基于自适应密度邻域关系的多标签在线流特征选择

张海翔[1],李培培[2],胡学钢[2]
(1. 蚌埠医学院附属合肥市第二人民医院 讯息处,安徽 合肥 230012;
2. 合肥工业大学 大数据知识工程教育部重点实验室,安徽 合肥 230601)

摘 要:流特征选择指从以流形式到来的特征数据中选出最优特征子集,现有方法大多在模型训练中需要事先学习领域信息并预设给定参数值。实际应用中,由于不同的数据集数据结构和来源不同,在模型学习过程中研究人员无法提前获取相关领域知识且针对不同类型数据集指定一个统一参数存在巨大挑战。基于此,提出一种基于自适应密度邻域关系的多标签在线流特征选择方法(multi-label online stream feature selection based on adaptive density neighborhood relation,ML-OFS-ADNR),基于邻域粗糙集理论,所提方法在特征依赖计算时无需任何先验领域信息。此外,提出了一种新的自适应密度邻域关系,使用周围实例的密度信息,可以在流特征选择过程中自动选择适当数量的邻域,不需要事先指定任何参数。通过模糊等价约束,ML-OFS-ADNR 可以选择高依赖低冗余度的特征。实验表明在 10 种不同类型的数据集上,所提方法在特征数量相同的情况下优于传统特征选择方法和先进的在线流特征选择方法。

关键词:多标签分类;流特征;邻域粗糙集;自适应密度邻域;在线流特征选择
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2024)01-0023-07
doi:10.3969/j.issn.1673-629X.2024.01.004

Multi-label Online Stream Feature Selection Based on Adaptive Density Neighborhood Relation

ZHANG Hai-xiang[1],LI Pei-pei[2],HU Xue-gang[2]
(1. Information Division,The Second People´s Hospital of Hefei Affiliated to Bengbu Medical College,Hefei 230012,China;
2. Key Laboratory of Knowledge Engineering with Big Data of Ministry of Education,Hefei University of Technology,Hefei 230601,China)

Abstract:Stream feature selection selects the optimal feature subset from the feature data arriving in the form of stream. Most existing methods require prior learning of domain information and presetting of given parameter values during model training. In real-world applications,due to the differences in data structure and source,researchers cannot obtain relevant domain information in advance during the model learning process for different datasets,and it is a huge challenge for them to specify a unified parameter for different types of datasets. Motivated by this,we propose a multi-label online stream feature selection based on adaptive density neighborhood relation (ML-OFS-ADNR). On the basis of the neighborhood rough set theory,the proposed method does not require any prior domain information in feature dependency calculation. Moreover,a new adaptive density neighborhood relationship is proposed,which can automatically select an appropriate number of neighborhoods in the streaming feature selection process using the density information of surrounding instances,and there is no need to specify any parameters in advance. By the fuzzy equal constraint,ML-OFS-ADNR can select features with high dependency and low redundancy. Experimental studies on ten different types of data sets show that the proposed method is superior to traditional feature selection methods with the same numbers of features and state-of-the-art online streaming feature selection algorithms in an online manner.

Key words: multi-label classification; streaming feature; neighborhood rough set; adaptive density neighborhood; online streaming

收稿日期:2023-03-14 **修回日期**:2023-07-18
基金项目:国家自然科学基金资助项目(61976077,62076085,62120106008);蚌埠医学院科技计划项目(2022byzd225sk)
作者简介:张海翔(1996-),男,工程师,硕士,通信作者,研究方向为数据挖掘与人工智能;李培培(1982-),女,副研究员,硕导,博士,CCF 会员(18097M),研究方向为数据挖掘与人工智能;胡学钢(1961-),男,教授,博导,博士,CCF 会员(13977S),研究方向为数据挖掘与人工智能。

feature selection

0 引 言

多标签分类处理特征数据对应多个标签问题[1-4]。例如老年人群体就医时可能患有多种常见基础病:高血压、糖尿病、血脂异常等。药物治疗时遗漏病因或过敏史都将产生严重医疗事故。同时多标签数据通常伴有高维度、冗余特性,在分类过程耗费更多时间和空间,分类精准度也会受到影响[5]。因而,在多标签学习预处理阶段需要采取特征降维操作。

特征降维方法分为两种:特征提取和特征选择。特征提取方法[6]将原空间映射或变换低维空间,该过程会破坏数据特征的原有结构,不利于后续操作。特征选择方法[7-8]从原空间提取具有代表意义的特征子集,保留原特征空间结构,分类时可直观体现性能与特征的关系,常见方法如:过滤器、包装器和嵌入技术。过滤技术[9-11]独立于算法根据训练数据的一般特性(如互信息、粗糙集理论)选择合适特征,基于一组标准对特征进行评分删除评分较低的特征。包装器技术[12]利用一个特定的算法作为特征选择过程的组成部分其结果更有效,但该方法耗时较长。嵌入式技术[13]将特征选择过程集成到学习模型的构建中,需要迭代矩阵反演计算。以上方法考虑数据事先全部获取,现实数据并非一次性全部到来。例如,随时间推移和用药影响患者病情和生命体征在不断变化,传统离线批处理特征降维方法无法满足病情诊断精准判断[14]。

流特征选择方法随着时间推移,特征一个接一个地流入,必须在每个时间戳中决定保留还是丢弃新到达的特征,并且在学习之前不知道整个特征空间的信息,最终从已到来数据中选出代表特征子集。例如,利用模糊互信息作为相关性和冗余度评价指标的 OSFS(在线流特征选择)[15]方法,该方法包括两个步骤:在线相关性分析决定最新到来的特征是否保留或丢弃;在线冗余分析去除已选特征子集中的冗余特征。根据这两步分析原理,OM-NRS[16]用邻域粗糙集代替模糊互信息作为评价指标。I-SFS[17]基于互信息的相关性和冗余度,利用多目标布谷鸟搜索技术优化两个目标,但以上多标签流特征方法需要预先设定参数值,不同数据集上如何确定统一参数阈值存在一定困难。基于改进邻域粗糙集的多标签流特征选择方法(ML-OFS-ANRS[18])采用新的 GAP 邻域关系,在学习前不需要域信息,也不需要预先指定任何参数,采用最大依赖和最大重要性标准进行特征冗余分析,且该方法的重要性最优阈值参数对不同数据集需要设置不同值。

基于此,该文提出自适应密度邻域关系的多标签在线流特征选择方法(ML-OFS-ADNR),在依赖关系计算时根据周围实例的密度信息自动确定邻域个数,不需要预先指定任何参数。同时采用模糊等价约束方法可以挑选出高依赖度低冗余度的特征,使得所选特征子集规模较小且具有判别性。迭代执行以上步骤,直到数据全部到来。该文主要贡献如下:(1)基于邻域粗糙集理论,提出自适应密度邻域关系的多标签在线流特征选择方法(ML-OFS-ADNR),该方法在特征依赖计算时无需领域知识;(2)根据周围实例的密度信息提出新的邻域关系,利用该关系可以在在线特征选择过程中自动选择适当数量的邻域,不需要预先指定任何参数;(3)ML-OFS-ADNR 采用模糊等价约束进行冗余分析,使所选特征子集具有较低的冗余度。基于粗糙集的特征选择方法通常采用特征显著性等于零的条件进行特征冗余分析。然而,在真实的数据集中,完全相等的约束过于严格。在模糊等价约束下,可以考虑更多的候选特征进行特征冗余分析,使得最终选择的特征子集小且具有鉴别性。10 个基准数据集实验结果显示,ML-OFS-ADNR 在相同数量的特征下,比传统特征选择方法和现有在线流特征选择方法具有更好的性能。

1 相关工作

1.1 批处理特征选择方法

特征选择在处理高维数据上可以实现更快的模型训练,降低过拟合的敏感性,抵消维数灾难带来的影响,减少数据分析期间的存储、内存资源消耗。近十年来,研究人员提出许多离线环境下的降维方法,可分为特征提取[19-20]和特征选择[21]。前者是通过对原始特征进行线性或非线性组合来构造一些辅助特征,后者是从给定原始特征中选择最具区分能力的子集。根据标签分类应用场景又可分为单标签特征选择和多标签特征选择[22]。然而上述大多数方法考虑离线环境下,数据全部已知的批处理模式,在流特征环境中这些批处理方法无法直接应用,该环境下数据集无法一次性全部获取,无法提前获取全部数据信息。为增强学习模型的可解释性,便于在现实环境中广泛应用,研究人员更加重视特征选择方法,进一步从静态数据扩展到流数据环境[15]。

1.2 多标签流特征选择方法

流数据环境下特征会随着时间的推移到来(理论上无限制),在按顺序提取新特征的过程中还要剔除已到来特征数据的冗余,确保每一轮获得最优特征子集。已有多标签流特征选择方法包括 OMGFS[23]和

ML - OFS - ANRS[18], SFSCI[24], PSO[10], G - SFS[17]。OMGFS 基于邻域对称不确定性和邻域互信息,考虑在线特征固有群结构,根据不同数据集类型设置相关性参数阈值。ML-OFS-ANRS 提出新的 GAP 邻域在线流特征方法,采用最大依赖最大重要性进行特征冗余分析,对数据集环境要求较高,且重要性阈值最优参数根据不同数据集需要设置不同值。针对类不平衡数据环境,SFSCI 根据特征和标签之间依赖关系进行特征选择,模型学习前需提前确定最近邻参数值。PSO 提出一种三相滤波过程,在多目标优化设置中将进化粒子群优化技术应用新到来特征组,检查当前组中选择特征对已选择特征的冗余性,丢弃已选择特征列表中相对新到来特征而言不重要的特征。G-SFS 将多目标布谷鸟搜索技术交替应用新到来特征组,从 I-SFS 构建相应的 G-SFS。

传统的特征降维方法需要事先获取数据集的内容,在流特征环境下由于特征并非一次性全部到来,因而传统方法无法直接应用。已有的流特征方法大多需要预设相关参数,在实际应用中不同数据集的来源和空间结构不同,无法做到为每一个数据集预设最优统一参数值,且改进邻域粗糙集方法在特征冗余筛选过程中对真实数据集过于严格导致所选特征子集中存在冗余特征,在参数设置上也无法做到完全自适应。针对现有问题,基于模糊粗糙集理论,所提出的方法在学习之前不需要指定任何参数,并被证明在处理现实世界的数据集时是有效的。

2　在线流特征选择多标签分类方法

定义流特征选择 $OSFS=(U,C\cup D,t)$,其中 U 为非空有限数据集,C 为条件属性集,D 为决策属性集。$C=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}\in R^{n\times d}$,由 d 维特征空间上 $F=[f_1,f_2,\cdots,f_d]^{\mathrm{T}}\in R^d$ 的 n 个样本组成。$D=[y_1,y_2,\cdots,y_n]^{\mathrm{T}}\in R^{n\times l}$,由决策特征空间上 $L=\{l_1,l_2,\cdots,l_m\}$ 的 n 个样本组成。时间戳 t 时,到来新特征 f,学习映射函数 $h:x_i\rightarrow L(x_i\in C)$,得到具有代表意义的最佳特征子集。

2.1　基于自适应密度邻域的多标签特征依赖

邻域关系现有技术分为两种:距离固定(δ 邻域)或邻域数固定(k 最近邻域)。不同数据集数据分布不同,无法做到预设统一的参数。借鉴数据实例分布关系确定参数值作为一种新特征依赖方法受到关注。文中邻域关系确定由实例周围的密度信息自动确定邻域个数。

$N_B(x_i)=<x_i^1,x_i^2,\cdots,x_i^j,\cdots,x_i^{n-1}>$ 表示 x_i 的所有邻居,依照特征子集 B 距离从近到远排序,满足距离关系 $\Delta_B(x_i,x_i^1)\leqslant\Delta_B(x_i,x_i^2)\leqslant\cdots\leqslant\Delta_B(x_i,x_i^{n-1})$。定义 x_i 到邻居 x_i^k 的密度 $\mathrm{Density}(x_i,x_i^k)=\dfrac{\Delta_B(x_i,x_i^k)}{k}$,从 x_i 到 x_i^{n-1} 假设密度值首先在 x_i^k 处减小,称 x_i^k 为拐点(Inflection Point)用 $\mathrm{IP}(x_i^k)$ 表示,使用 x_i 和拐点 x_i^k 间的样本作为 x_i 的最近邻居。任意对象 x_i 在特征子集 B 上的自适应邻域定义为:

$$\mathrm{IP}_B(x_i)=\{x_i^1,x_i^2,\cdots,x_i^{k-1}\}\tag{1}$$

基于这种新的密度邻域关系,在流特征多标签分类环境下提出新的依赖关系计算方法。由于不同标签种类会对特征筛选产生影响,所以在计算过程中考虑每一个标签下的特征依赖度值。新到来数据 f_t,计算 f_t 在 $D=[y_1,y_2,\cdots,y_n]^T\in R^{n\times l}$ 的依赖值,第一步,计算 f_t 中各个数据间的距离 $\Delta_t(x_i,x_i^k)(2\leqslant k\leqslant n)$ 并按照由近到远排序;第二步,根据密度邻域关系找出 f_t 中每个样本的邻居 $\{x_i^1,x_i^2,\cdots,x_i^k\}$;第三步,计算在所有标签 y_l 下 x_i 与密度邻域间的平均依赖值 $S_R^l(x_i)_{\mathrm{card}}$,其中 $S_R^l(x_i)_{\mathrm{card}}$ 表示 x_i 与密度邻居 $\{x_i^1,x_i^2,\cdots,x_i^k\}$ 在标签为 y_l 位置上各自对应的标签 $y_{x_i}^l$ 和 $y_{x_i^j}^l$ 相同的总数占总密度邻居数的比值,表示如下:

$$S_R^l(x_i)_{\mathrm{card}}=\frac{\sum_{j=1}^{k}[\![y_{x_i}^l==y_{x_i^j}^l]\!]}{k}\tag{2}$$

其中,k 为 x_i 密度邻居个数,$[\![y_{x_i}^l==y_{x_i^j}^l]\!]$ 表示在标签 y_l 下标签 $y_{x_i}^l$ 和 $y_{x_i^j}^l$ 相同时值为 1,不同为 0。第四步,根据 $S_R^l(x_i)_{\mathrm{card}}$ 求出 f_t 在标签 y_l 下的特征依赖值总和 Dep_t^l;最后重复三、四步得到每一个标签下 f_t 的依赖值 $\mathrm{Dep}_t=\{\mathrm{Dep}_t^{y_1},\mathrm{Dep}_t^{y_2},\cdots,\mathrm{Dep}_t^{y_l}\}$。

2.2　自适应密度邻域粗糙集的在线特征选择

该文在密度邻域粗糙集[25]基础上提出自适应密度粗糙集多标签流特征选择,实现对新到来的特征选择保留或抛弃,选出代表性的特征子集。新到来数据 f_t,计算多标签空间 D 下的依赖值 Dep_t,自适应设置特征依赖筛选参数,选出依赖值大的特征。由 $t-1$ 时刻 f_{t-1} 依赖值 $\mathrm{Dep}_{t-1}=\{\mathrm{Dep}_{t-1}^{y_1},\mathrm{Dep}_{t-1}^{y_2},\cdots,\mathrm{Dep}_{t-1}^{y_l}\}$ 得出平均依赖阈值 $\mathrm{Ave\_Dep}_{t-1}$(见式 3)作为下一时刻特征 f_t 筛选条件,初始 $\mathrm{Ave\_Dep}_{t-1}=0$。

$$\mathrm{Ave\_Dep}_{t-1}=\frac{\sum_{i=1}^{l}\mathrm{Dep}_{t-1}^i}{|l|},\quad t\geqslant 2\tag{3}$$

实现新到来特征筛选后,如果直接将其加入已筛选特征子集 F_t,容易产生较多冗余特征,导致模型性能降低,因而新特征筛选后还需对特征子集进行冗余优化。对新到来数据 f_t,根据依赖计算度计算方法得到依赖值 Dep_t。其次比较 $\mathrm{Ave\_Dep}_{t-1}$ 与 Dep_t,如果 Dep_t 中均小于 $\mathrm{Ave\_Dep}_{t-1}$,则认为 f_t 较已选特征子集 F_{t-1} 依赖性较低,丢弃。若满足,再比较已选特征集

F_{t-1} 与合并特征集 $F_{t-1} \cup f_t$ 的依赖关系，当 $\text{Ave\_Dep}_{F_{t-1}\cup f_t}$ 大于 Ave_Dep_{t-1}，意味着添加新特征 f_t 会增加已选特征集 F_{t-1} 的依赖性，那么将 f_t 加入 F_{t-1} 中，否则判断 Ave_Dep_{t-1} 与 $\text{Ave\_Dep}_{F_{t-1}\cup f_t}$ 之差对 Ave_Dep_{t-1} 的比值分析特征冗余性。对特征集 $F_{t-1} \cup f_t$ 中的每一特征从候选特征集合中随机选择一个特征 f，计算 f 显著性值（即 $\text{Ave\_Dep}_{F_{t-1}\cup f_t}$ 与 $\text{Ave\_Dep}_{(F_{t-1}\cup f_t)-f}$ 的差值），将显著性等于 0 的特征丢弃。通过这种新的在线流特征选择算法，可以选择高相关性、高依赖性和低冗余度的特征。

3 实验及其结果分析

3.1 实验数据集与评价指标

本节给出在实验数据集上所提方法的实验结果优势，其中选取的数据集均为常见多标签分类方法实验数据集，数据来源于 Mulan（http://mulan.sourceforge.net/datasets.html）和 Meka（https://waikato.github.io/meka/datasets/），详细信息见网站内容介绍。表 1 给出了 10 个实验数据集介绍，包括：样本数、特征数、标签数、数据领域。数据领域包括网页文本、电子邮件、音乐及基因功能，例如电子邮件 Enron 数据集共有 1 702个数据，每条数据至少从属 53 种标签的一种或多种。

表 1 数据集

Datasets	Instances	Features	Labels	Domain
Arts	5 000	462	26	Text
Birds	645	260	19	Audio
Education	5 000	550	33	Text
Emotions	593	72	6	Music
Enron	1 702	1 001	53	Text
Reference	5 000	793	33	Text
Scene	2 407	294	6	Image
Science	5 000	743	40	Text
Social	5 000	1 047	39	Text
Yeast	2 417	103	14	Biology

将所提方法与 5 个批处理方法和 2 个流特征选择方法进行对比。其中 MDDMspc[6] 通过判别特征与标签间依赖最大化实现降维。GLOCAL[26] 通过学习潜在标签和标签流形化实现降维并考虑标签的全局、局部相关性。LLSF[27] 学习各标签的特定数据实现多标签特征选择。LSML[28] 解决在缺失标签环境下的多标签特征选择问题。MCLS[29] 通过将原始逻辑标签转化数字标签实现实例相似度约束。SFSCI[24] 处理在类标签不平衡环境下流特征选择。ML-OFS-ANRS[18] 提出自适应邻域粗糙集流特征选择方法，利用最大依赖、最大重要指标筛选特征子集。所提方法模糊等价约束参数按照相关参数设置 0.05[25]，其余无需任何参数，使用 MLKNN 分类模型评估算法性能。

3.2 实验结果分析

3.2.1 所提方法与批处理方法对比

各实验指标含义如下：给定测试数据集 $T = \{(x_i, Y_i)\}_{i=1}^{n}$，测试样本 x_i，$Y_i \in \{0,1\}^l$ 表示标签集合，$h(x_i)$ 为样本 x_i 预测结果，$f(x_i, y)$ 表示 x_i 属于 y 的置信度。Average Precision 评估排名高于特定标签 $y \in y_i$ 的相关标签平均分数：

$$\text{AvePrec} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i^l|}\sum_{y_j\in Y_i^l}\frac{|\{y_q \in Y_i^l : R_i(y_q) \leqslant R_i(y_j)\}|}{R_i(y_j)} \tag{4}$$

其中，$R_i(y_j)$ 是样本 x_i 的标签 y_j 预测等级。

Ranking Loss 描述样本标签对被反向排序的平均比例：

$$\text{RL} = \frac{1}{n}\sum_{i=1}^{n}\frac{|(y_a, y_b) : R_i(y_a) > R_i(y_b), (y_a, y_b) \in Y_i^l \times Y_i^l|}{|Y_i^l||Y_i^l|} \tag{5}$$

Coverage 计算平均所需的步骤数，以向下推进标签列表并覆盖所有实例的适当标签：

$$\text{Coverage} = \frac{1}{n}\sum_{i=1}^{n}\max_{y\in y_i}\text{rank}_f(x_i, y) - 1 \tag{6}$$

One-error 计算一个不相关的标签被排在首位的次数：

$$\text{One-error} = \frac{1}{n}\sum_{i=1}^{n}\{[\arg\max_{y\in Y_i} f(x_i, y)] \notin Y_i\} \tag{7}$$

Hamming Loss 计算错误分类标签的平均得分：

$$\text{hLoss}(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{l}|h(x_i)\Delta Y_i| \tag{8}$$

Δ 表示两组间对称差。

为了更加直观体现各方法对比，进行 Friedman 检验[30]，表 2 给出各个实验指标临界值的 Friedman statistics F_F。根据 Nemenyi 检测方法对比所提方法是否优势显著，所提方法为控制算法，将算法间平均排名的差异与 CD 值进行比较，$\text{CD} = q_a\sqrt{\frac{k(k+1)}{6N}}$，且 CD = 2.848（$k=7, N=10$）时 $q_a = 2.948$，各实验指标下的 CD 结果见图 1。如果所提方法（ML-OFS-ADNR）与对比算法的平均排名相差 CD 且算法之间没有联系，认为两算法的性能差异显著。由图 1 可知，ML-OFS-

ADNR 的实验结果在各排名中均为首位,验证密度邻域粗糙集流特征选择方法的有效性,采用模糊等价约束进行冗余分析,可以较好保持特征子集具有低冗余、高依赖特性。

表 2　在 0.05 显著性水平条件下 F_F 每种评估方法的临界值

Metric	F_F	Critical value(α =0.05)
Average Precision	13.661 9	2.272
Hamming Loss	15.477 9	
One-Error	12.096 7	
Ranking Loss	16.544 9	
Coverage	10.733 8	

在 5 个评价指标上均显著优于 MDDMspc 和 MCLS,原因在于 MDDM 在标签处理上忽略了标签之间潜在的关联关系,所提方法引入密度邻域关系且利用平均依赖度最大限度考虑标签之间的内在联系,而 MCLS 在特征选择过程未对特征冗余进行过滤,取得的特征子集非最优,而 ML-OFS-ADNR 不仅对单个特征进行依赖度分析,而且分析新特征加入后是否对原有特征产生冗余。在线筛选过滤得到最优特征子集。

与 LLSF,GLOCAL 和 LSML 的对比中,虽然没有在所有指标上取得全部优势,但是无论在具体数据集还是 CD 图上综合排名所提方法最好。由于在密度邻域依赖度计算过程中仅考虑在密度信息范围内忽略密度范围以外的标签信息,无法最大化利用标签空间相关性信息,所提方法的 CD 非显著最优。

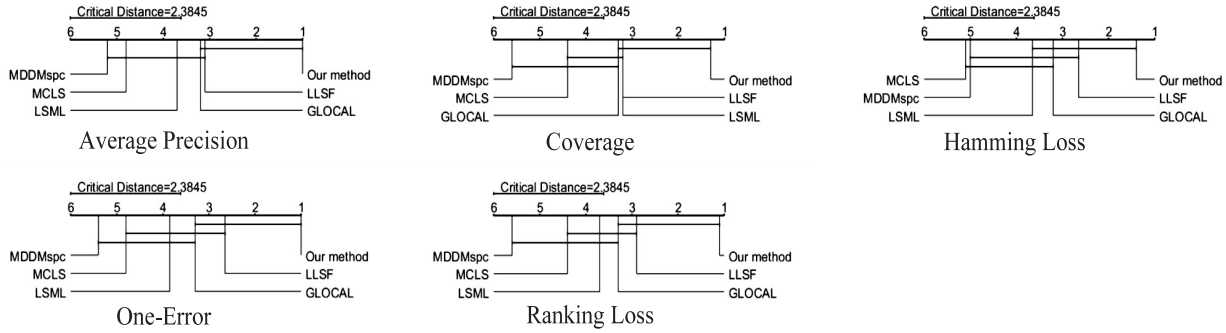

图 1　所提方法与对比算法的 Nemenyi 检验比较结果

3.2.2　所提方法与流特征选择方法对比

除与批处理方法对比外,还在多标签特征流环境下与最新的 SFSCI[24], ML-OFS-ANRS[18] 在指标 Average Precision 和 Ranking Loss 上进行比较,实验结果见表 3。

表 3　ML-OFS-ADNR 与 SFSCI,ML-OFS-ANRS 对比结果

Datasets	Average Precision(↑)			Ranking Loss(↓)		
	ML-OFS-ADNR	ML-OFS-ANRS	SFSCI	ML-OFS-ADNR	ML-OFS-ANRS	SFSCI
Arts	0.545	0.529	0.532 5	0.138 8	0.149 5	0.146 2
Birds	0.721	0.715 1	0.712 7	0.116 9	0.122 1	0.120 6
Education	0.575 4	0.561 5	0.567 3	0.088 4	0.091 2	0.088 6
Emotions	0.764 4	0.774 2	0.769 5	0.205	0.184 6	0.198 7
Enron	0.677 6	0.672 3	0.670 2	0.085 3	0.087 2	0.088 5
Reference	0.650 1	0.641	0.638 9	0.084 1	0.092 8	0.083 5
Scene	0.819 5	0.827 3	0.817 5	0.105 4	0.116 3	0.110 2
Science	0.490 2	0.482 6	0.497 4	0.130 6	0.137 1	0.128 1
Social	0.733 3	0.728 7	0.727 2	0.061 5	0.060 3	0.062 9
Yeast	0.768 1	0.761	0.756 5	0.170 7	0.169 3	0.174 1

根据表 3 可知:(1)基于密度邻域粗糙集无需预设参数,根据密度信息自动确定邻域个数而 SFSCI 在实验前需要手动设置最大邻居数 K,表明所提方法在 Average Precision 明显优于 SFSCI,Ranking Loss 上也有超过一半数据结果优势。(2)在其他数据集上,所提方法均优于 SCSFI,原因在于密度邻域关系可以根据数据集的不同类型自动选择合适的邻域个数,而 SCSFI 的 K 值设定忽略数据集内容,限制模型的可扩展性,在数据集上无法发挥全部优势。(3)与 ML-OFS-ANRS 对比,两种算法在邻域处理均采取自适应

的方法,保证了算法在不同数据都可最大化适应,在特征冗余筛选处理中,ML-OFS-ANRS 依赖最大重要性标准,且针对不同数据集需要找出相应最优重要性阈值参数值,而所提方法采用模糊等价约束,模糊等价约束参数统一且特征冗余性参数设置依赖前一个已选特征子集数据,实现了自适应变化。(4)相比所提方法采用模糊等价约束进行特征冗余分析,ML-OFS-ANRS 采用最大依赖和最大重要性标准进行特征冗余分析选择了更多的特征,造成一定的冗余。而在模糊等价约束下,可以考虑更多的候选特征进行特征冗余分析,使得最终所选特征子集最优。(5)图 2 给出实验过程中随着特征子集不断被筛选出来在数据集上的分类性能逐步提升。为保证实验效果图美观,仅展示了在 Average Precision 和 Ranking Loss 评价指标上的结果。可以看出在 Birds 和 Yeast 数据集上,所提方法分别筛选的特征子集数目达到 50 和 35 个左右时,表现出的分类性能趋于稳定。同特征子集数量下,所提方法在实验精度上优于另外两个对比算法。当其他算法特征子集筛选结果趋于稳定完成筛选,从实验精度上所提方法依旧领先。

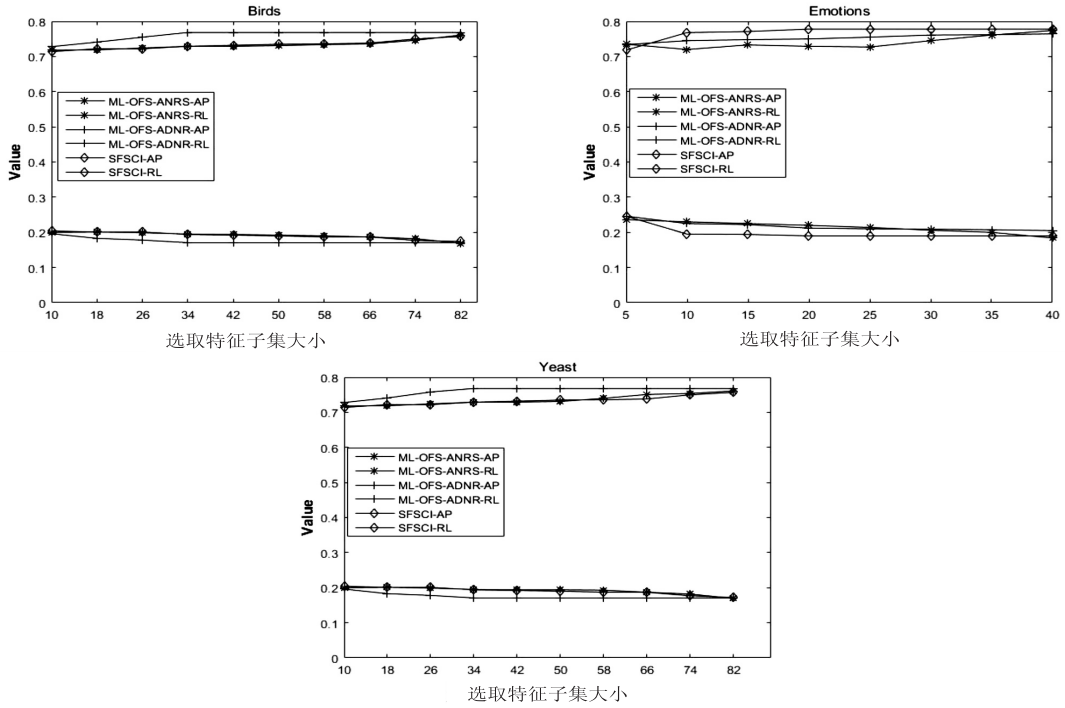

图 2 在 3 个代表数据集上调整特征子集所得到的实验结果

4 结束语

提出一种基于密度邻域粗糙集的多标签流特征选择方法(ML-OFS-ADNR)。该方法参照单标签密度邻域关系理论在多标签分类环境中予以拓展,在依赖计算时根据周围实例的密度信息自动决定邻域数量,不需要预先指定任何参数。同时,采用模糊等价约束进行冗余分析,使所选特征子集小且具有鉴别性。与 5 种传统特征选择和 2 种流特征选择算法相比,该方法在相同特征数量的情况下优于传统特征选择方法,并在在线方式下优于流特征选择算法。未来的工作中,考虑实现模糊等价约束参数阈值自适应变化和模糊粗糙集理论应用流特征选择,实现更加精准、自适应无参化的流特征方法研究。

参考文献:

[1] 武红鑫,韩 萌,陈志强,等. 监督和半监督学习下的多标签分类综述[J]. 计算机科学,2022,49(8):12-25.

[2] 曾艺祥,林耀进,范凯钧,等. 基于层次类别邻域粗糙集的在线流特征选择算法[J]. 南京大学学报:自然科学版,2022,58(3):506-518.

[3] 周慧颖,汪廷华,张代俐. 多标签特征选择研究进展[J]. 计算机工程与应用,2022,58(15):52-67.

[4] FAN Y,LIU J,WENG W,et al. Multi-label feature selection with local discriminant model and label correlations[J]. Neurocomputing,2021,442:98-115.

[5] WANG K. Robust cross-view embedding with discriminant structure for multi-label classification[J]. IEEE Access,2021,9:117596-117607.

[6] ZHANG Y,ZHOU Z H. Multi-label dimensionality reduction via dependence maximization[J]. ACM Transactions on Knowledge Discovery from Data,2010,4(3):1-21.

[7] FAN Y,CHEN B,HUANG W,et al. Multi-label feature selection based on label correlations and feature redundancy[J]. Knowledge-Based Systems,2022,241:108256.

[8] LI Y,HU L,GAO W. Label correlations variation for robust

multi - label feature selection [J]. Information Sciences, 2022,609:1075-1097.

[9] PENG L, WU Y, HUANG L. Opto-electric target tracking algorithm based on local feature selection and particle filter optimization[J]. Concurrency and Computation:Practice and Experience,2018,30(22):e4670.

[10] PAUL D, JAIN A, SAHA S, et al. Multi-objective PSO based online feature selection for multi-label classification [J]. Knowledge-Based Systems,2021,222:106966.

[11] WAN J,CHEN H,YUAN Z,et al. A novel hybrid feature selection method considering feature interaction in neighborhood rough set[J]. Knowledge-Based Systems,2021,227:107167.

[12] NOURI-MOGHADDAM B, GHAZANFARI M, FATHIAN M. A novel multi-objective forest optimization algorithm for wrapper feature selection[J]. Expert Systems with Applications,2021,175:114737.

[13] CHEN C W,TSAI Y H,CHANG F R,et al. Ensemble feature selection in medical datasets:combining filter,wrapper, and embedded feature selection results[J]. Expert Systems, 2020,37(5):e12553.

[14] XUE X,YAO M,WU Z. A novel ensemble-based wrapper method for feature selection using extreme learning machine and genetic algorithm[J]. Knowledge and Information Systems,2018,57:389-412.

[15] LIN Y,HU Q,LIU J,et al. Streaming feature selection for multilabel learning based on fuzzy mutual information[J]. IEEE Transactions on Fuzzy Systems,2017,25(6):1491-1507.

[16] LIU J,LIN Y,LI Y,et al. Online multi-label streaming feature selection based on neighborhood rough set[J]. Pattern Recognition,2018,84:273-287.

[17] PAUL D,KUMAR R,SAHA S,et al. Multi-objective cuckoo search-based streaming feature selection for multi-label dataset [J]. ACM Transactions on Knowledge Discovery from Data,2021,15(6):1-24.

[18] 张海翔,李培培,胡学钢. 基于自适应领域粗糙集的多标签在线流特征选择[J]. 微电子学与计算机,2022,39(7):44-53.

[19] SIBLINI W,KUNTZ P,MEYER F. A review on dimensionality reduction for multi-label classification[J]. IEEE Transactions on Knowledge and Data Engineering,2019,33(3):839-857.

[20] XU J,MAO Z H. Multilabel feature extraction algorithm via maximizing approximated and symmetrized normalized cross-covariance operator[J]. IEEE Transactions on Cybernetics, 2019,51(7):3510-3523.

[21] KASHEF S,NEZAMABADI-POUR H,NIKPOUR B. Multilabel feature selection:a comprehensive review and guiding experiments[J]. Wiley Interdisciplinary Reviews:Data Mining and Knowledge Discovery,2018,8(2):e1240.

[22] QIAN Y,WANG Q,CHENG H,et al. Fuzzy-rough feature selection accelerator[J]. Fuzzy Sets and Systems,2015,258:61-78.

[23] LIU J,LIN Y,WU S,et al. Online multi-label group feature selection[J]. Knowledge-Based Systems,2018,143:42-57.

[24] ZOU Y,HU X,LI P,et al. Multi-label streaming feature selection via class-imbalance aware rough set[C]//2021 international joint conference on neural networks (IJCNN). Shenzhen:IEEE,2021:1-9.

[25] ZHOU P, HU X, LI P, et al. OFS-density:a novel online streaming feature selection method[J]. Pattern Recognition, 2019,86:48-61.

[26] ZHU Y,KWOK J T,ZHOU Z H. Multi-label learning with global and local label correlation[J]. IEEE Transactions on Knowledge and Data Engineering,2017,30(6):1081-1094.

[27] HUANG J,LI G,HUANG Q,et al. Learning label-specific features and class-dependent labels for multi-label classification[J]. IEEE Transactions on Knowledge and Data Engineering,2016,28(12):3309-3323.

[28] HUANG J,QIN F,ZHENG X,et al. Improving multi-label classification with missing labels by learning label-specific features[J]. Information Sciences,2019,492:124-146.

[29] HUANG R, JIANG W, SUN G. Manifold-based constraint Laplacian score for multi-label feature selection[J]. Pattern Recognition Letters,2018,112:346-352.

[30] DEMŠAR J. Statistical comparisons of classifiers over multiple data sets [J]. The Journal of Machine Learning Research,2006,7:1-30.