较，表12和图4中分别给出了在不同的耦合阈值σ下OpenSSFS算法选取到的特征数量和分类精度变化图。本节的实验设置是在数据集上选取50%的已知类和50%的新类，并包含50%标记已知类数据样本。实验中选取了6个有代表意义的数据集，分别是较高维数据集Lung cancer，较低维数据集Zoo和Krkopt，较小规模数据集Dermatology和Soybean-large，以及较大规模数据集Letter。图4中的横轴为耦合阈值σ不同的取值，范围为0.02到0.18，纵轴是基于随机森林分类器的分类精度。根据表12和图4中的实验结果可得，Dermatology和Soybean-large两个数据集的维数和规模比较接近，参数σ的变化对模型性能的影响并不明显，分类精度的峰值在0.08到0.1之间，考虑到取0.08会增加选择到的特征数量，进而增加计算耗时，所以将0.1设定为最终阈值；Lung cancer数据集由于仅包含33个样本，但属于较高维度数据集，所以分类结果受参数影响较大，其分类精度在阈值为0.1时最高；Krkopt数据集在参数σ取值0.02到0.1时选取到的特征数量都是6，对应的分类精度也是相同的；因

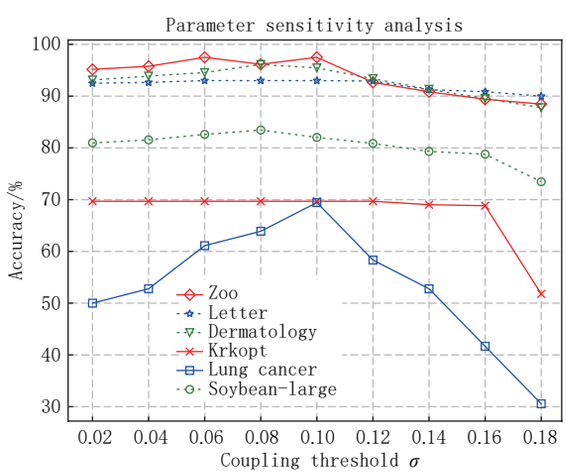

图4 耦合阈值分析

表12 不同耦合阈值特征选择数量分析

Dataset	Coupling threshold								
	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18
Soybean-large	26	25	24	24	23	22	21	21	18
Lung cancer	18	17	14	12	10	8	8	7	6
Zoo	11	11	10	9	9	8	7	7	7
Dermatology	18	17	16	14	13	11	10	9	8
Letter	15	14	13	12	12	12	11	11	10
Krkopt	6	6	6	6	6	5	5	5	4

此，图4中所有数据集的分类精度均在参数σ取值在0.08到0.1之间时最高。显然，参数σ越小，可以选取到越多的有效特征，但是会存在冗余；而参数σ越大，会遗漏一些关键特征，相应的分类精度会明显下降。综上分析，本文实验测试中将耦合阈值取σ设定为0.1，不仅可以得到较高的分类精度，同时也可选择较少的特征，减少模型计算量，降低计算耗时。

6.7　混淆矩阵分析

本节计算了OpenSSFS算法基于Soybean-large数据集在所有测试样本上的混淆矩阵，并依此进一步深入分析了OpenSSFS算法在已知类和新类上的性能，具体的结果见图5。本节实验的分析中，训练阶段Soybean-large数据集由50%的已知类和50%的新类组成，且有10%的已知类被标记，分类器使用随机森林(RandomForest)分类器，分类精度采用十折交叉验证方法确定最终结果。根据下图中的实验结果可得，OpenSSFS算法可以成功识别新类，而且不会将已知类与新类混淆，且对于包含大量新类的数据集也依然呈现了良好的性能。

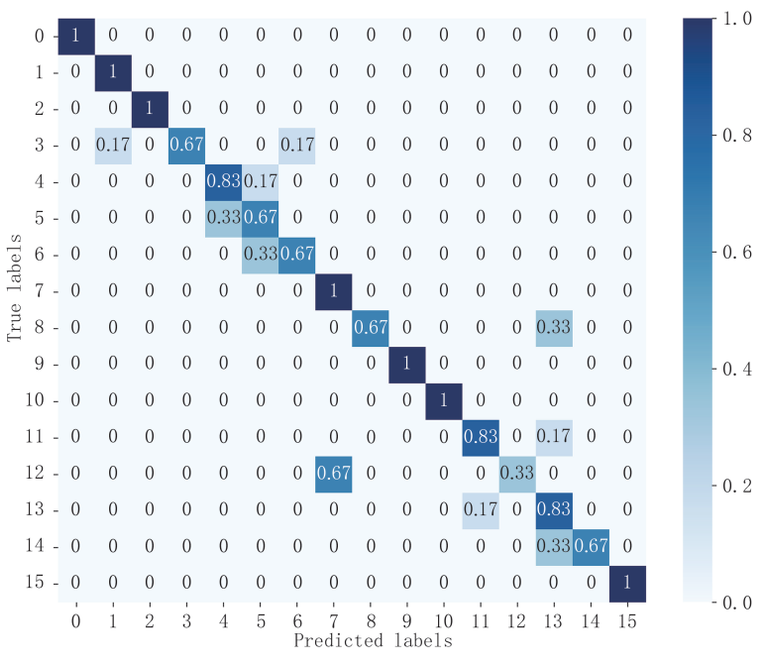

图5　基于Soybean-large数据集测试样本的混淆矩阵

7　结　论

本文针对开放世界中如何有效处理未标记样本所面临的挑战，通过构建面向未标记已知类样本的两级判别函数和面向未标记未知类样本的新类发现方法，设计了一种面向开放世界的半监督特征选择算法。相关的分析以及实验结果进一步验证了该算法不仅能够找到性能更优的有效特征子集，而且对已知类标记样本的标记率具有较低的敏感性，表现出较强的鲁棒性。本研究有望丰富开放世界学习的理论和方法，对开放环境下的机器学习方法具有重要的理论意义和科学价值。下一步的研究将重点是探索适用于更高维数据的更有效的开放世界学习方法和及其处理机制。

参　考　文　献

[1] Yang X L, Song Z X, King I, et al. A survey on deep semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(9): 8934-8954

[2] Kmita K, Kaczmarek-Majer K, Hryniewicz O. Explainable impact of partial supervision in semi-supervised fuzzy clustering. IEEE Transactions on Fuzzy Systems, 2024, 32(5): 3189-3198

[3] Wu H M, Li X M, Cheng K T. Exploring feature representation learning for semi-supervised medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(11): 16589-16601

[4] Sun N Z, Luo T J, Zhuge W Z, et al. Semi-supervised learning with label proportion. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(1): 877-890

[5] Ruff L, Kauffmann J.R, Vandermeulen R.A, et al. A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE, 2021, 109(5): 756-795

[6] Boukerche A, Zheng L N, Alfandi O. Outlier detection: Methods, models, and classification. ACM Computing Surveys, 2020, 53(3): 1-37

[7] Van Engelen J. E, Hoos H. H. A survey on semi-supervised learning. Machine Learning, 2020, 109(2): 373-440

[8] Jiang K, Xie W Y, Lei J, et al. Lren: Low-rank embedded network for sample-free hyperspectral anomaly detection// Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021: 4139-4146

[9] Geifman Y, El-Yaniv R. Selectivenet: A deep neural network with an integrated reject option//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 2151-2159

[10] Scheirer W J, de Rezende Rocha A, et al. Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(7): 1757-1772

[11] Geng C X, Huang S J, Chen S C. Recent advances in open set recognition: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3614-3631

[12] Yang H M, Zhang X Y, Yin F, et al. Convolutional prototype network for open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(5): 2358-2370

[13] Zhu Peng-Fei, Zhang Wan-Ying, Wang Yu, et al. Multi-granularity inter-class correlation based contrastive learning for open set recognition. Journal of Software, 2022, 33(4): 1156-1169 (in Chinese)
(朱鹏飞，张琬迎，王煜等. 考虑多粒度类相关性的对比式开放集识别方法. 软件学报，2022，33(4): 1156-1169)

[14] Liu Chang, Yang Chun, Yin Xu-Cheng. Open-set text recognition via part-based similarity. Acta Automatica Sinica, 2024, 50(10): 1977-1987 (in Chinese)
(刘畅，杨春，殷绪成. 基于文字局部结构相似度量的开放集文字识别方法. 自动化学报，2024，50(10): 1977-1987)

[15] Bendale A, Boult T. Towards open set deep networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1563-1572

[16] Hsu Y C, Lv Z Y, Zsolt K. Learning to cluster in order to transfer across domains and tasks//Proceedings of the International Conference on Learning Representations. Toulon, France: ICLR, 2018, DOI:10.48550/arXiv.1711.10125

[17] Han K, Vedaldi A, Zisserman A. Learning to discover novel visual categories via deep transfer clustering//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 8400-8408

[18] Zhong Z, Fini E, Roy S, et al. Neighborhood contrastive learning for novel class discovery//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 10862-10870

[19] Guo L Z, Zhang Y G, Wu Z F, et al. Robust semi-supervised learning when not all classes have labels//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022

[20] Han K, Rebuffi S A, Ehrhardt S, et al. Autonovel: Automatically discovering and learning novel visual categories. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6767-6781

[21] Feng Yao-Gong, Yu Jian, Sang Ji-Tao, et al. Survey on knowledge-based zero-shot visual recognition. Journal of Software, 2021, 32(2): 370-405 (in Chinese)
(冯耀功，于剑，桑基韬等. 基于知识的零样本视觉识别综述. 软件学报，2021，32(2): 370-405)

[22] Chen Z, Luo Y D, Qiu R H, et al. Semantics disentangling for generalized zero-shot learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 8692-8700

[23] Li J J, Jing M M, Lu K, et al. Leveraging the invariant side of generative zero-shot learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7394-7403

[24] Yu Y L, Ji Z, Han J G, et al. Episode-based prototype generating network for zero-shot learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA,2020, 14032-14041

[25] Li J, Lan X G, Long Y, et al. A joint label space for generalized zero-shot classification. IEEE Transactions on Image Processing, 2020, 29: 5817-5831

[26] Chen S M, Xie G S, Peng Q M, et al. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning//Proceedings of the Advances in Neural Information Processing Systems. Online, 2021: 16622-16634

[27] Liu J M, Wang Y Q M, Zhang T Z, et al. Open-world semi-supervised novel class discovery//Proceedings of the International Joint Conference on Artificial Intelligence. Cape Town, South Africa, 2023

[28] Cao K, Brbic M, Leskovec J. Open-world semi-supervised learning// Proceedings of the International Conference on Learning Representations. Online, 2022

[29] Vaze S, Han K, Vedaldi A, et al. Generalized category discovery//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 7482-7491

[30] Rizve M. N, Kardan N, Khan S, et al. Openldn: Learning to discover novel classes for open-world semi-supervised learning// Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel, 2022: 382-401

[31] Rizve M. N, Kardan N, Shah M. Towards realistic semi-supervised learning//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel. 2022: 437-455

[32] Zhao T H, Lin Y T, Wu Y, et al. Promote knowledge mining towards open-world semi-supervised learning. Pattern Recognition, 2024, 149: 110259

[33] Zhou P, Zhang Y Y, Ling Z L, et al. Online heterogeneous streaming feature selection without feature type information. IEEE Transactions on Big Data, 2014, 10(4): 470-485

[34] Fan W, Liu K P, Liu H, et al. Interactive reinforcement learning for feature selection with decision tree in the loop. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(2): 1624-1636

[35] Li Yong-Hao, Hu Liang, Gao Wan-Fu. Multi-label feature selection based on sparse coefficient matrix reconstruction. Chinese Journal of Computer, 2022, 45(9): 1827-1841 (in Chinese)
(李永豪，胡亮，高万夫. 基于稀疏系数矩阵重构的多标记特征选择. 计算机学报，2022，45(9): 1827-1841)

[36] Qian W B, Xu F K, Huang J T, et al. A novel granular ball computing-based fuzzy rough set for feature selection in label distribution learning. Knowledge-Based Systems, 2023, 278: 110898

[37] Chen H T, Wang Y, Hu Q H. Multi-granularity regularized re-balancing for class incremental learning. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(7): 7263-7277

[38] Ma N, Hu Q H, Wu K J, et al. A dissimilarity measure powered feature weighted fuzzy C-means algorithm for gene expression data. IEEE Transactions on Fuzzy Systems, 2025, 33(1): 192-202

[39] Wang Feng, Yao Zhen, Liang Jiye. Multi-granulation incremental feature selection algorithm for dynamic hybrid data. Journal of Software, 2025, 36(3): 1186-1201 (in Chinese)
(王锋，姚珍，梁吉业. 面向动态混合数据的多粒度增量特征选择算法. 软件学报，2025，36(3): 1186-1201)

[40] Wang F, Liang J Y, Song P. Coupling learning for feature selection in categorical data. International Journal of Machine Learning and Cybernetics, 2023, 14(7):2455-2465

[41] Shi D, Zhu L, Li J J, et al. Binary label learning for semi-supervised feature selection. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(3): 2299-2312

[42] Li X L, Zhang Y X, Z R. Semisupervised feature selection via generalized uncorrelated constraint and manifold embedding. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(9): 5070-5079

[43] Li G J, Yu Z W, Yang K X, et al. P. Exploring feature selection with limited labels: a comprehensive survey of semi-supervised and unsupervised approaches. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 6124-6144

[44] Huang Z X. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304

[45] Robnik- Šikonja M, Kononenko I. Theoretical and empirical analysis of relieff and rrelieff. Machine Learning, 2003, 53(1-2): 23-69

[46] Quinlan J. R. Induction of decision trees. Machine Learning, 1986, 1: 81-106

[47] Battiti R. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, 1994, 5(4): 537-550

[48] Estevez P A, Tesmer M, Perez C A, et al. Normalized mutual information feature selection. IEEE Transactions on Neural Networks, 2009, 20(2): 189-201

WANG Feng, Ph. D., associate professor. Her main research interests include data mining, machine learning, and granular computing.

WU Wen-Qiang, M. S. candidate. His main research interests include dynamic data mining and opening set recognition.

LIANG Ji-Ye, Ph. D., professor. His main research interests include granular computing, data mining and machine learning.

Background

Existing semi-supervised learning methodologies typically operate under the closed-world assumption, wherein category information remains static throughout the learning process; that is, the labeled data utilized for model training encompasses all categories. However, this premise is frequently challenging to satisfy in practical applications due to the presence of numerous unknown class samples within unlabeled datasets. Consequently, researchers have identified a highly demanding research avenue in recent years: extending semi-supervised learning not only to effectively identify unlabeled samples from known classes but also to learn new unknown classes, thereby establishing an open-world semi-supervised learning framework. To tackle this challenge, this paper introduces a semi-supervised feature selection algorithm tailored for open-world scenarios based on symbolic data (OpenSSFS). The algorithm incorporates coupled learning into the measurement of symbolic sample similarity and category relevance analysis, and subsequently constructs an adaptive pseudo-label generation algorithm for unlabeled known-class data, a granulation and novel class discovery algorithm for unlabeled unknown-class data, as well as a feature selection algorithm based on category relevance.