计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
50(6): 1177-1184, 2013

一类基于信息熵的多标签特征选择算法

张振海　李士宁　李志刚　陈　昊
(西北工业大学计算机学院　西安　710072)
(htty326@163.com)

Multi-Label Feature Selection Algorithm Based on Information Entropy

Zhang Zhenhai, Li Shining, Li Zhigang, and Chen Hao
(*School of Computer Science, Northwestern Polytechnical University, Xi'an* 710072)

Abstract　Multi-label classification is the learning problem where each instance is associated with a set of labels. Feature selection is capable of eliminating redundant and irrelevant features in multi-label classification, which leads to performance improvement of multi-label classifiers. However the existing feature selection methods have high computation complexity and are not able to give a reasonable feature subset. Hence a novel multi-label feature selection algorithm based on information entropy is proposed in this paper. It assumes that features are independent of each other. Its main ideas are: 1) The information gain between the feature and label set is derived from the information gain between the feature and the label, and employed to measure the correlation degree between them; 2) An threshold selection method is used to choose a reasonable feature subset from original features. The proposed algorithm firstly computes the information gain between each feature and label set, and then removes the irrelevant and redundant features according to the selected information gain value determined by threshold selection method. The experiment is conducted on four different datasets and two different classifiers. The experimental results and their analysis show that the proposed algorithm can effectively promote the performance of multi-label classifiers in multi-label classification.

Key words　Internet of things; data processing; information theory; multi-label classification; feature selection; information gain; dimensionality reduction

摘　要　在多标签分类问题中,特征选择是提升多标签分类器性能的一种重要手段.针对目前多标签特征选择算法计算复杂度大和无法给出一个合理的特征子集的问题,提出了一种基于信息熵的多标签特征选择算法.该算法假设特征之间相互独立,使用特征与标签集合之间的信息增益来衡量特征与标签集合之间的重要程度,并据此提出一种信息增益阈值选择方法.首先计算每一个特征与标签集合之间的信息增益,然后使用信息增益阈值选择算法得到一个合理的阈值,最后根据阈值删除不相关的特征,得到一组合理的特征子集.在2个不同分类器和4个多标签数据集上的实验结果表明:特征选择算法能够有效地提升多标签分类器的分类性能.

关键词　物联网;数据处理;信息论;多标签分类;特征选择;信息增益;特征降维

中图法分类号　TP18; TP391

收稿日期:2012-12-31;**修回日期**:2013-04-08
基金项目:"核高基"国家科技重大专项基金项目(2012ZX03005007)
通信作者:李士宁(lishining@nwpu.edu.cn)

近年来,随着物联网的发展和相关应用的不断深入,产生的数据越来越多,形式也多种多样.智能信息处理技术作为物联网的关键技术之一,可以从大量的数据中挖掘有用的信息,实现对物理世界的精准管理和作出科学的决策.分类是实现物联网信息智能化处理的一种重要手段.传统的分类主要研究单标签分类问题,目标是将一个实例准确地划分到某一个(类别)标签中,然而这已无法满足物联网中相关的实际应用需求.例如,在精准农业应用中,需要通过传感器采集回来的土壤温度、土壤湿度、土壤中氮等营养成分的浓度、空气湿度、二氧化碳浓度等信息作出是否需要开启灌水开关、农作物生长环境是否出现异常、是否需要施肥、是否需要开启风扇进行降温等决策,像这类将一个实例(土壤温度等数据形成的特征向量)同时划分到多个(类别)标签中的分类问题称为多标签分类问题.多标签分类和单标签分类的区别是:单标签分类中的每个实例只能属于一个标签,而多标签分类中的一个实例可以同时属于多个标签,即一个标签集合.近年来,多标签分类引起了广泛的关注[1-4].

在多标签数据中,维数灾难会严重影响多标签分类器的分类性能[5].特征降维技术可以降低特征维数,提高分类器的分类性能.在单标签分类中,特征降维已经得到了广泛的研究,而目前针对多标签分类中的特征降维技术的研究还比较少.特征降维可以分为特征提取和特征选择.本文主要研究多标签分类问题中的特征选择问题.

多标签分类问题中的每一个实例可以同时划分到多个标签中,并且标签之间具有相关性,这使得多标签特特征选择问题比单标签的特征选择问题更复杂.例如,在单标签特征选择中用于表示特征与标签之间关系的方法在多标签特征选择中已不适用,需要找出一种有效的方法来表示特征与标签集合之间的关系.Li 等人和 You 等人[6-7]提出了一种嵌入式特征选择算法,以预报风险作为评价特征的准则,对于每一个特征,计算特征属性值被它的平均值代替前后的分类精度的差值,差值最小的特征将被删除.Shao 等人[8]提出了一种混合优化的特征选择算法(hybrid optimization based multi-label feature selection, HOML),该算法综合了模拟退火算法和遗传算法以及贪婪算法寻找最优特征子集.这两种方法直接使用分类器来评价特征与标签集合之间的关系,使得算法的计算复杂度较大,在实际中难以应用.Lee 等人[9]使用特征集合包含某个特征与否之间的信息熵之差来表示该特征的重要性,通过信息熵之差最大化和正向搜索的方法选择特征子集;相对于前两种方法,该算法的计算复杂度明显减小,然而该算法没有解决信息熵阈值的选择的问题,因此无法给出一个合理的特征子集,不利于实际应用.

为此,本文提出一种基于信息熵的多标签特征选择算法(multi-label feature selection algorithm based on information entropy, MLFSIE).该算法的主要思想是:采用特征与标签集合的信息增益来衡量一个特征与标签集合之间的关系,并提出一种信息增益阈值选择方法;首先计算每一个特征与标签集合的信息增益值,然后根据信息增益阈值选择方法设定的阈值删除不相关的特征,达到降维的目的.MLFSIE 算法处理对象是离散型特征变量,并且算法的复杂度不依赖于任何分类器,实验结果表明,本文所提出的特征选择算法能够有效地提升多标签分类器的预测性能.

本文的主要贡献有:1)提出使用特征与标签集合的信息增益来衡量一个特征对标签集合的重要程度;2)提出一种信息增益阈值选择方法,提高特征选择算法的适应能力,便于实际应用.

1 相关工作

特征降维包括特征提取和特征选择.特征提取是将特征从高维空间映射到低维空间,对原始的特征信息进行融合.在多标签分类的特征提取方面,近年来已有相关研究成果.Zhang 等人[10]提出了一种多标签分类方法,该方法集成了特征提取和特征选择方法,首先使用主成分分析(principal component analysis, PCA)对原始数据进行特征提取,接着在提取出来的特征上结合汉明损失(Hamming loss)和次序损失(rank loss)作为适应度函数,使用遗传算法选择最优特征子集.Zhang 等人[5]提出了一种基于最大依赖维的多标签特征提取方法,该方法采用希尔伯特-斯密特准则(Hilbert-Schmidt independence criterion, HSIC)作为依赖性的评价标准,通过最大化数据集特征与类标签之间的依赖性,以有监督的方式将数据集特征从高维空间映射到低维空间中,从而达到降维的目的.Park 等人[11]通过对多标签数据集进行分解等方式对单标签问题中的降维方法线性判别分析(linear discriminant analysis, LDA)进行了改进和扩展,使其能够处理多标签中的特征降维问题.Yu 等人[12]通过在目标函数中加入标签优

化的因素将单标签中的无监督的降维方法潜在语义索引(latent semantic indexing, LSI)扩展成可以处理多标签问题的有监督降维方法. Ji 等人[13]提出了一种线性降维方法,该方法将降维过程和分类过程结合在一起,为每一个标签计算一个线性函数,将最小二乘损失(least squares loss)和铰链损失(hinge loss)作为目标函数的因素,通过最小化目标函数找到合适的特征. 然而,经过特征提取处理后的特征已不是原始的特征,没有原始特征的物理含义,这对理解所研究的问题造成了很大的困难.

特征选择是根据某一评价准则从原始特征中去除冗余或者不相关的特征,得到一组使得评价准则最优的特征子集. 特征选择的结果能够保持原始特征的物理含义,可以得到低维的原始特征子集,让人很容易理解特征在所研究问题中的意义. 特征选择方法近年来也有一些相关研究[6-8]. 然而,这些特征选择算法对特征子集的评价与具体的分类器紧密相关,因此特征选择算法的结果和计算复杂度也依赖于具体的分类器,并且计算复杂度通常比较大.

2 多标签问题描述和信息增益

2.1 多标签问题描述

为了对本文的特征选择算法进行描述,统一概念,首先给出多标签问题的形式化描述.

定义 1. 设一个实例 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$,其中 $x_i\in\mathbb{R}$;候选标签集合 $L=\{l_1,l_2,\cdots,l_m\}$,该实例所对应的标签集合 $Y=\{l_1,l_2,\cdots,l_p\}$,$p\leqslant m$,即 $Y\subseteq L$;则包含 t 个样本的多标签数据集可表示为

$$D=\{(\boldsymbol{X}_i,Y_i)\mid 1\leqslant i\leqslant t,X_i\in\mathbb{R}^n,\ Y_i\subseteq L\},\tag{1}$$

其中,X_i 表示一个实例,也称为样本的特征向量,Y_i 表示该实例对应的标签集合.

2.2 信息增益[14]

定义 2. 设集合 $A=\{a_1,a_2,\cdots,a_m\}$,$p(a_i)$ 为元素 a_i 的先验概率,则称

$$H(A)=-\sum_{i=1}^{n}p(a_i)\log_r p(a_i)\tag{2}$$

为集合 A 的信息熵. 信息熵的值越大说明集合的不确定程度越大.

定义 3. 设集合 $A=\{a_1,a_2,\cdots,a_m\}$,集合 $B=\{b_1,b_2,\cdots,b_n\}$,则在给定集合 A 的条件下集合 B 的条件熵为

$$H(B\mid A)=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(a_ib_j)\log_r p(b_j\mid a_i).\tag{3}$$

条件熵用于衡量在集合 A 出现的条件下集合 B 的不确定程度的大小.

定义 4. 设集合 $A=\{a_1,a_2,\cdots,a_m\}$,集合 $B=\{b_1,b_2,\cdots,b_n\}$,则集合 A 和集合 B 的联合熵为

$$H(AB)=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(a_ib_j)\log_r p(a_ib_j),\tag{4}$$

并且与信息熵、条件熵满足以下关系:

$$H(AB)=H(A)+H(B\mid A)=H(B)+H(A\mid B).\tag{5}$$

定义 5. 设集合 $A=\{a_1,a_2,\cdots,a_m\}$,集合 $B=\{b_1,b_2,\cdots,b_n\}$,则信息增益为

$$IG(B\mid A)=H(B)-H(B\mid A).\tag{6}$$

信息增益用于衡量集合 A 对集合 B 的相关程度的大小,信息增益值越大,说明集合 A 对集合 B 的相关程度越大. 由信息论理论可知 $H(B|A)\geqslant 0$,因此 $IG(B|A)\leqslant H(B)$. 同理,$IG(A|B)\leqslant H(A)$.

将式(5)代入式(6)可以得到:

$$IG(B\mid A)=H(A)+H(B)-H(AB).\tag{7}$$

信息增益具有如下性质[15]:

性质 1. 如果集合 A 和集合 B 相互独立,则信息增益取最小值.

性质 2. 如果集合 A 完全由集合 B 决定,则信息增益取最大值.

由于信息增益具有上述两个性质,可以有效地区分特征与标签之间的关系,因此 MLFSIE 算法使用信息增益来研究特征与标签集合之间的相关性.

3 本文的多标签特征选择算法

在特征选择中,如何衡量特征与标签之间的关系是一个重要的问题,只有明确了特征与标签之间的关系才能找出令人满意的特征子集. 在传统的单标签问题中,可以直接使用信息增益衡量一个特征与标签之间的相关性. 而在多标签问题中,一个特征不仅与某一个标签相关,还可能与多个标签相关,为此需要在信息增益的基础上研究一个特征与标签集合之间的关系.

本节首先介绍特征与标签集合的相关性,然后对算法 MLFSIE 进行描述.

3.1 特征与标签集合之间的相关性

定义 6. 给定一个特征 x 和标签集合 $L=\{l_1,l_2,\cdots,l_m\}$,令 $IG(l_i\mid x)$ 为特征 x 与标签 l_i 的信息

增益,则称

$$IGS(Y \mid x) = \sum_{i=1}^{m} IG(l_i \mid x) \tag{8}$$

为特征 x 与标签集合 L 的信息增益. 由于 $IG(l_i \mid x) \geqslant 0$,因此 $IGS(L \mid x) \geqslant 0$. 由信息论的理论可知,$IGS(L \mid x) \leqslant \sum_{i=1}^{m} H(l_i)$.

特征与标签集合的信息增益具有以下性质.

性质 3. 如果一个特征 x 与标签集合 $L=\{l_1, l_2, \cdots, l_m\}$ 中的每一个标签都是相互独立的,则特征 x 与标签集合 L 的信息增益取最小值.

证明. 由性质 1 以及信息增益的非负性可知,如果标签 l_i 与特征 x 是相互独立的,则信息增益 $IG(l_i \mid x)=0$,结合定义 6 可知此时 $IGS(L \mid x)=0$. 又因为 $IGS(L \mid x) \geqslant 0$,故特征 x 与标签集合 L 的信息增益取值最小. 证毕.

性质 4. 如果标签集合 $L=\{l_1, l_2, \cdots, l_m\}$ 中的每一个标签 l_i 均由特征 x 完全决定,则特征 x 与标签集合 L 的信息增益取最大值.

证明. 由性质 2 及定义 5 的分析可知,如果标签 l_i 完全由特征 x 决定,则 $IG(l_i \mid x)=H(l_i)$,结合定义 6 可得 $IGS(L \mid x) \leqslant \sum_{i=1}^{m} H(l_i)$. 由于 $IGS(L \mid x) \leqslant \sum_{i=1}^{m} H(l_i)$, 故特征 x 与标签集合 L 的信息增益取最大值. 证毕.

由性质 3 可知,$IGS(L \mid x)$可以找出与对标签集合 L 毫无作用的特征. 性质 3 和性质 4 说明了设定一个合理的阈值,可以去除与标签集合 L 相关性不大的特征.

在给出特征与标签集合的相关性定义之前,为了方便计算阈值,首先对不同特征 x 与标签集合 L 的信息增益进行变换,假设信息增益的分布服从正态分布,通过

$$IGZ(L \mid x) = \frac{IGS(L \mid x) - \mu}{\sigma} \tag{9}$$

将其变换成标准的正态分布,其中 μ 表示特征与标签集合的信息增益的均值;σ 表示特征与标签集合的信息增益的标准差. 在标准变换的基础上给出特征与标签集合的相关性定义.

定义 7. 给定一个阈值 $\delta>0$,如果变换后的特征 x 与标签集合 L 的信息增益值的绝对值 $|IGZ(L \mid x)| \geqslant \delta$,则称特征 x 与标签集合是相关的,否则为不相关.

在 MLFSIE 中,根据定义 7,与标签集合不相关的特征将会被删除,相关的特征被留下.

3.2 归一化信息增益

为了使得各个特征与标签的信息增益有相同的衡量范围,在计算特征与标签集合的信息增益之前,首先对特征与标签的信息增益进行归一化处理. Yu 等人[16]证明了信息增益具有对称性,即 $IG(B \mid A) = IG(A \mid B)$,并且 $IG(B \mid A) \leqslant H(B)$, $IG(A \mid B) \leqslant H(A)$. 使用

$$SU(A,B) = 2\left[\frac{IG(B \mid A)}{H(A)+H(B)}\right] \tag{10}$$

对特征与标签之间的信息增益 $IG(B \mid A)$进行归一化处理. $SU(A,B) \in [0,1]$,如果 $SU(A,B)=0$,表示 A 和 B 是相互独立的,$SU(A,B)=1$ 表示可以通过 A 和 B 当中的一个确定另外一个的值.

3.3 阈值选择

由 $IGS(L \mid x) = \sum_{i=1}^{m} IG(l_i \mid x)$ 可知,特征与标签集合的信息增益和标签数量相关,而不同数据集的标签数量不一样,因此设定某一个特定的信息增益阈值无法保证特征选择算法在不同的应用场合中都能得出一个合理的特征子集,为了解决这个问题,需要设计一种能够根据不同应用场合自动计算信息增益阈值的方法.

在计算特征与标签集合的信息增益过程中,假设候选标签数量 $m=20$,对于 2 个不同的特征 x_1 和 x_2,有 $IGS(L \mid x_1)=0.5$, $IGS(L \mid x_2)=1.6$,根据 3.1 节的描述可以得出特征 x_2 的重要性要大于特征 x_1,因此会选择留下特征 x_2. 然而在极端情况下:如果 $IG(l_1 \mid x_1)=0.5$,并且对于余下的 $i \in [2,20]$ 有 $IG(l_i \mid x_1)=0$,同时对于 $i \in [1,20]$ 有 $IG(l_i \mid x_2) = 1.6/20 = 0.08$,此时可以明显地看出特征 x_1 对标签 l_1 的识别能力远大于特征 x_2,而特征 x_2 对于任何一个标签 l_i 的识别能力都比较弱,因此留下特征 x_1 是一种更明智的选择. 为了减少上述情况的出现,在标准正态变换的基础上采用

$$\delta = \frac{1}{n}\sum_{i=1}^{n} |IGZ(L \mid x_i)| \tag{11}$$

对阈值进行设定,其中,$|IGZ(L \mid x_i)|$表示信息增益 $IGZ(L \mid x_i)$的绝对值. 式(11)采用变换后的信息增益绝对值的数学期望作为当前的阈值,增加了算法的自适应能力.

3.4 算法描述

MLFSIE 算法使用特征与标签集合的信息增益来衡量特征与标签集合的相关性,假设特征之间相互独立,因此只需要计算单个特征与标签集合之间的关系. 主要思想为:计算每一个特征与标签集合的

信息增益,并根据阈值选择方法自适应地设定阈值 δ,如果信息增益小于设定的阈值 δ,则该特征与标签集合是不相关的,删除该特征,否则保留该特征.

算法 1. 基于信息熵的多标签特征选择算法 MLFSIE.

输入:多标签数据集 D;

输出:降维后的多标签数据集 D'.

① $IGS=\varnothing$;
② for each $x_i \in X$
③ $S=\varnothing$;
④ for each $l_j \in L$
⑤ 根据式(7)计算 $IG(l_j|x_i)$;
⑥ 根据式(10)对 $IG(l_j|x_i)$进行归一化,得 s_j;
⑦ $S=S\bigcup\{s_j\}$;
⑧ end
⑨ 根据 $S=\{s_1,s_2,\cdots,s_m\}$,用式(8)计算 IGS_i;
⑩ $IGS=IGS\bigcup\{IGS_i\}$;
⑪ end
⑫ 根据 $IGS=\{IGS_1,IGS_2,\cdots,IGS_n\}$,用式(9)进行变换,得 $IGZ=\{IGZ_1,IGZ_2,\cdots,IGZ_n\}$
⑬ $\delta=\frac{1}{n}\sum_{i=1}^{n}|IGZ_i|$;
⑭ for each $IGZ_i \in IGZ$
⑮ if $|IGZ_i|<\delta$ then $X=X-\{x_i\}$;
⑯ end
⑰ 输出删除不相关特征后的多标签数据集 D'.

在算法 1 中,步骤④~⑧计算一个特征与每一个标签的信息增益并进行归一化处理,得到包含 m 个元素的集合 S,其中 m 为定义 1 中候选标签集合的数量,集合中的元素 s_j 是特征与标签 l_j 经过归一化后的信息增益. 步骤②~⑪计算每一个特征与标签集合的信息增益,得到包含 n 个元素的集合 IGS,其中 n 为定义 1 中特征的数量. 步骤⑫将集合 IGS 变换成标准正态分布的集合 IGZ,步骤⑬根据集合 IGZ 自适应地设定阈值,最后步骤⑭~⑰根据阈值删除不相关的特征.

3.5 算法性能分析

在 MLFSIE 算法中,步骤②~⑪计算每一个特征与标签集合的信息增益,计算复杂度为 $O(n\times m)$;步骤⑫是对每个特征与标签集合的信息增益进行变换,计算复杂度为 $O(n)$;步骤⑬计算阈值,计算复杂度为 $O(n)$;步骤⑭~⑰根据阈值删除不相关特征,计算复杂度为 $O(n)$. 由上述分析可知 MLFSIE 算法的计算复杂度为 $O(n\times m)$,其中 n 为特征数量,m 为候选标签数量.

根据上述分析可知,多标签特征选择算法 MLFSIE 的计算复杂度只与特征数量和候选标签数量有关,并不依赖于任何具体的分类器.

4 实 验

为了评测 MLFSIE 算法的性能,本文在 Mulan[17] 平台上实现了 MLFSIE 算法,Mulan 是一个基于 Weka[18] 的开源项目. 实验将 MLFSIE 与 PCA[19] 和 LP-CHI[20] 进行对比,采用分类算法 LP[2](label powerset)和 MLkNN[21](multi-label k-nearest neighbor)对降维后的数据集进行验证. PCA 是一种经典有效的无监督降维方法,可以直接对多标签数据进行降维. LP-CHI 在降维的同时考虑标签之间的关系,在实际应用中是一种效果比较好的有监督降维方法. LP 和 MLkNN 是两种总体性能比较好的多标签分类器,在相关研究中经常被用来进行对比. 实验中 PCA 的方差比例参数设置为 0.95,LP-CHI 中保留的特征维数与 MLFSIE 降维后的维数相同,MLkNN 中的 k 设置为 10,LP 使用 C4.5[14] 决策树作为基础分类器. 实验过程首先使用特征降维算法对数据集进行降维,然后使用分类算法对降维后的数据集以 10 次交叉的方式进行验证.

4.1 数据集

采用数据集 enron,medical,corel5k 和 genbase① 作为实验数据集,相关信息如表 1 所示:

Table 1 Multi-Label Datasets and Their Statistics
表 1 数据集的相关信息

Name	Features	Labels	Instances	Cardinality
enron	1 001	53	1 702	3.378
medical	1 449	45	978	1.245
corel5k	499	374	5 000	3.522
genbase	1 186	27	662	1.252

4.2 评价指标

本文采用 *MicroFMeasure* 和 *HammingLoss* 作为分类性能的评价指标. 令分类器对一个未知实例的预测标签集合为 Z,结合定义 1,各个评价指标的定义如下.

① 数据集均来自 http://mulan.sourceforge.net/datasets.html

MicroFMeasure[2]是准确率和召回率的调和平均数,定义为

$$MicroFMeasure = \frac{2 \times tp}{2 \times tp + fp + fn}, \quad (12)$$

其中,tp,fp 和 fn 分别表示所有特征向量中所有标签 true positives,false positives 和 false negatives 的数目. *MicroFMeasure* 的值越大表明分类的性能越好.

HammingLoss[2]用于评估一个实例被误分的平均次数,定义为

$$HammingLoss = \frac{1}{t}\sum_{i=1}^{t}\frac{|Y_i \Delta Z_i|}{m}, \quad (13)$$

其中,Δ 表示 2 个集合之间对应元素的差异. 该指标的值越小说明分类的性能越好.

4.3 实验结果及分析

实验结果包括特征降维算法在 LP 和 ML*k*NN 上的性能评测结果. 结果采用评价指标值的平均值和标准差表示. 在表 2~5 中,ALL 表示采用所有特征,MLFSIE,PCA 和 LP-CHI 分别表示相应特征降维算法降维后的特征. 符号 ↑ 表示指标的值越大分类性能越好;符号 ↓ 表示指标的值越小分类性能越好. 加粗的数字表示最优值.

Table 2 Comparative Results of LP Classifier in Terms of *MicroFMeasure*(↑)

表 2 降维后 LP 分类器的 *MicroFMeasure* 值比较(↑)

Datasets	MLFSIE	PCA	LP-CHI	ALL
enron	**0.440 3±0.026 2**	0.426 7±0.022 0	0.405 4±0.027 8	0.430 8±0.022 6
medical	**0.786 2±0.028 3**	0.534 7±0.026 0	0.776 6±0.038 7	0.752 6±0.027 0
core15k	**0.110 8±0.008 1**	0.107 6±0.007 6	0.105 9±0.013 0	0.109 7±0.007 5
genbase	**0.981 4±0.018 1**	0.965 3±0.022 5	0.981 4±0.018 1	0.980 1±0.017 6

Table 3 Comparative Results of LP Classifier in Terms of *HammingLoss*(↓)

表 3 降维后 LP 分类器的 *HammingLoss* 值比较(↓)

Datasets	MLFSIE	PCA	LP-CHI	ALL
enron	**0.070 4±0.003 6**	0.072 2±0.002 7	0.070 9±0.002 9	0.071 7±0.002 8
medical	**0.011 5±0.001 6**	0.025 5±0.001 6	0.012 1±0.002 3	0.013 5±0.001 6
core15k	**0.016 6±0.000 2**	0.016 8±0.000 2	0.016 7±0.000 3	0.016 8±0.000 2
genbase	**0.001 8±0.001 9**	0.003 2±0.002 1	0.001 8±0.001 9	0.001 9±0.001 9

Table 4 Comparative Results of ML*k*NN Classifier in Terms of *MicroFMeasure*(↑)

表 4 降维后 ML*k*NN 分类器的 *MicroFMeasure* 值比较(↑)

Datasets	MLFSIE	PCA	LP-CHI	ALL
enron	**0.478 0±0.028 8**	0.312 5±0.035 5	0.355 8±0.019 9	0.477 8±0.022 6
medical	**0.733 9±0.035 8**	0.582 6±0.038 9	0.728 1±0.032 8	0.680 0±0.040 1
core15k	**0.014 7±0.005 6**	0.027 6±0.009 6	0.004 1±0.002 3	0.032 0±0.010 3
genbase	**0.952 2±0.026 8**	0.780 4±0.070 5	0.951 5±0.027 1	0.946 2±0.031 9

Table 5 Comparative Results of ML*k*NN Classifier in Terms of *HammingLoss*(↓)

表 5 降维后 ML*k*NN 分类器的 *HammingLoss* 值比较(↓)

Datasets	MLFSIE	PCA	LP-CHI	ALL
enron	**0.051 5±0.002 0**	0.058 2±0.002 1	0.057 1±0.001 6	0.052 3±0.002 2
medical	**0.013 0±0.001 7**	0.019 1±0.001 6	0.013 2±0.001 6	0.015 1±0.001 8
core15k	**0.009 4±0.000 1**	0.009 4±0.000 1	0.009 4±0.000 1	0.009 4±0.000 1
genbase	**0.004 3±0.002 5**	0.018 6±0.006 3	0.004 4±0.002 5	0.004 8±0.003 0

1) LP 分类器上的性能评测结果及分析

MLFSIE 在 LP 分类器上的性能评测结果如表 2 和表 3 所示.

从表 2 和表 3 的结果可以看出,MLFSIE 在 4 个数据集上的降维性能均优于 PCA,LP-CHI 和 ALL. 由于 PCA 在降维过程中忽略了特征与标签

之间的关系,因此降维后 LP 分类器的性能在 4 个数据集都出现了下降,而 LP-CHI 在降维后在大多数情况下能够提高 LP 分类器的性能.

使用显著水平为 0.05 的双尾配对 t 检验进行分析可知,在 *MicroFMeasure* 值上,MLFSIE 在数据集 enron 和 medical 上的性能显著优于 PCA,在数据集 enron 上的性能显著优于 LP-CHI,在数据集 medical 上的性能显著优于 ALL. 在 *HammingLoss* 值上,MLFSIE 在数据集 enron,medical 和 core15k 上的性能显著优于 PCA,在数据集 medical 上的性能显著优于 ALL.

2) MLkNN 分类器上的性能评测结果及分析

MLFSIE 在 MLkNN 分类器上的性能评测结果如表 4 和表 5 所示.

在表 4 的结果中,MLFSIE 在 3 个数据集上的降维性能优于 PCA,LP-CHI 和 ALL,而在数据集 core15k 的降维性能低于 PCA 和 ALL. 在表 5 的结果中,MLFSIE 在 4 个数据集上的降维性能均优于 PCA,LP-CHI 和 ALL. PCA 由于没有考虑标签的信息,在 4 个数据集上使得 MLkNN 分类器的性能出现了下降,而 LP-CHI 在大多数情况下能够提升 MLkNN 分类器的性能.

使用显著水平为 0.05 的双尾配对 t 检验进行分析可知,在 *MicroFMeasure* 值上,MLFSIE 在数据集 enron,medical 和 genbase 上的性能显著优于 PCA,而在数据集 core15k 上的性能则出现了显著下降;在数据集 enron 和 core15k 上的性能显著优于 LP-CHI;在数据集 medical 上的性能显著优于 ALL,而在数据集 core15k 上的性能则出现了显著下降. 在 *HammingLoss* 值上,MLFSIE 在数据集 enron,medical 和 genbase 上的性能显著优于 PCA,在数据集 enron 上的性能显著优于 LP-CHI,在数据集 enron 和 medical 上显著优于 ALL.

5 总结与展望

本文研究多标签问题中的特征选择问题,针对离散型特征变量,以特征与标签集合的信息增益作为衡量特征与标签集合之间相关性的准则,提出了一种基于信息熵的多标签特征选择算法 MLFSIE. 首先给出了特征与标签集合的信息增益的定义,接着给出了一种信息增益的阈值选择方法,根据设定的阈值删除不相关的特征,最后使用实验验证了本文提出的特征选择算法的有效性.

由于本文提出的多标签特征选择算法 MLFSIE 假设特征之间相互独立,忽略了特征之间的关系对标签的影响,因此 MLFSIE 不能处理特征之间相关性比较大的情况. 本文的下一步工作将研究特征之间的关系对标签的影响,进一步提高特征选择的效果,提升分类器的分类性能. 此外,本文提出的多标签特征选择算法 MLFSIE 的处理对象是离散型特征变量,对于连续性特征变量在使用该算法之前需要对特征变量进行离散化处理.

参 考 文 献

[1] Li Yufeng, Huang Shengjun, Zhou Zhihua. Regularized semi-supervised multi-label learning [J]. Journal of Computer Research and Development, 2012, 49(6): 1272-1278 (in Chinese)
(李宇峰, 黄圣君, 周志华. 一种基于正则化的半监督多标记学习方法 [J]. 计算机研究与发展, 2012, 49(6): 1272-1278)

[2] Tsoumakas G, Katakis I, Vlahavas I. Data Mining and Knowledge Discovery Handbook [M]. Berlin: Springer, 2010: 667-685

[3] Zheng Wei, Wang Chaokun, Liu Zhang, et al. A multi-label classification algorithm based on random walk model [J]. Chinese Journal of Computers, 2010, 33(8): 1418-1426 (in Chinese)
(郑伟, 王朝坤, 刘璋, 等. 一种基于随机游走模型的多标签分类算法 [J]. 计算机学报, 2010, 33(8): 1418-1426)

[4] Kong Xiangnan, Li Ming, Jiang Yuan, et al. A transductive multi-label classification method for weak labeling [J]. Journal of Computer Research and Development, 2010, 47(8): 1392-1399 (in Chinese)
(孔祥南, 黎铭, 姜远, 等. 一种针对弱标记的直推式多标记分类方法 [J]. 计算机研究与发展, 2010, 47(8): 1392-1399)

[5] Zhang Y, Zhou Z H. Multi-label dimensionality reduction via dependence maximization [C] // Proc of the 23rd AAAI Conf on Artificial Intelligence and the 20th Innovative Applications of Artificial Intelligence Conference. Menlo Park: American Association for Artificial Intelligence, 2008: 1503-1505

[6] Li G Z, You M, Ge L, et al. Feature selection for semi-supervised multi-label learning with application to gene function analysis [C] // Proc of the 2010 ACM Int Conf on Bioinformatics and Computational Biology. New York: Association for Computing Machinery, 2010: 354-357

[7] You M Y, Liu J M, 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

[8] Shao H, Li G, Liu G, et al. Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine [J]. Science China Information Sciences, 2012, 54(1): 1-13

[9] Lee J, Lim H, Kim D W. Approximating mutual information for multi-label feature selection [J]. Electronics Letters, 2012, 48(15): 929-930

[10] Zhang M L, Pena J M, Robles V. Feature selection for multi-label naive Bayes classification [J]. Information Sciences, 2009, 179(19): 3218-3229

[11] Park C H, Lee M. On applying linear discriminant analysis for multi-labeled problems [J]. Pattern Recognition Letters, 2008, 29(7): 878-887

[12] Yu K, Yu S, Tresp V. Multi-label informed latent semantic indexing [C] // Proc of the 28th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2005: 258-265

[13] Ji S, Ye J. Linear dimensionality reduction for multi-label classification [C] // Proc of the 21st Int Joint Conf on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2009: 1077-1082

[14] Quinlan J R. C4.5: Programs for Machine Learning [M]. San Mateo: Morgan Kaufmann Publishers Inc, 1993

[15] Xu Yan, Li Jintao, Wang Bin, et al. A category resolve power-based feature selection method [J]. Journal of Software, 2008, 19(1): 82-89 (in Chinese)
(徐燕, 李锦涛, 王斌, 等. 基于区分类别能力的高性能特征选择方法 [J]. 软件学报, 2008, 19(1): 82-89)

[16] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution [C] // Proc of the 12th Int Conf on Machine Learning. Menlo Park: American Association for Artificial Intelligence, 2003: 856-863

[17] Tsoumakas G, Spyromitros-xioufis E, Vilcek J, et al. MULAN: A Java library for multi-label learning [J]. Journal of Machine Learning Research, 2011, 12: 2411-2414

[18] Bouckaert R R, Frank E, Hall M A, et al. WEKA - experiences with a java open-source project [J]. Journal of Machine Learning Research, 2010, 11: 2533-2541

[19] Wang J. Geometric Structure of High-Dimensional Data and Dimensionality Reduction [M]. Beijing: Higher Education Prass and Springer, 2011: 95-114

[20] Tsoumakas K T G, Kalliris G, Vlahavas I. Multi-label classification of music into emotions [C] // Proc of the 9th Int Conf of Music Information Retrieval. Philadelphia: Drexel University, 2008: 325-330

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

Zhang Zhenhai, born in 1984. PhD candidate. His research interests include wireless sensor networks and data processing(htty326@163.com).

Li Shining, born in 1967. Received PhD degree in computer science in 2005. Professor and PhD supervisor of Northwestern Polytechnical University. Senior member of China Computer Federation. His research interests include wireless sensor networks and mobile computing(lishining@nwpu.edu.cn).

Li Zhigang, born in 1975. Received PhD degree in computer science in 2006. Associate professor and master supervisor of Northwestern Polytechnical University. His research interests include wireless sensor networks and mobile computing(lizhigang@nwpu.edu.cn).

Chen Hao, born in 1988, Master. His research interests include wireless sensor networks(sniy@foxmail.com).