Journal of Computer Applications
计算机应用,2020,40(2):503 - 509
ISSN 1001-9081
CODEN JYIIDU
2020-02-10
http://www.joca.cn

文章编号:1001-9081(2020)02-0503-07
DOI:10.11772/j.issn.1001-9081.2019091626

基于专家特征的条件互信息多标记特征选择算法

程玉胜[1,2*],宋　帆[1],王一宾[1,2],钱　坤[1]
(1.安庆师范大学 计算机与信息学院,安徽 安庆 246011;　2.安徽省高校智能感知与计算重点实验室,安徽 安庆 246011)
(*通信作者电子邮箱 chengyshaq@163.com)

摘　要:特征选择对于分类器的分类精度和泛化性能起重要作用。目前的多标记特征选择算法主要利用最大相关性最小冗余性准则在全部特征集中进行特征选择,没有考虑专家特征,因此多标记特征选择算法的运行时间较长、复杂度较高。实际上,在现实生活中专家依据几个或者多个关键特征就能够直接决定整体的预测方向。如果提取关注这些信息,必将减少特征选择的计算时间,甚至提升分类器性能。基于此,提出一种基于专家特征的条件互信息多标记特征选择算法。首先将专家特征与剩余的特征相联合,再利用条件互信息得出一个与标记集合相关性由强到弱的特征序列,最后通过划分子空间去除冗余性较大的特征。该算法在7个多标记数据集上进行了实验对比,结果表明该算法较其他特征选择算法有一定优势,统计假设检验与稳定性分析进一步证明了所提出算法的有效性和合理性。

关键词:特征选择;专家特征;条件互信息;多标记学习;局部子空间
中图分类号:TP391　　**文献标志码**:A

Multi-label feature selection algorithm based on conditional mutual information of expert feature

CHENG Yusheng[1,2*], SONG Fan[1], WANG Yibin[1,2], QIAN Kun[1]
(1. *School of Computer and Information, Anqing Normal University, Anqing Anhui* 246011, *China*;
2. *University Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing Anhui* 246011, *China*)

Abstract: Feature selection plays an important role in the classification accuracy and generalization performance of classifiers. The existing multi-label feature selection algorithms mainly use the maximum relevance and minimum redundancy criterion to perform feature selection in all feature sets without considering expert features, therefore, the multi-label feature selection algorithm has the disadvantages of long running time and high complexity. Actually, in real life, experts can directly determine the overall prediction direction based on a few or several key features. Paying attention to and extracting this information will inevitably reduce the calculation time of feature selection and even improve the performance of classifier. Based on this, a multi-label feature selection algorithm based on conditional mutual information of expert feature was proposed. Firstly, the expert features were combined with the remaining features, and then the conditional mutual information was used to obtain a feature sequence of strong to weak relativity with the label set. Finally, the subspaces were divided to remove the redundant features. The experimental comparison was performed to the proposed algorithm on 7 multi-label datasets. Experimental results show that the proposed algorithm has certain advantages over the other feature selection algorithms, and the statistical hypothesis testing and the stability analysis further illustrate the effectiveness and the rationality of the proposed algorithm.

Key words: feature selection; expert feature; conditional mutual information; multi-label learning; local subspace

0　引言

多标记学习[1]作为机器学习等领域的研究热点之一,较之传统的单标记学习中的一个对象只能局限于单个标记,多标记学习框架更具有实际性和广泛性。在真实世界中,一个对象可能隶属于多个标记[2],此时,单个标记就难以表述对象的完整性。所以,多标记学习的研究对于多义性对象的学习建模更具有实际运用意义,近年来已成为一个新的研究热点[3-4]。

在多标记学习问题中,由于数据的高维性会引起维数灾难,导致分类器精度降低[5]。而特征选择作为一种普遍的降维手段,对于分类器的分类精度和泛化性能起重要作用。特征选择的首要目的是在样本数据集中找到一个特征子集,且使得找到的特征子集蕴含尽可能多的区分类别信息,同时要考虑子集内部的冗余性尽量小[6]。而信息论中的互信息理论,作为不确定性的一种有效度量方式,被广泛用于多标记特征选择,因此许多学者在此方面进行了研究。例如Zhang等[7]

收稿日期:2019-08-30;**修回日期**:2019-09-24;**录用日期**:2019-10-09。
基金项目:安徽省高校重点科研项目(KJ2017A352);安庆师范大学科研创新团队建设计划项目。
作者简介:程玉胜(1969—),男,安徽安庆人,教授,博士,CCF会员,主要研究方向:大数据、粗糙集、特征选择的机器学习;　宋帆(1992—),男,安徽铜陵人,硕士研究生,CCF会员,主要研究方向:多标记学习、神经网络;　王一宾(1970—),男,安徽安庆人,教授,硕士,主要研究方向:多标记学习、机器学习、软件安全;　钱坤(1995—),男,安徽滁州人,硕士研究生,CCF会员,主要研究方向:多标记学习、机器学习、数据统计。

提出的基于最大相关性的多标记维度约简(Multi-label Dimensionality reduction via Dependence Maximization, MDDM)算法。Lee等[8]通过多元互信息最大化已选特征与标记集合的相关性,提出了基于多变量互信息的多标记特征选择算法PMU(Pairwise Multivariate mutual information)。Lin等[9]提出了基于邻域互信息的多标记特征选择。刘景华等[10]通过互信息排序已选特征和标记的相关性,提出了基于局部子空间的多标记特征选择算法(Multi-label Feature Selection algorithm based on Local Subspace, MFSLS)。

上述算法的多标记特征选择算法判定特征是否冗余的标准单一,如信息熵方法仅考虑特征和标记间的相关性,未考虑特征和特征间的关系[11];联合互信息虽然考虑了整体互信息大小,但未考虑单个特征和标记间的相关性。这些算法都没有提取关注专家特征,在整个特征集中进行特征选择,因此时间复杂度很高,如PMU算法在大数据集中执行时间很长,基于信息熵的多标签特征选择(Multi-Label Feature Selection based on Information Entropy, MLFSIE)算法虽然提高了执行速度,但未考虑特征间的冗余性。

除此之外,包括上述算法在内的大多数多标记特征选择算法均未考虑到优先挑选出专家特征的现实意义,忽略了一个关键的问题:在现实生活中,人们针对分类问题时,通常根据专家经验选取几个或者多个最重要的特征,然后再通过相关评价准则建立特征向任务目标的映射进行多标记分类。例如,在医院中专家医师看病人的病情时,往往根据自己的多年临床经验先确定几个最重要的病症(即对结果不可或缺起到重要作用的特征(专家特征)),然后再在专家特征的基础上进行各种身体检查、血液化验、分析病历,最后分析汇总来确诊。同时也要考虑某些看似不显眼的症状(即与标记空间相关性较弱的特征),因为忽略某些重要性较次要的特征也会产生误诊的可能。

基于此,再结合信息论中的互信息[12-14],本文提出一种基于专家特征的条件互信息多标记特征选择算法(Multi-label Feature Selection algorithm based on Conditional Mutual Information of Expert Feature, MFSEF)。该算法在最小冗余最大相关性前提下,通过子空间划分,考虑了重要性较次要的特征可能对分类性能产生的影响,受现实生活中实际问题的启发,兼顾考虑了可能决定整体的预测方向的专家特征,从而提升多标记分类性能。首先通过瀑布图联合互信息选出几个最关键的专家特征;再以该专家特征作条件,保持专家特征不变,与余下的特征作并集,构建融合一个新的特征空间,然后计算新特征与标记集合之间的互信息,再进行排序形成新的特征排序集合;借鉴MFSLS的思想,最后进行特征选择。实验在7个多标记数据集上测试,同其他常用的多标记特征选择算法进行比较,通过4个评价指标的结果可以看出,本文算法优于通用的多标记特征选择算法。最后,还通过统计假设检验进一步证实了本文方法的合理性与稳定性。

1 理论介绍

1.1 多标记学习

由于真实世界的对象具有多义性,多标记学习框架作为一种多义性对象学习建模工具由此产生[15]。在该框架下,每个对象由一个示例描述,每个示例具有多个但有限的类别标记,学习的目的是为每个未知示例赋予正确的标记。在数学语言中,多标记问题可描述为:假定$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}^{\mathrm{T}}\in\mathbf{R}^{n*d}$表示有$n$个样本且每个样本特征维度为$d$,$\boldsymbol{Y}=\{1,2,\cdots,Q\}$表示样本对应的标记集合[16]。$T=\{(\boldsymbol{x}_1,\boldsymbol{Y}_1),(\boldsymbol{x}_2,\boldsymbol{Y}_2),\cdots,(\boldsymbol{x}_m,\boldsymbol{Y}_m)\}(\boldsymbol{x}_i\in\boldsymbol{X},\boldsymbol{Y}_i\in\boldsymbol{Y})$表示训练集,多标记学习目的是得到映射关系$f:\boldsymbol{X}\to\{-1,1\}^Q$,从而对新样本进行标记的预测。

1.2 条件互信息

定义1 设集合$\boldsymbol{A}=\{a_1,a_2,\cdots,a_m\}$,令$p(a_i)$表示元素$a_i$的先验概率,则集合$\boldsymbol{A}$的信息熵为:

$$\boldsymbol{H}(\boldsymbol{A})=-p(a_i)\mathrm{lb}\,p(a_i) \tag{1}$$

信息熵可以度量集合不确定性的程度,信息熵越大表示集合的不稳定性越大。对于多标记特征选择算法,常通过信息熵来选择特征空间中与标记空间互信息较大的特征。

定义2 设集合$\boldsymbol{A}=\{a_1,a_2,\cdots,a_m\}$,$\boldsymbol{B}=\{b_1,b_2,\cdots,b_n\}$,则在给定集合$\boldsymbol{A}$的条件下集合$\boldsymbol{B}$的条件熵为:

$$\boldsymbol{H}(\boldsymbol{B}|\boldsymbol{A})=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(a_i,b_j)\mathrm{lb}\,p(b_j|\ a_i) \tag{2}$$

条件熵可以度量在集合$\boldsymbol{A}$出现的条件下集合$\boldsymbol{B}$的不确定程度的大小。

定义3 设集合$\boldsymbol{A}=\{a_1,a_2,\cdots,a_m\}$,集合$\boldsymbol{B}=\{b_1,b_2,\cdots,b_n\}$,则集合$\boldsymbol{A}$与$\boldsymbol{B}$的联合熵为:

$$\boldsymbol{H}(\boldsymbol{A},\boldsymbol{B})=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(a_i,b_j)\mathrm{lb}\,p(a_i,b_j) \tag{3}$$

信息熵、条件熵及联合熵的关系为:

$$\boldsymbol{H}(\boldsymbol{A},\boldsymbol{B})=\boldsymbol{H}(\boldsymbol{A})+\boldsymbol{H}(\boldsymbol{B}|\boldsymbol{A})=\boldsymbol{H}(\boldsymbol{B})+\boldsymbol{H}(\boldsymbol{A}|\boldsymbol{B}) \tag{4}$$

定义4 给定集合$\boldsymbol{A}$和集合$\boldsymbol{B}$,定义集合$\boldsymbol{A}$和$\boldsymbol{B}$之间的互信息为:

$$\boldsymbol{I}(\boldsymbol{A};\ \boldsymbol{B})=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(a_i,b_j)\mathrm{lb}\left[\frac{p(a_i,b_j)}{p(a_i)p(b_j)}\right] \tag{5}$$

互信息被广泛用于度量随机变量间相关性的大小,即$\boldsymbol{I}(\boldsymbol{A};\ \boldsymbol{B})$表示集合$\boldsymbol{A}$和集合$\boldsymbol{B}$间的相关性大小。$\boldsymbol{I}(\boldsymbol{A};\ \boldsymbol{B})$越大,表示两者间的相关性越大。另有$\boldsymbol{I}(\boldsymbol{A};\ \boldsymbol{B})=\boldsymbol{I}(\boldsymbol{B};\ \boldsymbol{A})$,且满足:

$$\begin{aligned}\boldsymbol{I}(\boldsymbol{A};\ \boldsymbol{B})&=\boldsymbol{H}(\boldsymbol{A})+\boldsymbol{H}(\boldsymbol{B})-\boldsymbol{H}(\boldsymbol{A},\boldsymbol{B})=\\&\boldsymbol{H}(\boldsymbol{A})-\boldsymbol{H}(\boldsymbol{A}|\boldsymbol{B})=\boldsymbol{H}(\boldsymbol{B})-\boldsymbol{H}(\boldsymbol{B}|\boldsymbol{A})\end{aligned} \tag{6}$$

当$\boldsymbol{I}(\boldsymbol{A};\ \boldsymbol{B})=0$时,集合$\boldsymbol{A}$和集合$\boldsymbol{B}$无相关性,集合$\boldsymbol{A}$和集合$\boldsymbol{B}$之间未提供任何信息。

定义5 设集合$\boldsymbol{A}=\{a_1,a_2,\cdots,a_m\}$,$\boldsymbol{B}=\{b_1,b_2,\cdots,b_n\}$,$\boldsymbol{C}=\{c_1,c_2,\cdots,c_t\}$,则在集合$\boldsymbol{C}$条件下集合$\boldsymbol{A}$和$\boldsymbol{B}$间的条件互信息[17]为:

$$\boldsymbol{I}(\boldsymbol{A};\ \boldsymbol{B}|\boldsymbol{C})=\boldsymbol{H}(\boldsymbol{A}|\boldsymbol{B})-\boldsymbol{H}(\boldsymbol{A}|\boldsymbol{B},\boldsymbol{C}) \tag{7}$$

联合互信息可由式(6)和式(7)得出:

$$\boldsymbol{I}(\boldsymbol{A},\boldsymbol{C};\ \boldsymbol{B})=\boldsymbol{I}(\boldsymbol{A};\ \boldsymbol{B}|\boldsymbol{C})+\boldsymbol{I}(\boldsymbol{B};\ \boldsymbol{C}) \tag{8}$$

联合互信息是考虑$\boldsymbol{A}$、$\boldsymbol{C}$整体同$\boldsymbol{B}$之间的关系,由上式可知条件互信息和互信息之和为联合互信息,根据式(5)得出联合互信息为:

$$\boldsymbol{I}(\boldsymbol{A},\boldsymbol{C};\ \boldsymbol{B})=-\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{t}p(a_i,b_j,c_k)\mathrm{lb}\left[\frac{p(a_i,b_j,c_k)}{p(a_i,c_k)p(b_j)}\right] \tag{9}$$

联合互信息$\boldsymbol{I}(\boldsymbol{A},\boldsymbol{C};\ \boldsymbol{B})$越大,则表示$\boldsymbol{A}$、$\boldsymbol{C}$整体同$\boldsymbol{B}$间的

相关性越强。另外关于条件互信息还可变形表示如下:

$$\boldsymbol{I}(\boldsymbol{A};\ \boldsymbol{B}|\boldsymbol{C})=\boldsymbol{H}(\boldsymbol{A},\boldsymbol{C})+\boldsymbol{H}(\boldsymbol{B},\boldsymbol{C})-\boldsymbol{H}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})-\boldsymbol{H}(\boldsymbol{C}) \tag{10}$$

2　MFSEF

在通过互信息考虑特征与标记之间的相关性来进行多标记特征选择中,先给定f表示描述样本的特征,l表示样本的类别标记,则$I(f;\ l)$虽然仅可以在单标记中描述在样本中特征和类别标记之间的相关性程度,而在多标记中,一个样本是由多个特征向量表示且隶属于多个类别标记,故给出以下定义。

定义 6　给定特征f和标记空间$\boldsymbol{L}=\{l_1,l_2,\cdots,l_n\}$,$\forall l_i\in\boldsymbol{L}(i=1,2,\cdots,n)$,$I(f;\ l_i)$为特征$f$和标记$l_i$的互信息,那么特征$f$和标记空间集$\boldsymbol{L}$的互信息可定义为:

$$\boldsymbol{FMI}(f;\ \boldsymbol{L})=\sum_{i=1}^{n}\boldsymbol{I}(f;\ l_i) \tag{11}$$

定义 7　给定一个特征子集为$\boldsymbol{S}=\{f_1,f_2,\cdots,f_m\}$,特征$f_i$与特征子集空间的互信息定义为:

$$\boldsymbol{FMI}(f_i;\ \boldsymbol{S})=\sum_{j=1,j\neq i}^{m}\boldsymbol{I}(f_i,f_j) \tag{12}$$

特征与标记集合之间的互信息大小描述了两个集合间的相关性程度,特征和标记集合的互信息越大,表明该特征越重要;反之,表明该特征重要性越弱,当特征和标记集合的互信息为零时,表明该特征和每个类别标记相互独立,此时特征和标记集合的互信息也取得最小值。

给定训练样本$\boldsymbol{U}=\{x_1,x_2,\cdots,x_n\}$和其构成样本的特征集合$\boldsymbol{F}=\{f_1,f_2,\cdots,f_d\}$,以及标记空间集合$\boldsymbol{L}=\{l_1,l_2,\cdots,l_t\}$。

由于专家特征在现实生活中是通过人的经验主观性选定,而本文实验所采用数据集为常用的多标记数据集,若单纯地人为指定专家特征,可能会影响实验的可靠性和有效性,故可事先通过数据集画出对应的特征值瀑布图,然后根据瀑布图联合互信息理论挑选出几个特征作为专家特征。图1和图2为常用多标记数据集所画的部分代表性瀑布图。

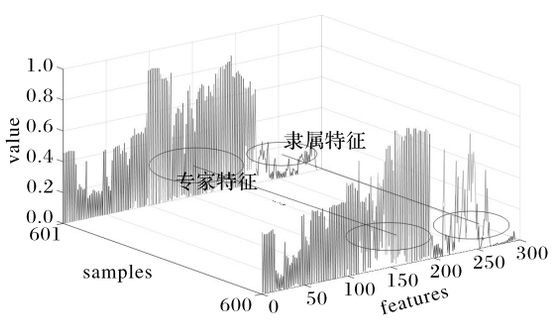

图1　第600和601个样本中所有特征对应的特征值构成的瀑布图

Fig. 1　Waterfall plot of eigenvalues corresponding to all features in the 600th and 601th samples

其中图1展示了在第600个样本到第601个样本中,每个特征所对应的特征值大小的对比,由图可看出:两个样本对应的前100个特征的特征值基本较小且小于0.5,两样本特征值基本相同,表明该段特征对标记的影响甚微;而在第200到300之间的特征所对应的特征值数值有大有小,分布极其不均衡,波动较大,说明该段特征对样本类别的划分起到决定性作用,即称隶属特征;在第100和200之间的特征所对应的特征值基本数值较大,且分布均衡,无较大波动,表明该段特征对标记的影响至关重要,即称为专家特征。

在图2中展示了在所有样本中,每个特征所对应的特征值大小的对比。针对100~200的专家特征,明显看出所有样本的特征值基本较大接近1,这表明这部分特征对于标记的划分不可或缺,但这部分特征值基本相同,表明全部样本基本上均具有该特征,所以如果事先把这些特征挑选部分出来,作为条件与剩余特征相联合,然后再针对剩余特征进行后续特征选择操作,无疑可以避免重复计算,提高特征选择速度,而且更符合现实生活中面对分类问题时,常优先选出专家特征的操作习惯,因此更具有实际应用价值。

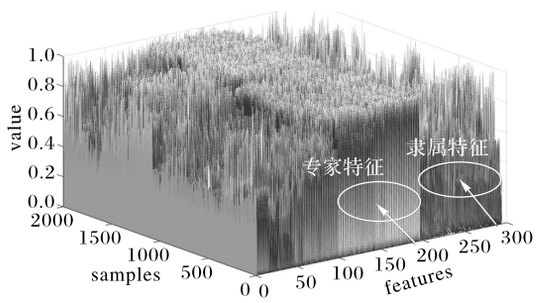

图2　全部样本所有特征对应的特征值构成的整体瀑布图

Fig. 2　Overall waterfall plot composed of eigenvalues corresponding to all features of all samples

先通过瀑布图观察出专家特征的大致分布,再通过互信息考虑特征空间和标记空间的相关性大小,由互信息大小降序挑选靠前的特征,最后综合考虑两要素选出若干个特征作为专家特征。本文实验取前四个特征作专家特征,记作$\boldsymbol{E}=\{e_1,e_2,e_3,e_4\}$。

传统的基于互信息的多标记特征选择算法仅考虑特征空间$\boldsymbol{F}$与标记空间$\boldsymbol{L}$的相关性:$\boldsymbol{F}\rightarrow\boldsymbol{L}$,本文以专家特征$\boldsymbol{E}$为条件,保持专家特征不变,将专家特征和每个原始特征作并集构建新的特征空间,再考虑其与标记空间的相关性:$\boldsymbol{E}\cup\boldsymbol{F}\rightarrow\boldsymbol{L}$。

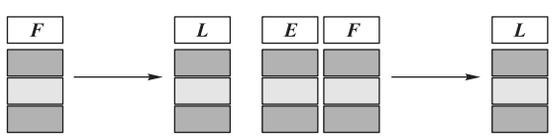

图3　专家特征联合的图解描述

Fig. 3　Graphical description of combining expert features

由定义6可知特征和标记集合的互信息越大,表明该特征越重要;反之,表明该特征重要性越弱。故用互信息大小进行降序排列得到一组新的特征序列:

$$\boldsymbol{S}=\{f_1',f_2',\cdots,f_d'\}$$

对于多标记特征选择,由于每个标记隶属于不同特征空间,因此对于特征和标记集合的相关性通过互信息大小进行计算时,相关性强的特征之间可能有比较大的冗余性,而相关性弱的特征也不一定对判别标记类别不起作用,也有可能某个相关性弱的特征往往对最后分类结果起决定性作用。考虑到此情景,可以通过建立局部子空间模型来解决此问题[10]。文献[18]中说明,当数据集的特征维度较小的时候,子空间个数可以划分2、3、4,考虑到较多保留相关性强特征,同时兼顾对某些类别标记贡献较大的但是特征与标记相关性较弱的特征,故每个特征子空间的采样比例可以设置为由大到小,例如对于2个子空间,采样比例可为:{0.6,0.4}、{0.7,0.3}、

{0.9,0.2}。又由文献[10]中实验证明3个子空间的预测效果最好,故可以将已经通过专家特征进行互信息大小降序排列后的特征序列划分为三个子空间,再通过{0.6,0.3,0.1}的采样比例进行进一步的特征选择。

局部子空间的详细过程如下:

给定有d维特征空间,三个子空间,故每个子空间特征维数为$\lceil d/3\rceil$,由定义6可计算出在子空间中每个特征和剩余特征的互信息大小:此时互信息越大,表明其特征的冗余性越高;反之,特征间互信息越小,冗余性越低。因此将通过定义6新得到的三个子空间中的特征进行升序,三个子空间关于特征间的冗余性排列分别为:$\boldsymbol{S}'_1=\{f'_{11},f'_{12},\cdots,f'_{1\lceil d/3\rceil}\}$,$\boldsymbol{S}'_2=\{f'_{21},f'_{22},\cdots,f'_{2\lceil d/3\rceil}\}$,$\boldsymbol{S}'_3=\{f'_{31},f'_{32},\cdots,f'_{3\lceil d/3\rceil}\}$。

再通过采样比例分别在三个子空间中选择冗余性比较小的特征,由于采样比例为{0.6,0.3,0.1},故三个子空间通过比例选择出的特征个数分别为:$N_1=\lceil 0.6*\lceil d/3\rceil\rceil$,$N_2=\lceil 0.3*\lceil d/3\rceil\rceil$,$N_3=\lceil 0.1*\lceil d/3\rceil\rceil$。

通过采样比例分别挑选出$\boldsymbol{S}'_1$、$\boldsymbol{S}'_2$和$\boldsymbol{S}'_3$的前N_1、N_2和N_3个特征,记作:$\boldsymbol{S}''_1=\{f'_{11},f'_{12},\cdots,f'_{1N_1}\}$,$\boldsymbol{S}''_2=\{f'_{21},f'_{22},\cdots,f'_{2N_2}\}$,$\boldsymbol{S}''_3=\{f'_{31},f'_{32},\cdots,f'_{3N_3}\}$。最后将$\boldsymbol{S}''_1$、$\boldsymbol{S}''_2$和$\boldsymbol{S}''_3$合并,就成了新的特征选择后的最终特征子集。

模拟现实世界中分类问题,引入专家特征作为条件,同原始特征作并集,通过信息论中的互信息理论背景,计算特征与标记空间的相关性大小,再结合局部子空间模型的划分,最后实现特征选择。这样既考虑了多标记特征选择在现实社会中的合理性和实用性,也避免了传统特征选择只是根据相关准则选择较强的特征导致的特征间的冗余性,最后实验结果也显示了较好的分类性能。

算法1 MFSEF。

输入 多标记数据集$\boldsymbol{D}$,专家特征$\boldsymbol{E}$;

输出 特征子集$\boldsymbol{Sub}$。

1) $\boldsymbol{FMI}=\varnothing$;
2) $\boldsymbol{E}=\{e_1,e_2,e_3,e_4\}$;
3) for each $f_i\in\boldsymbol{L}$
4) $\boldsymbol{S}=\varnothing$;$\boldsymbol{Sub}=\varnothing$;
5) for each $l_j\in\boldsymbol{L}$
6) 通过定义6计算$\boldsymbol{FMI}(f_i\cup\boldsymbol{E};\ l_j)$
7) end
8) 根据$\sum_{j=1}^{n}\boldsymbol{FMI}(f_i\cup\boldsymbol{E};\ l_j)$计算每个特征$f_i$在专家特征$\boldsymbol{E}$的条件下对所有标记互信息大小;
9) end
10) 通过第8)步计算出来的特征空间和标记空间互信息大小,对特征进行一个降序,从而得到新特征序列$\boldsymbol{S}$;
11) 将特征集合$\boldsymbol{S}$均分成3个子空间$\boldsymbol{S}_1$、$\boldsymbol{S}_2$和$\boldsymbol{S}_3$;
12) 对三个子空间分别通过定义7计算特征和特征的互信息大小,然后进行升序排列,再通过采样比例{0.6,0.3,0.1}在三个子空间分别挑选出新的特征子集;
13) 将四个专家特征和新得到的三个特征子集合并,按顺序依次放入$\boldsymbol{Sub}$。

3 实验及其结果分析

3.1 实验数据集

为验证本文算法的有效性,选取了Entertainment、Recreation、Artificial、Reference、Health、Business、Computer共7个数据集,详细信息见表1。

表1 多标记数据集

Tab. 1 Multi-label datasets

数据集	样本数	特征数	类别数	训练数	测试数
Health	5000	612	32	2000	3000
Recreation	5000	606	22	2000	3000
Artificial	5000	462	26	2000	3000
Reference	5000	793	33	2000	3000
Entertainment	5000	640	21	2000	3000
Business	5000	438	30	2000	3000
Compute	5000	681	33	2000	3000

3.2 实验环境及评价指标

本实验代码在Matlab 2016a中运行;硬件环境为Intel Core i5-2525M 2.50 GHz CPU,8 GB内存;操作系统为Windows 10。实验选取多标记常用的4种性能评价指标[19-20],即平均精度(Average Precision, AP)、海明损失(Hamming Loss, HL)、排序损失(Ranking Loss, RL)和1-错误率(One Error, OE)来综合评价多标记学习算法性能,且分别简写为:AP↑、HL↓、RL↓和OE↓,其中:↑代表指标数值越高越好,↓代表指标数值越低越好。设:多标记分类器$h(\cdot)$,预测函数$f(\cdot,\cdot)$,排序函数$rank_f$,多标记数据集$\boldsymbol{D}=\{(\boldsymbol{x}_i,\boldsymbol{Y}_i)\mid 1\leqslant i\leqslant n)\}$。上述4种评价指标AP、HL、RL和OE形式化定义如下:

1) Average Precision:评估在特定标记$\boldsymbol{y}\in\boldsymbol{Y}_i$排列的正确标记的平均分数。

$$AP_D(f)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left|\boldsymbol{Y}_i\right|}\sum_{\boldsymbol{y}\in\boldsymbol{Y}_i}\frac{\left|\{rank_f(\boldsymbol{x}_i,\boldsymbol{y}')\leqslant rank_f(\boldsymbol{x}_i,\boldsymbol{y}),\boldsymbol{y}'\in\boldsymbol{Y}_i\}\right|}{rank_f(\boldsymbol{x}_i,\boldsymbol{y})}\tag{13}$$

2) Hamming Loss:用于度量样本在单个标记的真实标记和预测标记的错误匹配情况。

$$HL_D(h)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left|\boldsymbol{Y}\right|}\left|h(\boldsymbol{x}_i)\neq\boldsymbol{Y}_i\right|\tag{14}$$

3) One Error:评估对象最高排位标记并未正确标记的次数情况。

$$OE_D(f)=\frac{1}{n}\sum_{i=1}^{n}\left[\arg\max_{\boldsymbol{y}\in\boldsymbol{Y}}f(\boldsymbol{x}_i,\boldsymbol{y})\notin\boldsymbol{Y}_i\right]\tag{15}$$

4) Ranking Loss:用来考察样本的不相关标记的排序低于相关标记的排序的情况。

$$RL_D(f)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left|\boldsymbol{Y}_i\right|\left|\overline{\boldsymbol{Y}}_i\right|}\left|\left\{(\boldsymbol{y}_1,\boldsymbol{y}_2)\mid f(\boldsymbol{x}_i,\boldsymbol{y}_1)\leqslant f(\boldsymbol{x}_i,\boldsymbol{y}_2),(\boldsymbol{y}_1,\boldsymbol{y}_2)\in\boldsymbol{Y}_i\times\overline{\boldsymbol{Y}}_i\right\}\right|\tag{16}$$

3.3 算法选择与相关参数设置

为验证新提出的特征选择算法性能,实验将MFSEF算法与4个经典多标记特征选择算法进行对比,分别是MDDMspc、MDDMproj、PMU及MFSLS。表2到表5中的第2列Original表示在原始特征空间下仅通过基本的经典多标记分类器ML-kNN的分类性能;MDDM是基于最大相关性的多标记维度约简算法,而MDDM又可划分为MDDMspc和MDDMproj,PMU是通过多元互信息最大化已选特征与标记集合的相关性,提出基于多变量互信息的多标记特征选择算法。其中MDDMspc和MDDMproj算法需先进行原始数据归一化,再进行特征选择,MFSLS和PMU是针对离散型数据进行处理,鉴于此,为了实验的严谨和合理性,以MFSLS和PMU离散

化方法为基准,对本文实验数据先两折离散化。由文献[18]可知,当选择多标记数据集的特征维度不高时,将子空间划分为3个,特征采样比例设置为{0.6,0.3,0.1},专家特征个数k_1设为4。另本实验最后分类器采用ML-kNN,故相关参数选择默认值,近邻个数k取10,平滑参数s取1。

3.4　实验结果

表2到表5给出了本文算法和其他4种算法在7个多标记数据集上实验结果,最好的结果加粗表示;同时,每种方法在所有数据集上的平均排序结果列在最后一行;数据右下角括号数字表示6种算法分别在评价指标下的排序序号。

表2　各算法在7个数据集上的平均精度测试结果

Tab. 2　AP(↑) results of each algorithm on 7 datasets

数据集	Original	MDDMspc	MDDMproj	PMU	MFSLS	MFSEF
Health	0.681 2$_{(3)}$	0.679 4$_{(4)}$	0.651 6$_{(6)}$	0.670 9$_{(5)}$	0.723 7$_{(2)}$	**0.728 0**$_{(1)}$
Recreation	0.454 7$_{(4)}$	0.449 7$_{(5)}$	0.462 8$_{(3)}$	0.444 1$_{(6)}$	0.510 2$_{(2)}$	**0.522 5**$_{(1)}$
Artificial	0.509 4$_{(3)}$	0.497 4$_{(4)}$	0.484 8$_{(6)}$	0.490 9$_{(5)}$	0.536 3$_{(2)}$	**0.536 4**$_{(1)}$
Reference	0.619 4$_{(3)}$	0.601 4$_{(5)}$	0.599 2$_{(6)}$	0.614 5$_{(4)}$	0.630 4$_{(2)}$	**0.634 7**$_{(1)}$
Entertainment	0.602 3$_{(3)}$	0.551 3$_{(6)}$	0.558 8$_{(5)}$	0.567 1$_{(4)}$	0.603 2$_{(2)}$	**0.604 0**$_{(1)}$
Business	**0.879 8**$_{(1)}$	0.870 7$_{(5)}$	0.873 1$_{(4)}$	0.862 8$_{(6)}$	0.876 2$_{(3)}$	0.876 5$_{(2)}$
Computer	0.633 5$_{(3)}$	0.631 9$_{(4)}$	0.625 0$_{(6)}$	0.625 2$_{(5)}$	0.640 5$_{(2)}$	**0.642 4**$_{(1)}$
均值	0.625 8	0.611 7	0.607 9	0.610 8	0.645 8	0.649 2

表3　各算法在7个数据集上的海明损失测试结果

Tab. 3　HL(↓) results of each algorithm on 7 datasets

数据集	Original	MDDMspc	MDDMproj	PMU	MFSLS	MFSEF
Health	0.045 8$_{(5)}$	0.043 8$_{(4)}$	0.046 2$_{(6)}$	0.043 5$_{(3)}$	0.041 0$_{(2)}$	**0.038 6**$_{(1)}$
Recreation	0.061 8$_{(3)}$	0.063 3$_{(5)}$	0.061 9$_{(4)}$	0.063 7$_{(6)}$	0.059 8$_{(2)}$	**0.058 8**$_{(1)}$
Artificial	0.061 2$_{(4)}$	0.061 6$_{(6)}$	0.060 9$_{(3)}$	0.061 5$_{(5)}$	**0.058 7**$_{(1)}$	0.059 4$_{(2)}$
Reference	0.031 4$_{(4)}$	0.032 4$_{(6)}$	0.031 1$_{(3)}$	0.030 7$_{(2)}$	0.031 5$_{(5)}$	**0.028 8**$_{(1)}$
Entertainment	0.061 2$_{(4)}$	0.062 4$_{(6)}$	0.062 0$_{(5)}$	0.060 7$_{(3)}$	0.059 4$_{(2)}$	**0.059 1**$_{(1)}$
Business	**0.026 9**$_{(1)}$	0.028 0$_{(4.5)}$	0.028 0$_{(4.5)}$	0.028 5$_{(6)}$	0.027 4$_{(3)}$	0.027 2$_{(2)}$
Computer	0.041 2$_{(6)}$	0.040 8$_{(4.5)}$	0.040 8$_{(4.5)}$	0.040 5$_{(3)}$	0.040 1$_{(2)}$	**0.040 0**$_{(1)}$
均值	0.047 1	0.047 5	0.047 3	0.047 0	0.045 4	0.044 6

表4　各算法在7个数据集上的排序损失测试结果

Tab. 4　RL(↓) results of each algorithm on 7 datasets

数据集	Original	MDDMspc	MDDMproj	PMU	MFSLS	MFSEF
Health	0.060 5$_{(3)}$	0.065 2$_{(5)}$	0.069 9$_{(6)}$	0.063 5$_{(4)}$	**0.056 3**$_{(1)}$	0.056 5$_{(2)}$
Recreation	0.191 4$_{(5)}$	0.189 2$_{(3)}$	0.192 4$_{(6)}$	0.189 5$_{(4)}$	0.177 0$_{(2)}$	**0.173 0**$_{(1)}$
Artificial	0.152 0$_{(3)}$	0.153 9$_{(5)}$	0.157 6$_{(6)}$	0.153 0$_{(4)}$	0.146 8$_{(2)}$	**0.145 7**$_{(1)}$
Reference	0.091 9$_{(4)}$	0.092 5$_{(5)}$	0.093 3$_{(6)}$	0.087 0$_{(2)}$	0.088 3$_{(3)}$	**0.085 5**$_{(1)}$
Entertainment	0.115 4$_{(3)}$	0.126 4$_{(6)}$	0.125 8$_{(5)}$	0.122 6$_{(4)}$	0.115 0$_{(2)}$	**0.113 1**$_{(1)}$
Business	**0.037 4**$_{(1)}$	0.043 3$_{(5)}$	0.041 6$_{(4)}$	0.045 9$_{(6)}$	0.040 2$_{(2)}$	0.040 7$_{(3)}$
Computer	0.092 2$_{(4)}$	0.091 9$_{(3)}$	0.094 5$_{(5)}$	0.095 2$_{(6)}$	0.089 6$_{(2)}$	**0.089 4**$_{(1)}$
均值	0.105 8	0.108 9	0.110 7	0.108 1	0.101 9	0.100 6

表5　各算法在7个数据集上的1-错误率测试结果

Tab. 5　OE(↓) results of each algorithm on 7 datasets

数据集	Original	MDDMspc	MDDMproj	PMU	MFSLS	MFSEF
Health	0.420 7$_{(4)}$	0.405 3$_{(3)}$	0.446 3$_{(6)}$	0.431 3$_{(5)}$	0.347 7$_{(2)}$	**0.340 0**$_{(1)}$
Recreation	0.706 3$_{(4)}$	0.714 3$_{(5)}$	0.688 3$_{(3)}$	0.719 0$_{(6)}$	0.616 3$_{(2)}$	**0.608 3**$_{(1)}$
Artificial	0.632 7$_{(3)}$	0.647 0$_{(4)}$	0.667 0$_{(6)}$	0.657 0$_{(5)}$	0.584 0$_{(2)}$	**0.583 7**$_{(1)}$
Reference	0.473 0$_{(3)}$	0.497 3$_{(6)}$	0.496 0$_{(5)}$	0.491 0$_{(4)}$	0.461 7$_{(2)}$	**0.454 0**$_{(1)}$
Entertainment	0.529 7$_{(2)}$	0.602 0$_{(6)}$	0.599 7$_{(5)}$	0.584 7$_{(4)}$	**0.525 7**$_{(1)}$	0.533 0$_{(3)}$
Business	**0.121 3**$_{(1)}$	0.128 3$_{(5)}$	0.126 3$_{(4)}$	0.136 0$_{(6)}$	0.122 7$_{(2.5)}$	0.122 7$_{(2.5)}$
Computer	0.436 3$_{(3)}$	0.445 3$_{(4)}$	0.453 7$_{(6)}$	0.449 7$_{(5)}$	0.432 3$_{(2)}$	**0.430 7**$_{(1)}$
均值	0.474 3	0.491 4	0.496 8	0.495 5	0.441 5	0.438 9

实验结果显示:MFSEF在Health、Recreation、Artificial、Reference、Entertainment、Business、Computer等7个多标记数据集上实验结果的平均排序都占优,其中,对于AP指标,平均数值越大,算法性能越优,对于其他三个评价指标,平均数值越小,算法性能越优。从表2到表5可发现:

① 在AP指标中,MFSEF仅在Business数据集中AP不是最优,排名第二。对比4种算法和原始特征空间,MFSEF在其他数据集中AP值最大,即分类性能达到最优。

② 在HL指标中,MFSEF在Business和Artificial数据集中HL值排名第二,对比4种算法和原始特征空间,MFSEF在其

他数据集中HL值最小,即分类性能达到最优。

③ 在RL指标中,MFSEF在Health和Business数据集中分别排第二和第三,在其他5个数据集对比4种算法和原始特征空间,MFSEF的RL值最小,即分类性能达到最优。

④ 在OE指标中,MFSEF在Entertainment数据集上排第三,在Business数据集和MFSLS对比算法并列排第二,在其他5个数据集对比4种算法和原始特征空间,MFSEF的OE值最小,即分类性能达到最优。

上述实验分析充分表明,通过本文算法特征选择后的特征子集在后续分类性能上,对比其他4种算法和原始特征空间在7个多标记数据集中多数占优,验证了本文算法的有效性和鲁棒性。

图4是对比其他4种算法,随着选择后的特征数目的逐渐变多,其分类性能的变化情况。针对每一种算法,都有28种对比结果。介于篇幅所限,本文只选取了Artificial数据集的曲线图进行分析,分别展示了在AP、HL、RL和OE四种评价指标时,5种算法在特征数逐渐变大时分类性能的变化情况。对比原始特征空间和PMU、MDDMspc、MDDMproj、MFSLS这4种算法的分类性能,在Artificial数据集上,MFSEF基本上占优。基本上在前80个特征范围类,本文算法在四个评价指标上均明显优于其他对比算法,并且往往能通过较少的特征数更快地达到更好的分类性能。另外,在其他未展示的数据集上,本文算法的分类性能曲线变化也多数占优,这充分地表明MFSEF的有效性和合理性。

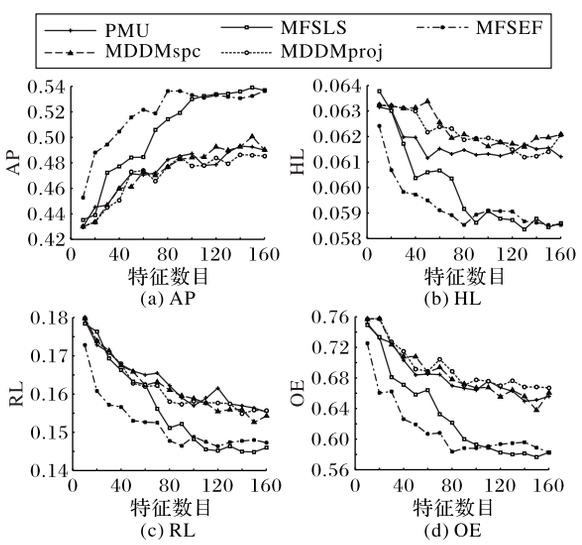

图4 Artificial数据集的各个评价指标的性能变化

Fig. 4 Changes in performance of various evaluation indicators on Artificial dataset

3.5 统计假设检验

在上述实验分析中,本文算法达到了显而易见的效果,下面将运用统计学中统计假设检验[21]进一步说明本文算法的有效性和合理性。

统计假设检验:在上述7个数据集上采用显著性水平为5%的Nemenyi检验[22-23]来对比MFSEF算法与其他对比算法。如果两个算法在所有数据集上的平均排序的差值小于或者等于临界差值(Critical Difference, CD),那么这两个算法之间没有显著性差异,反之存在显著性差异。如图5所示,在最上行为临界值CD=2.850 0时,若两个算法之间没有显著性差异则用实线连接。在图5中,随着坐标轴上的数值增大其算法性能依次降低。

图5展示了各算法在AP、RL、HL、OE四个指标上的CD图,从中可以看出,本文算法在4个指标上性能均处于首位。具体在平均精度指标上,如图5(a)、(b)、(d)所示,在Average Precision、Ranking Loss和One Error三个指标上,MLSEF与MDDMspc、PMU、MDDMproj均有显著差异,且优于这三种算法;如图5(c)在Hamming Loss指标上,本文算法与MDDMproj和MDDMspc有显著差异,且优于这两种算法。与其他算法相比,在统计上,本文算法有45%的概率与其他算法无显著性差异。

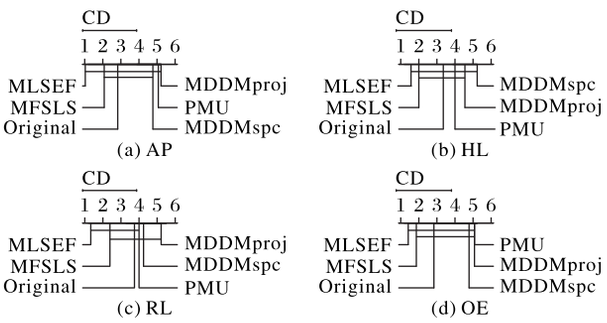

图5 算法综合性能比较

Fig. 5 Performance comparison of algorithms.

通过以上针对图5的分析,可知本文算法综合性能最为优秀,在统计上也优于其他对比算法。基于以上的实验结果和统计分析再次充分表明本文算法的优越性。

4 结语

与通过相关准则挑选特征和标记相关性强的多标记特征选择算法相比,本文不仅考虑了重要性较次要的特征可能对分类性能产生的影响,还考虑了可能决定整体预测方向的最关键特征。模拟现实生活中的实际情况,通过经验优先挑选出部分专家特征与剩余特征相联合,利用条件互信息和联合互信息理论得出一个与标记集合相关性由强到弱的特征序列,再通过划分子空间去除冗余性较大的特征,最后保留专家特征和挑选出的新的特征作为最后的特征子集。在已公开的多个基准多标记数据集中的实验结果表明,该算法在实验中较其他对比的多标记特征选择算法有一定优势和较好的稳定性,且更具有实际应用价值。

本文算法在专家特征的选取上还可以进一步探讨,目前只是通过瀑布图联合互信息理论模拟选出几个专家特征,所以最后结果可能由于个人选取专家特征的不同,实验结果和预期效果存在一定的误差,但是针对具体问题分析数据,合理选择专家特征,还是可以有效减少误差。

参考文献 (References)

[1] GIBAJA E, VENTURA S. A tutorial on multilabel learning [J]. ACM Computing Surveys, 2015, 47(3): 1-38.

[2] 何志芬,杨明,刘会东. 多标记分类和标记相关性的联合学习[J]. 软件学报, 2014, 25(9): 1967-1981. (HE Z F, YANG M, LIU H D. Joint learning of multi-label classification and label correlations[J]. Journal of Software, 2014, 25(9): 1967-1981.)

[3] WANG Z, CHEN T, LI G, et al. Multi-label image recognition by recurrently discovering attentional regions [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway:

IEEE, 2017: 464-472.

[4] OZONAT K, YOUNG D. Towards a universal marketplace over the web: statistical multi-label classification of service provider forms with simulated annealing [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009:1295-1304.

[5] SHU X, LAI D, XU H, et al. Learning shared subspace for multi-label dimensionality reduction via dependence maximization [J]. Neurocomputing, 2015, 168: 356-364.

[6] PEREIRA R B, PLASTINO A, ZADROZNY B, et al. Categorizing feature selection methods for multi-label classification[J]. Artificial Intelligence Review, 2018, 49(1): 57-78.

[7] ZHANG Y, ZHOU Z. Multilabel dimensionality reduction via dependence maximization [J]. ACM Transactions on Knowledge Discovery from Data, 2010, 4(3): No. 14.

[8] LEE J, KIM D W. Feature selection for multi-label classification using multivariate mutual information [J]. Pattern Recognition Letters, 2013, 34(3): 349-357.

[9] LIN Y, HU Q, LIU J, et al. Multi-label feature selection based on neighborhood mutual information [J]. Applied Soft Computing, 2016, 38: 244-256.

[10] 刘景华,林梦雷,王晨曦,等. 基于局部子空间的多标记特征选择算法[J]. 模式识别与人工智能, 2016, 29(3): 240-251. (LIU J H, LIN M L, WANG C X, et al. Multi-label feature selection algorithm based on local subspace [J]. Pattern Recognition and Artificial Intelligence, 2016, 29(3): 240-251.)

[11] 王晨曦,林耀进,唐莉,等. 基于信息粒化的多标记特征选择算法[J]. 模式识别与人工智能, 2018, 31(2): 123-131. (WANG C X, LIN Y J, TANG L, et al. Multi-label feature selection based on information granulation [J]. Pattern Recognition and Artificial Intelligence, 2018, 31(2): 123-131.)

[12] LEE J, LIM H, KIM D W. Approximating mutual information for multi-label feature selection [J]. Electronics Letters, 2012, 48(15): 929-930.

[13] YU S, HUANG T. Exponential weighted entropy and exponential weighted mutual information [J]. Neurocomputing, 2017, 249: 86-94.

[14] LI F, MIAO D, PEDRYCZ W. Granular multi-label feature selection based on mutual information[J]. Pattern Recognition, 2017, 67: 410-423.

[15] ZHANG M, ZHOU Z. ML-*k*NN: a lazy learning approach to multi-label learning[J]. Pattern recognition, 2007, 40(7): 2038-2048.

[16] ZHANG M, ZHOU Z. A review on multi-label learning algorithms [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8):1819-1837.

[17] FLEURET F. Fast binary feature selection with conditional mutual information[J]. Journal of Machine Learning Research, 2004, 5: 1531-1555.

[18] 杨明,王飞. 一种基于局部随机子空间的分类集成算法[J]. 模式识别与人工智能, 2012, 25(4): 595-603. (YANG M, WANG F. A classifier ensemble algorithm based on local random subspace [J]. Pattern Recognition and Artificial Intelligence, 2012, 25(4): 595-603.)

[19] TSOUMAKAS G, VLAHAVAS I. Random *k*-labelsets: an ensemble method for multilabel classification [C]// Proceedings of the 18th European Conference on Machine Learning, LNCS 4701. Berlin: Springer, 2007: 406-417.

[20] ZHANG M, PEÑA J M, ROBLES V. Feature selection for multi-label naive Bayes classification[J]. Information Sciences, 2009, 179(19): 3218-3229.

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

[22] ZHANG M, WU L. LIFT: multi-label learning with label-specific features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1): 107-120.

[23] 程玉胜,钱坤,王一宾,等. 融合萤火虫方法的多标签懒惰学习算法[J]. 计算机应用,2019,39(5):1305-1311. (CHENG Y S, QIAN K, WANG Y B, et al. Multi-label lazy learning algorithm based on firefly method [J]. Journal of Computer Applications, 2019, 39(5): 1305-1311.)

This work is partially supported by the Key University Natural Science Research Project of Anhui Province (KJ2017A352), the Program for Innovative Research Team in Anqing Normal University.

CHENG Yusheng, born in 1969, Ph. D, professor. His research interests include big data, rough set, machine learning for feature selection.

SONG Fan, born in 1992, M. S. candidate. His research interests include multi-label learning, neural network.

WNAG Yibin, born in 1970, M. S., professor. His research interests include multi-label learning, machine learning, software security.

QIAN Kun, born in 1995, M. S. candidate. His research interests include multi-label learning, machine learning, data statistics.