基于标签关系改进的多标签特征选择算法

陈福才 李思豪 张建朋 黄瑞阳
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要 多标签特征选择是应对数据维度灾难现象的主要方法之一,可以在降低特征维度的同时提高学习效率,优化分类性能。针对目前特征选择算法没有考虑标签间的相互关系,以及信息量的衡量范围存在偏差的问题,提出一种基于标签关系改进的多标签特征选择算法。首先引入对称不确定性对信息量进行归一化处理,然后用归一化的互信息量作为相关性的衡量方法,并据此定义标签的重要性权重,对依赖度和冗余度中的标签相关项进行加权处理;进而提出一种特征评分函数作为特征重要性的评价指标,并依次选择出评分最高的特征组成最佳特征子集。实验结果表明,与其他算法相比,该算法在提取出更加精确的低维特征子集后,不仅能够有效提高面向实体信息挖掘的多标签学习算法的性能,也能提高基于离散特征的多标签学习算法的效率。

关键词 多标签特征选择,标签关系,依赖度,冗余度,特征评分
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.06.041

Multi-label Feature Selection Algorithm Based on Improved Label Correlation

CHEN Fu-cai LI Si-hao ZHANG Jian-peng HUANG Rui-yang
(National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002, China)

Abstract Multi-label feature selection is one of the essential methods to overcome the curse of dimensionality. It reduces the feature dimension, improves the learning efficiency, and optimizes the classification performance. However, many existing feature selection algorithms hardly take label correlation into consideration, and the range of information entropies are biased within different data sets. To address those problems, this paper proposed a multi-label feature selection algorithm based on the improved label correlation. The algorithm firstly uses symmetrical uncertainty to normalize the information entropy, and takes normalized mutual information as relationship measurement to define the label importance, with which the label-related items in dependency and redundancy are weighted. In the end, the score function is put forward to evaluate the feature importance, and the best feature subset is selected with the highest score. Experiments demonstrate that after selecting out the concise and accurate feature subset, the multi-label classification is accelerated in terms of the performance and the efficiency with disperse features.

Keywords Multi-label feature selection, Label correlation, Dependency, Redundancy, Feature score

1 引言

多标签学习算法(Multi-Label Learning,MLL)已经成为基因工程、多媒体语义标注、文本分类等领域研究和应用的热点。近年来,数据更新速度的不断加快及其来源范围的不断拓展,使得数据特征维度越来越高,进一步导致多标签学习面临过拟合、计算复杂度高等挑战,即维度灾难现象[1-2]。作为MLL的一种数据预处理过程,多标签特征选择(Multi-Label Feature Selection,MLFS)通过从数据集中移除冗余的、不相关的和有噪声的特征,使用更加精简和准确的特征子集代替原始特征集进行标签学习,达到了降低特征维度、提高学习效率、优化分类性能和减小算法消耗的目的[3-6]。

根据特征子集重要性衡量方法的不同,MLFS主要分为过滤式、封装式和嵌入式3类[5,7-8]。过滤式通过利用数据本身固有的分布特性来计算特征评分,并以此作为特征选择的依据,如mRMR算法[9];封装式分别针对不同的特征子集执行分类算法,并以分类算法的性能来评价该特征子集,如HOML算法[10];嵌入式则直接将特征选择过程融入到标签

到稿日期:2017-04-25 返修日期:2017-07-29 本文受国家重点研发计划项目(2016YFB0800101),国家自然科学基金创新研究群体项目(61521003)资助。

陈福才(1974-),男,研究员,硕士生导师,主要研究方向为网络大数据分析与电信网信息关防,E-mail:1242100831@qq.com(通信作者);**李思豪**(1991-),男,硕士生,主要研究方向为网络大数据分析,E-mail:michaelbournelisihao@outlook.com;**张建朋**(1988-),男,博士生,主要研究方向为网络大数据分析与数据流挖掘;**黄瑞阳**(1986-),男,博士,助理研究员,主要研究方向为网络大数据分析。

学习算法中，如 MEFS 算法[11]。由于具有独立于分类算法、比封装式方法的复杂度更低等优点，过滤多标签特征选择受到普遍关注。

在对 MLFS 的研究中，很多学者利用香农信息熵理论来度量数据的分布特性。Doquireand 和 Verleysen[12] 首先利用 PPT 算法将多标签问题转化为单标签问题，再利用基于互信息的贪婪搜索策略选择与标签相关性最大的特征。张振海等人提出的 MLFSIE 算法[13]则利用特征与标签集合之间的信息增益来衡量特征的重要程度，并根据所提出的信息增益阈值选择方法删除不相关特征。上述两种方法只考虑了特征和标签集之间的关系，Hanchuan 等人的 mRMR 算法[9]则定义了特征之间的冗余度指标以及特征与标签集之间的相关性指标，进而选择出冗余度低而相关性高的特征子集。基于该思想，Monalisa 和 Anirban[14]将特征相关性指标和冗余度指标之比作为特征评分，每次迭代挑选出评分最高的特征，直至所选特征集规模达到预先指定的阈值；更进一步，Lin 和 Hu 等人[15]提出的 MDMR 算法在冗余度的指标中还考虑了特征间的条件互信息量。

综合而言，目前对特征选择算法的研究还普遍存在两个问题：1）仅考虑了特征间以及特征与标签间的关系，忽略了多标签学习中不同标签之间存在的相关性；2）没有考虑由数据分布不同引起的信息量衡量范围的偏差，信息量的可比性影响了结果的可靠性。针对上述问题，本文提出一种基于标签关系改进的多标签特征选择算法 MLFSLC。该算法首先用归一化的互信息量作为关联度的衡量方法，定义了标签重要性，并以此为权重来改进特征依赖度和冗余度的计算方法，从而考虑了标签的相关性；然后，结合依赖度和冗余度指标提出一种规范化的特征评分函数，将不同的特征在统一的指标下进行评价；最后，依据特征评分选择出最佳特征子集。通过 3 种评价指标的实验结果验证了 MLFSLC 算法的有效性。

2　相关知识

2.1　多标签数据集

为方便后文描述特征选择算法，首先对多标签数据集进行定义。

定义 1　设在 d 维特征空间中，$\boldsymbol{X}\subset\mathbb{R}^d$ 表示输入样本集，$\boldsymbol{x}=\{x_1,x_2,\cdots,x_d\}\in\boldsymbol{X}$ 表示一个样本，x_i 表示样本 $\boldsymbol{x}$ 的第 i 维特征，$\boldsymbol{L}=\{l_1,l_2,\cdots,l_q\}$ 表示标签集，q 为标签数，则定义多标签数据集如式(1)所示。

$$\boldsymbol{D}=\{(\boldsymbol{x},\boldsymbol{Y})\mid\boldsymbol{x}\in\boldsymbol{X},\boldsymbol{Y}\subseteq\boldsymbol{L}\} \tag{1}$$

其中，$\boldsymbol{Y}=\{y_1,y_2,\cdots,y_q\}$ 是样本 $\boldsymbol{x}$ 对应的标签子集，当且仅当样本 $\boldsymbol{x}$ 有标签 l_i 时，$y_i=1$，否则 $y_i=0$。

2.2　互信息量

信息熵(Information Entropy)描述了集合的不确定程度。设有集合 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_m\}$ 和 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_n\}$，$p(x_i)$为元素 x_i 的先验概率，则集合的信息熵、集合间的联合信息熵和条件信息熵的计算分别如式(2)－式(4)所示。

$$H(\boldsymbol{X})=-\sum_{i=1}^{m}p(x_i)\log_2 p(x_i) \tag{2}$$

$$H(\boldsymbol{X},\boldsymbol{Y})=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\log_2 p(x_i,y_j) \tag{3}$$

$$\begin{aligned}H(\boldsymbol{Y}\mid\boldsymbol{X})&=H(\boldsymbol{X},\boldsymbol{Y})-H(\boldsymbol{X})\\&=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\log_2 p(y_j\mid x_i)\end{aligned} \tag{4}$$

因此，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的互信息量(Mutual Information)表示已知 $\boldsymbol{Y}$ 的条件下，$\boldsymbol{X}$ 不确定性的改变量。其从统计学的角度反映了 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的关联程度，如式(5)所示。

$$\begin{aligned}MI(\boldsymbol{X};\boldsymbol{Y})&=H(\boldsymbol{X})-H(\boldsymbol{X}\mid\boldsymbol{Y})\\&=\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\log_2\frac{p(x_i\mid y_j)}{p(x_i)}\end{aligned} \tag{5}$$

互信息量具有如下性质：

性质 1　互信息量具有对称性，即：

$$MI(\boldsymbol{X};\boldsymbol{Y})=MI(\boldsymbol{Y};\boldsymbol{X})=H(\boldsymbol{X})+H(\boldsymbol{Y})-H(\boldsymbol{X},\boldsymbol{Y})$$

设集合 $\boldsymbol{Z}=\{z_1,z_2,\cdots,z_t\}$，则将已知 $\boldsymbol{Z}$ 的条件下 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的互信息量称为条件互信息量(Conditional Mutual Information)，如式(6)所示。

$$MI(\boldsymbol{X};\boldsymbol{Y}\mid\boldsymbol{Z})=\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{t}p(x_i,y_j,z_k)\log_2\frac{p(x_i\mid y_j,z_k)}{p(x_i\mid z_k)} \tag{6}$$

条件互信息量具有如下性质：

性质 2　$MI(\boldsymbol{X};\boldsymbol{Y}\mid\boldsymbol{Z})=MI(\boldsymbol{X};\boldsymbol{Y},\boldsymbol{Z})-MI(\boldsymbol{X};\boldsymbol{Z})$

鉴于信息熵理论具有较高的计算效率和更好的可解释性，本文将其作为衡量特征重要性的方法。

2.3　特征依赖度与冗余度

依赖度(Dependency)与冗余度(Redundancy)是特征重要性的两个方面[7]。下面分别给出这两个因素的定义并对其进行分析。

首先，依赖度衡量特征与标签的关联程度，如定义 2 所示。

定义 2　在多标签特征选择算法中，设 f 表示待选特征，$\boldsymbol{L}=\{l_1,l_2,\cdots,l_q\}$ 表示标签集，q 为标签数，则依赖度定义为 f 和 $\boldsymbol{L}$ 的平均互信息量，如式(7)所示。

$$Dp(f)=\frac{1}{q}\sum_{i=1}^{q}MI(f;l_i) \tag{7}$$

结合互信息量的定义可知，若 f 和 $\boldsymbol{L}$ 相互独立，则 Dp 取值最小；若 f 和 $\boldsymbol{L}$ 完全相关，则 Dp 取值最大。因此，依赖度越高，特征和标签的关联程度越强，特征越重要。

然而，与标签集依赖度高的特征之间也可能存在信息重叠，重叠的信息对分类器的性能无益。信息重叠的多少用冗余度来衡量，如定义 3 所示。

定义 3　设 $\boldsymbol{S}=\{s_1,s_2,\cdots,s_t\}$ 表示已选特征子集，t 为已选特征子集的大小，则特征冗余度定义为待选特征 f 和 $\boldsymbol{S}$ 的平均互信息量，如式(8)所示。

$$Rd(f)=\frac{1}{t}\sum_{i=1}^{t}MI(f;s_i) \tag{8}$$

文献[15]还在冗余度的计算中考虑了 f，$\boldsymbol{L}$ 和 $\boldsymbol{S}$ 的条件互信息量，如式(9)所示。

$$Rd(f)=\frac{1}{t}\sum_{i=1}^{t}\left[MI(f;s_i)-\sum_{j=1}^{q}MI(f;l_j\mid s_i)\right] \tag{9}$$

由式(9)可知，冗余度越大，特征间的信息重叠越多，移除该特征对系统的影响越小，特征的重要性越低。

3 基于标签关系改进的多标签特征选择算法的相关知识

3.1 互信息量的归一化处理

由于数据分布不同，不同的互信息量可能有不同的衡量范围。为使它们之间具有更好的可比性，在计算特征评分前，首先选择对称不确定性(Symmetrical Uncertainty)[16]对互信息量进行归一化处理，如式(10)所示。

$$NMI(\boldsymbol{X},\boldsymbol{Y})=2\left[\frac{MI(\boldsymbol{X};\boldsymbol{Y})}{H(\boldsymbol{X})+H(\boldsymbol{Y})}\right] \tag{10}$$

易证明 $NMI(\boldsymbol{X},\boldsymbol{Y})\in[0,1]$，且 $NMI(\boldsymbol{X},\boldsymbol{Y})=0$ 表示 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 相互独立，$NMI(\boldsymbol{X},\boldsymbol{Y})=1$ 表示可通过 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之一确定另一个。

3.2 标签的重要性及对特征依赖度和冗余度的改进

如果一个标签的重要性越强，那么在特征重要性的计算过程中，包含该标签的相关项也应具有较大的权重。根据该思想，本文利用标签重要性对特征依赖度和冗余度进行改进。

标签重要性与标签间的相关性密切相关，首先给出标签重要性的定义，如定义 4 所示。

定义 4 将标签 l_i 的重要性定义为标签 l_i 与标签集的关联度占所有标签与标签集的关联度之和的比重。

标签与标签集的关联度 $Ld(l_i)$ 用互信息量进行衡量，如式(11)所示。

$$Ld(l_i)=\frac{1}{q-1}\sum_{l_j\in\boldsymbol{L},j\neq i}NMI(l_i;l_j) \tag{11}$$

其中，$\boldsymbol{L}=\{l_1,l_2,\cdots,l_q\}$ 表示标签集，q 为标签数，$l_k\in\boldsymbol{L}(1\leqslant k\leqslant q)$ 表示标签集中的标签。

根据定义 4，标签 l_i 的重要性 $IMP(l_i)$ 的计算方法如式(12)所示。

$$IMP(l_i)=\frac{Ld(l_i)}{\sum_{j=1}^{q}Ld(l_j)}=\frac{\sum_{l_j\in\boldsymbol{L},j\neq i}NMI(l_i;l_j)}{\sum_{i=1}^{q}\sum_{l_j\in\boldsymbol{L},j\neq i}NMI(l_i;l_j)} \tag{12}$$

然后以标签重要性为权重对式(7)和式(9)中的标签相关项进行加权，使得改进后的依赖度和冗余度在体现特征重要性的同时，也考虑了标签的关系，如式(13)和式(14)所示。

$$Dp(f)=\frac{1}{q}\sum_{i=1}^{q}IMP(l_i)NMI(f;l_i) \tag{13}$$

$$Rd(f)=\frac{1}{t}\sum_{i=1}^{t}\left[NMI(f;s_i)-\sum_{j=1}^{q}IMP(l_i)NMI(f;l_j|s_i)\right] \tag{14}$$

3.3 融合标签关系的特征评分函数及其计算

由 2.3 节的分析可知，依赖度和冗余度分别从特征与标签集关系的角度和特征与已选特征集关系的角度体现了特征的重要性，而多标签特征选择的目的就是过滤出依赖度尽可能高而冗余度尽可能低的特征。因此，本文基于依赖度和冗余度指标具有同等重要性的假设，将两者的差值作为特征评分函数来度量特征的重要性，如式(15)所示。

$$score(f)=Dp(f)-Rd(f) \tag{15}$$

将式(13)和式(14)代入上式，即可得到式(16)。

$$score(f)=\frac{1}{q}\sum_{i=1}^{q}IMP(l_i)NMI(f;l_i)-\frac{1}{t}\sum_{i=1}^{t}\left[NMI(f;s_i)-\sum_{j=1}^{q}IMP(l_i)NMI(f;l_j|s_i)\right] \tag{16}$$

由式(15)和式(16)可以看出，特征评分函数分为两项。第一项对应依赖度，标签集不会因特征选择结果的不同而不同，因此对候选特征 f 而言，该项是常数，可预先计算得到。注意到第二项中，每一个候选特征 f 的冗余度都与前一个已选特征 $s\in\boldsymbol{S}$ 密切相关。首先用 $Rd(f;s)$ 表示 f 和 s 的冗余度，如式(17)所示。

$$Rd(f;s)=NMI(f;s)-\sum_{j=1}^{q}IMP(l_i)NMI(f;l_j|s) \tag{17}$$

由于 $\boldsymbol{S}$ 中的特征是通过选择顺序依次排列的，令 $\boldsymbol{S}_n$ 表示第 n 次选择的特征，则评分函数的冗余度项可重写为如式(18)所示的形式。

$$\begin{aligned}R(f)&=\frac{1}{|\boldsymbol{S}|}\sum_{n=1}^{|\boldsymbol{S}|}Rd(f;\boldsymbol{S}_n)\\&=\frac{1}{|\boldsymbol{S}|}\left[Rd(f;S_{|\boldsymbol{S}|})+\sum_{n=1}^{|\boldsymbol{S}|-1}Rd(f;\boldsymbol{S}_n)\right]\end{aligned} \tag{18}$$

令 $R_t(f)$ 表示第 t 次特征选择后特征 f 的冗余度项，则由式(18)可得，当 $t\neq0$ 时，有：

$$\begin{aligned}R_t(f)&=\frac{1}{t}\left[Rd(f;\boldsymbol{S}_t)+(t-1)R_{t-1}(f)\right]\\&=R_{t-1}(f)+\frac{Rd(f;\boldsymbol{S}_t)-R_{t-1}(f)}{t}\end{aligned} \tag{19}$$

因此，对特征评分冗余项的迭代计算公式如式(20)所示，其中 $t=0$ 表示尚未进行特征选择，故冗余度为 0。

$$R_t(f)=\begin{cases}R_{t-1}(f)+\dfrac{Rd(f;\boldsymbol{S}_t)-R_{t-1}(f)}{t}, & t\geqslant1\\0, & t=0\end{cases} \tag{20}$$

通过式(20)，在计算第 t 次迭代中候选特征 f 的冗余度项时，只需利用第 $t-1$ 次的迭代结果，并计算 f 和最新已选特征的冗余度即可。

3.4 算法描述与分析

MLFSLC 算法使用一种融合了标签关系的特征评分函数作为特征选择依据。如图 1 所示，令 $\boldsymbol{F}=\{f_1,f_2,\cdots,f_m\}$ 表示原始多标签特征集，$\boldsymbol{S}$ 表示已选特征集，则算法思想为：从原始多标签特征集开始，选择出候选特征集中满足 $\max\limits_{f\in\boldsymbol{F}-\boldsymbol{S}}score(f)$ 的特征，并依次加入到已选特征集，直至已选特征集达到指定大小为止。

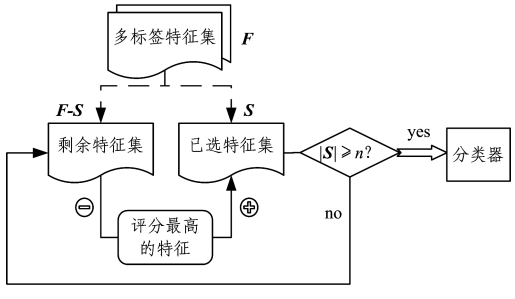

图 1 MLFSLC 算法示意图

Fig. 1 Schematic diagram of MLFSLC algorithm

MLFSLC 算法的伪代码如算法 1 所示。

算法 1　MLFSLC 算法

输入：原始特征集 **F**，标签集 **L**，特征选择比例 δ

输出：已选特征子集 **S**

1. 计算预定的已选特征子集大小 n=|**F**|×δ；
2. 初始化 **S** 为空集；
3. 对于每一个标签 $l_i \in \mathbf{L}$，利用式(12)计算标签重要性 IMP(l_i)；
4. 对于每一个特征 f∈**F**，利用式(13)计算依赖度项 $D_f \leftarrow Dp(f)$，且初始化冗余度项 $R_f \leftarrow 0$；
5. while |**S**|<n do
6. 　当|**S**|>0 时，对于每一个特征 f∈**F**，通过式(20)来更新冗余度项，即 $R_f \leftarrow R_f + [Rd(f;\mathbf{S}_{|\mathbf{S}|}) - R_f]/|\mathbf{S}|$；
7. 　利用式(15)计算每一个特征 f∈**F** 的特征评分，即 score(f)=$D_f - R_f$；
8. 　选择特征评分最大的特征 f_M，即 $f_M \leftarrow \max_{f \in F-S} \text{score}(f)$；
9. 　将 f_M 加入到已选特征集中，即 $\mathbf{S} \leftarrow \mathbf{S} + \{f_M\}$；
10. 将 f_M 从原始特征集中删除，即 $\mathbf{F} \leftarrow \mathbf{F} - \{f_M\}$；
11. end while
12. end

下面分析算法的时间复杂度。令 q 为标签数，$m=|\boldsymbol{F}|$ 为原始特征集的大小，$n=|\boldsymbol{F}|\times\delta=m\delta$ 为预定的已选特征子集的大小，则步骤 3 的复杂度为 $O(q^2)$，步骤 4 的复杂度为 $O(mq)$；步骤 5－步骤 12 的复杂度为 $O(n^2+mn)=O((\delta^2+\delta)m^2)$。算法总的复杂度为 $O((\delta^2+\delta)m^2+q^2+mq)$，与特征选择比例、原始特征集大小和标签数有关。值得注意的是，步骤 3 和步骤 4 计算得到的标签重要性和特征依赖度项对数据集而言是常量，如果其在运行算法前已知，则算法总的复杂度为 $O((\delta^2+\delta)m^2)$，与原始特征集大小和特征选择比例有关。

4　实验结果与分析

本节首先给出实验数据集来源、实验过程的相关设定和算法性能的评价指标，然后重点进行两方面的实验，包括：

1）分析特征选择比例 δ 对 MLFSLC 算法性能的影响；

2）验证 MLFSLC 算法能够有效提高多标签学习算法的性能，且与 MLFSIE[13]，mRMR[9]，MDMR[15] 等算法相比，其性能更好。

4.1　实验数据集与相关说明

本文在 Emotions，Enron 和 Medical 等常用多标签数据集上进行实验。所有数据均来源于 Mulan Java 开源工程[1)]，各数据的样本总数、特征维数、特征类型、特征维数和特征基数等基本信息如表 1 所列。

表 1　实验所用数据集及部分信息

Table 1　Partial information of datasets utilized in experiments

数据集	样本总数	特征维数	特征类型	标签维数	标签基数	来源领域
Emotions	593	72	连续型	6	1.869	音乐
Enron	1702	1001	离散型	53	3.378	文本
Medical	978	1449	离散型	45	1.245	文本

其中，标签基数(cardinality)表示平均每个样本被赋予多少个标签。令 $|\boldsymbol{X}|$ 表示样本数(下同)，则标签基数的计算如式(21)所示。

$$card=\sum_{i=1}^{|\boldsymbol{X}|}\frac{|\boldsymbol{Y}_i|}{|\boldsymbol{X}|} \tag{21}$$

为验证 MLFSLC 算法的有效性，对实验做如下说明：

1）鉴于计算信息量的需要，参考文献[15]所采用的等宽策略(equal-width strategy)对数据集中的连续型特征进行离散化处理。

2）利用十折交叉验证的方法进行实验。

4.2　评价指标

采用通用的 Hamming Loss，MicroFMeasure 和 Coverage 作为算法性能的评价指标。令 $\boldsymbol{Y}_i^p \subseteq \boldsymbol{L}$ 和 $\boldsymbol{Y}_i \subseteq \boldsymbol{L}$ 分别表示对第 i 个样本预测出的标签子集和真实的标签子集，则结合定义 1，对各评价指标说明如下。

Hamming Loss 用于衡量一个样本被误分给错误标签的平均次数，数值越小，表示算法的性能越好。该指标的计算如式(22)所示。

$$Hamming\ Loss=\frac{1}{|\boldsymbol{X}|}\sum_{i=1}^{|\boldsymbol{X}|}\frac{|\boldsymbol{Y}_i^p \oplus \boldsymbol{Y}_i|}{q} \tag{22}$$

其中，⊕表示异或运算，q 为所有可能的标签数目。

MicroFMeasure 是准确率和召回率的调和平均，能平衡分类系统的查准率和查全率，数值越大，表示算法性能越好。该指标的计算如式(23)所示。

$$MicroFMeasure=\frac{1}{|\boldsymbol{X}|}\sum_{i=1}^{|\boldsymbol{X}|}\frac{2\times|\boldsymbol{Y}_i^p \odot \boldsymbol{Y}_i|}{|\boldsymbol{Y}_i^p|+|\boldsymbol{Y}_i|} \tag{23}$$

其中，⊙表示同或运算。

Coverage 表示依次对排序后的样本预测标签子集进行遍历，得到所有真实标签所需要的最小遍历深度，数值越小，表示算法性能越好。该指标的计算如式(24)所示。

$$Coverage=\frac{1}{|\boldsymbol{X}|}\left[\sum_{i=1}^{|\boldsymbol{X}|}\max_{\lambda \in \boldsymbol{X}_i} rank(\lambda)-1\right] \tag{24}$$

其中，$rank(\lambda)$ 表示 λ 在序列中的排位，若 $\lambda_1>\lambda_2$，则 $rank(\lambda_1)<rank(\lambda_2)$。

4.3　实验结果及分析

实验首先利用 MLFSLC 算法对原始数据集进行特征选择，然后利用常见的 MLkNN[17]，RAkEL[18]，CC[19] 和 HOMER[20] 等算法进行多标签分类，再通过分类结果反映特征选择算法的性能。其中，RAkEL，HOMER 和 CC 算法使用决策树 C4.5 作为基础分类器，MLkNN 算法设定最近邻个数 $k=10$。

4.3.1　特征选择比例 δ 对算法性能的影响

为了充分分析特征选择比例 δ 较高和较低时对 MLFSLC 算法性能的影响，实验分别从 0.05 至 0.1 以 0.01 为步长、从 0.1 至 1.0 以 0.1 为步长来调整参数 δ 的值，分析 3 种评价指标的变化，结果如图 2 所示。

图 2 的各子图中，横轴表示特征选择比例，纵轴表示各评价指标刻度。可以看出，随着特征选择比例 δ 的减小，MLFSLC 算法的性能并非单调变化，当 δ=1 时，其不进行特征选

1) http://mulan.sourceforge.net/datasets-mlc.html

择，直接使用原始数据集进行实验，未取得最佳的分类效果；而且结合 3.4 节的算法复杂度分析可知，此时特征选择算法的复杂度最高；而随着 δ 的减小，用于分类的特征子集的规模也逐渐减小，但当 δ 减小到一定程度时，总能取得最好的分类效果，本文称此时的 δ 值为最佳特征选择比例，并将每种数据集在不同评价指标下的最佳特征选择比例统计于表 2 中。上述事实反映出，通过 MLFSLC 算法选择出的特征子集，能同时提高多标签分类算法的性能和处理效率。

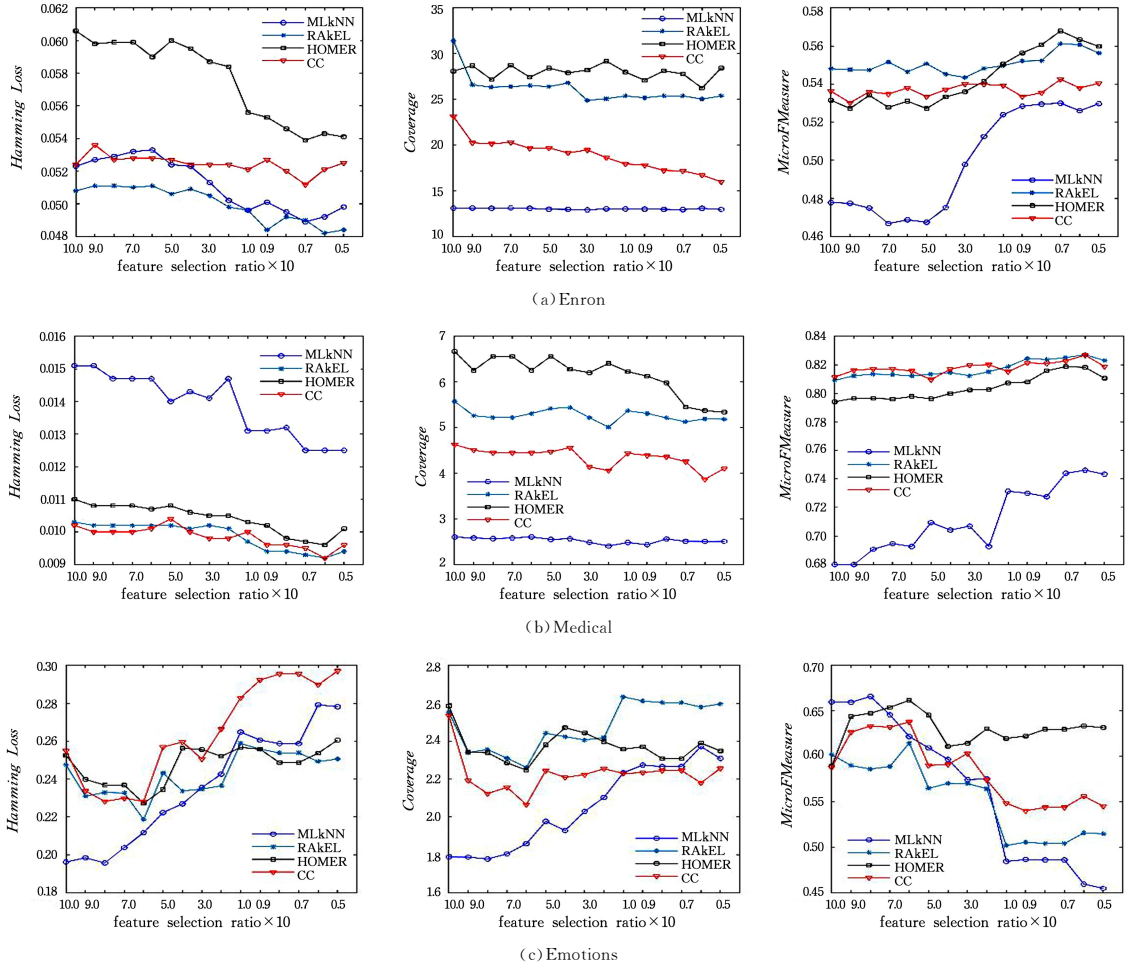

图 2 特征选择比例对算法性能的影响

Fig. 2 Effects of feature selection ratio on performance of algorithm

表 2 不同数据集的最佳特征选择比例

Table 2 Optimal feature selection ratio for different datasets

	Hamming Loss				*Coverage*				*MicroFMeasure*			
	MLkNN	RAkEL	HOMER	CC	MLkNN	RAkEL	HOMER	CC	MLkNN	RAkEL	HOMER	CC
Enron	0.07	0.06	0.07	0.07	0.30	0.30	0.06	0.05	0.07	0.07	0.07	0.07
Medical	0.06	0.06	0.06	0.06	0.20	0.20	0.05	0.06	0.06	0.06	0.07	0.06
Emotions	0.80	0.60	0.60	0.60	0.80	0.60	0.60	0.60	0.80	0.60	0.60	0.60

对于 Enron 和 Medical 数据集，在 Hamming Loss 和 MicroFMeasure 评价指标下，δ 增长至 0.5 以后，分类性能总体逐渐提高，表 2 中，两种数据集分别在 $\delta=0.07$ 和 $\delta=0.06$ 时性能最佳；在 Coverage 评价指标下，对于 MLkNN 和 RAkEL 两种分类器，两种数据集分别在 $\delta=0.3$ 和 $\delta=0.2$ 时性能最佳，而 HOMER 和 CC 分类器的两种数据集在 δ 为 0.05～0.06 时性能最好。对于 Emotions 数据集，MLkNN 分类器的性能在 $\delta=0.8$ 时最佳，而其他分类器的性能在 $\delta=0.6$ 时最好。特征选择算法达到最佳性能时，冗余度、依赖度和信息损失达到一个很好的平衡状态，因此不同数据集有不同的最佳特征选择比例，而 Emotions 数据集的最佳特征选择比例比 Enron 和 Medical 数据集高，这很大程度上是由于连续特征离散化已造成了信息损失。

4.3.2 MLFSLC 算法与其他算法的性能比较

为了验证算法的有效性，将 MLFSLC 与 MLFSIE[13]，mRMR[9] 和 MDMR[15] 等算法进行对比，各性能评价指标的结果分别列于表 3－表 5 中。其中，每个表分别表示 MLkNN，RAkEL，HOMER 以及 CC 4 种分类器在对应性能评价指标下的值，表中数据以“每种算法性能评价指标的最优值/该算法对应的最佳特征选择比例”的形式列出，特征选择比例以百分比(%)为单位，实验中各评价指标的最优值和各特征选择比例的最小值分别加粗表示。

表 3　不同分类器下 4 种特征选择算法的 *Hamming Loss* 值比较

Table 3　Comparison of *Hamming Loss* values between four feature selection algorithms under different classifiers

	MLkNN			RAkEL		
	Enron	Medical	Emotions	Enron	Medical	Emotions
MLFSLC	**0.0489/7**	**0.0125/6**	**0.1957**/80	**0.0482/6**	**0.0092/6**	**0.2186**/60
MLFSIE	0.0518/27	0.0130/7.8	0.1966/**19.4**	0.0494/27	0.0094/7.8	0.2192/19.4
mRMR	0.0511/20	0.0132/7	0.1959/90	0.0497/80	0.0093/7	0.2195/**7**
MDMR	0.0509/20	0.0133/8	0.1982/80	0.0493/40	0.0096/7	0.2195/**7**
	HOMER			CC		
	Enron	Medical	Emotions	Enron	Medical	Emotions
MLFSLC	**0.0539/7**	**0.0096/6**	**0.2272**/60	**0.0517/7**	**0.0092/6**	**0.2282**/60
MLFSIE	0.0565/27	0.0099/7.8	0.2284/19.4	0.0521/27	0.0102/7.8	0.2291/19.4
mRMR	0.0572/20	0.0102/7	0.2313/**7**	0.0521/80	0.0095/7	0.2284/50
MDMR	0.0562/10	0.0102/7	0.2274/10	0.0522/10	0.0097/7	0.2338/**7**

表 4　不同分类器下 4 种特征选择算法的 *Coverage* 值比较

Table 4　Comparison of *Coverage* values between four feature selection algorithms under different classifiers

	MLkNN			RAkEL		
	Enron	Medical	Emotions	Enron	Medical	Emotions
MLFSLC	**12.9057**/30	**2.4057**/20	**1.7778**/80	**24.8904**/30	**5.0088**/20	**2.2615**/60
MLFSIE	12.9653/27	2.4388/**7.8**	1.7923/**19.4**	27.7152/**27**	5.3184/**7.8**	2.4730/19.4
mRMR	13.0217/60	2.4651/20	1.8114/90	32.0431/80	5.6689/20	2.5977/50
MDMR	12.9084/**10**	2.4456/20	1.7877/90	26.2721/80	5.2775/70	2.5297/**7**
	HOMER			CC		
	Enron	Medical	Emotions	Enron	Medical	Emotions
MLFSLC	**26.2354/6**	**5.3398/5**	2.2468/60	15.9758/**5**	**3.8680/6**	**2.0651**/60
MLFSIE	26.2677/27	5.8759/7.8	2.2479/19.4	16.0019/27	4.2050/7.8	2.0709/19.4
mRMR	26.2632/7	6.1007/7	2.2597/**6**	16.2892/**5**	4.5214/**6**	2.0831/9
MDMR	26.2744/9	6.0971/7	**2.2427/6**	**15.7406/5**	4.2256/**6**	2.0674/**5**

表 5　不同分类器下 4 种特征选择算法的 *MicroFMeasure* 值比较

Table 5　Comparison of *MicroFMeasure* values between four feature selection algorithms under different classifiers

	MLkNN			RAkEL		
	Enron	Medical	Emotions	Enron	Medical	Emotions
MLFSLC	**0.5300/7**	**0.7461/6**	**0.6659**/80	**0.5612/7**	**0.8271/6**	**0.6147**/60
MLFSIE	0.4950/27	0.7339/7.8	0.6635/**19.4**	0.5577/27	0.8242/7.8	0.6139/19.4
mRMR	0.5024/20	0.7278/7	0.6649/80	0.5543/80	0.8246/7	0.6135/**7**
MDMR	0.5051/20	0.7253/8	0.6596/80	0.5577/40	0.8209/10	0.6135/**7**
	HOMER			CC		
	Enron	Medical	Emotions	Enron	Medical	Emotions
MLFSLC	**0.5679/7**	**0.8188/7**	**0.6617**/60	0.5425/**7**	**0.8268/6**	**0.6379**/60
MLFSIE	0.5545/27	0.8142/7.8	0.6567/19.4	0.5396/27	0.8132/7.8	0.6313/19.4
mRMR	0.5391/40	0.8095/**7**	0.6525/**7**	0.5383/80	0.8238/7	0.6250/50
MDMR	0.5551/9	0.8090/**7**	0.6587/10	**0.5432**/80	0.8185/7	0.6364/**5**

从表 3－表 5 可以看出，一方面，与其他特征选择算法相比，除 Emotions 数据集在 HOMER 分类器下的 Coverage 指标以及 Enron 数据集在 CC 分类器下的 Coverage 和 MicroFMeasure 评价指标逊于 MDMR 外，MLFSLC 在各评价指标下都体现出了更好的分类效果；另一方面，由于 Emotions 数据集具有连续型特征，利用 MLFSLC 算法进行特征选择时需要对其连续特征进行离散化处理，这将导致其最佳特征选择比例较其他特征选择算法高；而对于 Enron 和 Medical 数据集，除在 MLkNN 和 RAkEL 分类器的 Coverage 评价指标外，其最佳 MLFSLC 特征选择比例都低于其他特征选择算法，更低的最佳特征选择比例意味着可以用规模更小的特征子集实现更好的分类效果，从而达到更高的分类效率。以上结果表明，充分考虑了数据特征集与标签集关系的 MLFSLC 算法能够有效提高多标签分类算法的性能和基于离散特征的多标签分类算法的效率。

结束语　当前的多标签特征选择算法虽然能有效应对数据的维度灾难问题，但大多忽略了标签之间的相互关系和由不同数据分布引起的信息量衡量范围的偏差。基于此，本文的多标签特征选择算法 MLFSLC 首先定义了标签的重要性，并以此为权重对特征依赖度和冗余度加权，然后提出一种规范化的特征评分函数，并据此依次选择出最佳特征子集。实验验证了本文所提特征选择算法能够有效提高多标签学习算法的效率。但是由于互信息量计算方法的限制，算法的处理对象是离散型特征变量，在对具有连续型特征变量的数据集进行离散化处理的过程中，会对数据结构造成破坏，影响分类结果。此外，如何通过设定阈值自适应地选择最佳特征选择比例，也是本算法有待改进的方向之一。

参 考 文 献

[1]　WU X，ZHU X，WU G Q，et al. Data mining with big data[J]. IEEE Transactions on Knowledge and Data Engineering，2014，26(1)：97-107.

[2] ZHANG J J,FANG M,LI X. Multi-label learning with discriminative features for each label[J]. Neurocomputing,2015,154:305-316.

[3] JIANG S,WANG L. Efficient feature selection based on correlation measure between continuous and discrete features[J]. Information Processing Letters,2016,116(2):203-215.

[4] ZHANG Y X,SUN Y,YANG J H,et al. Feature importance analysis for spammer detection in SinaWeibo[J]. Journal on Communications,2016,37(8):24-33. (in Chinese)
张宇翔,孙菀,杨家海,等. 新浪微博反垃圾中特征选择的重要性分析[J]. 通信学报,2016,37(8):24-33.

[5] XIE J Y,XIE W X. Several Feature Selection Algorithms Based on the Discernibility of a Feature Subset and Support Vector Machines[J]. Chinese Journal of Computers,2014,37(8):1704-1718. (in Chinese)
谢娟英,谢维信. 基于特征子集区分度与支持向量机的特征选择算法[J]. 计算机学报,2014,37(8):1704-1718.

[6] LIU H,LI X,ZHANG S. Learning instance correlation functions for multilabel classification[J]. IEEE Transactions on Cybernetics,2017,47(2):499-510.

[7] TANG J L,ALELYANI S,LIU H. Feature selection for classification:A review[M]// Data Classification:Algorithms and Applications. CRC Press,Chapman,2014:313-334.

[8] SILVA A M D,LEONG P H W. Grammar-based feature generation for time-series prediction[M]. Singapore:Springer Singapore,2015:13-23.

[9] PENG H,LONG F,DING C. Feature selection based on mutual information criteria of max-dependency,max-relevance,and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2005,27(8):1226-1238.

[10] SHAO H,LI G Z,LIU G P,et al. Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine[J]. Science China Information Sciences,2013,56(5):1-13.

[11] YOU M,LIU J,LI G Z,et al. Embedded feature selection for multi-label classification of music emotions[J]. International Journal of Computational Intelligence Systems,2012,5(4):668-678.

[12] DOQUIRE G,VERLEYSEN M. Mutual information-based feature selection for multi-label classification[J]. Neurocomputing,2013,122:148-155.

[13] ZHANG Z H,LI S N,LI Z G,et al. Multi-Label Feature Selection Algorithm Based on Information Entropy[J]. Journal of Computer Research and Development,2013,50(6):1177-1184. (in Chinese)
张振海,李士宁,李志刚,等. 一类基于信息熵的多标签特征选择算法[J]. 计算机研究与发展,2013,50(6):1177-1184.

[14] MANDAL M,MUKHOPADHYAY A. An improved minimum redundancy maximum relevance approach for feature selection in gene expression data[J]. Procedia Technology,2013,10(1):20-27.

[15] LIN Y,HU Q,LIU J,et al. Multi-label feature selection based on max-dependency and min-redundancy[J]. Neurocomputing,2015,168(C):92-103.

[16] WITTEN I H,FRANK E,HALL M A,et al.. Data mining:Practical machine learning tools and techniques[M]. Burlington:Morgan Kaufmann,2016:143-186.

[17] ZHANG M L,ZHOU Z H. ML-KNN:A lazy learning approach to multi-label learning[J]. Pattern Recognition,2007,40(7):2038-2048.

[18] TSOUMAKAS G,KATAKIS I,VLAHAVAS I. Random k-labelsets for multilabelclassification[J]. IEEE Transactions on Knowledge and Data Engineering,2011,23(7):1079-1089.

[19] READ J,PFAHRINGER B,HOLMES G,et al. Classifier chains for multi-label classification[J]. Machine Learning,2009,85(3):254-269.

[20] TSOUMAKAS G,KATAKIS I,VLAHAVAS I. Effective and efficient multilabel classification in domains with large number of labels[C]// Proccessing of ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08). Antwerp,Belgium,2008:30-44.

(上接第 207 页)

[10] MIRZA P. Extracting Temporal and Causal Relations between Events[C]// Proceedings of the ACL 2014 Student Research Workshop. Baltimore,Maryland USA,2014:10-17.

[11] MIRZA P,TONELLI S. CATENA:CAusal and TEmporal relation extraction from NAtural language texts[C]// The 26th International Conference on Computational Linguistics. 2016:64-75.

[12] MANI I,VERHAGEN M,WELLNER B,et al. Machine learning of temporal relations[C]// Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,2006:753-760.

[13] CHAMBERS N,WANG S,JURAFSKY D. Classifying temporal relations between events[C]// Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics,2007:173-176.

[14] CHAMBERS N,DAN J. Jointly Combining Implicit Constraints Improves Temporal Ordering[C]// Conference on Empirical Methods in Natural Language Processing. 2009:698-706.

[15] ZHENG X,LI P F,ZHU Q M. Global Inference for Temporal Relations Between Chinese Events[J]. Journal of Chinese Information Processing,2016,30(5):129-135. (in Chinese)
郑新,李培峰,朱巧明. 基于全局优化的中文事件时序关系推理方法[J]. 中文信息学报,2016,30(5):129-135.

[16] HUANG Y L,LI P F,ZHU Q M. The Construction of Chinese Relevant Event Corpus and Its Recognition Approach[J]. Computer Engineering and Science,2015,37(12):2306-2311. (in Chinese)
黄一龙,李培峰,朱巧明. 中文事件相关性语料库构建及识别方法[J]. 计算机工程与科学,2015,37(12):2306-2311.

[17] ZHENG X,LI P F,ZHU Q M. Annotation and Classification of Temporal Relation Between Chinese Events[J]. Computer Science,2015,42(7):276-279. (in Chinese)
郑新,李培峰,朱巧明,等. 中文事件时序关系的标注和分类方法[J]. 计算机科学,2015,42(7):276-279.