计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
49(2): 372-382, 2012

基于互信息的无监督特征选择

徐峻岭[1,2] 周毓明[2,3] 陈 林[2,3] 徐宝文[2,3]
[1](东南大学计算机科学与工程学院 南京 210096)
[2](计算机软件新技术国家重点实验室(南京大学) 南京 210093)
[3](南京大学计算机科学与技术系 南京 210093)
(junlingxu@gmail.com)

An Unsupervised Feature Selection Approach Based on Mutual Information

Xu Junling[1,2], Zhou Yuming[2,3], Chen Lin[2,3], and Xu Baowen[2,3]
[1](*School of Computer Science and Engineering, Southeast University, Nanjing* 210096)
[2](*State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing* 210093)
[3](*Department of Computer Science and Technology, Nanjing University, Nanjing* 210093)

Abstract In data analysis, feature selection can be used to reduce the redundancy of features, improve the comprehensibility of models, and identify the hidden structures in high-dimensional data. In this paper, we propose a novel unsupervised feature selection approach based on mutual information called UFS-MI. In UFS-MI, we use a feature selection criterion, UmRMR, to evaluate the importance of each feature, which takes into account both relevance and redundancy. The relevance and redundancy respectively use mutual information to measure the dependence of features on the latent class and the dependence between features. In the new algorithm, features are selected or ranked in a stepwise way, one at a time, by estimating the capability of each specified candidate feature to decrease the uncertainty of other features (i.e. the capability of retaining the information contained in other features). The effectiveness of UFS-MI is confirmed by the theoretical proof which shows it can select features highly correlated with the latent class. An empirical comparison between UFS-MI and several traditional feature selection methods are also conducted on some popular data sets and the results show that UFS-MI can attain better or comparable performance and it is applicable to both numerical and non-numerical features.

Key words feature selection; unsupervised feature selection; mutual information; minimum redundancy and maximum relevance; unsupervised minimum redundancy and maximum relevance

摘 要 在数据分析中,特征选择可以用来降低特征的冗余,提高分析结果的可理解性和发现高维数据中隐藏的结构.提出了一种基于互信息的无监督的特征选择方法(UFS-MI),在UFS-MI中,使用了一种综合考虑了相关度和冗余度的特征选择标准UmRMR(无监督最小冗余最大相关)来评价特征的重要性.相关度和冗余度分别使用互信息来度量特征与潜在类别变量之间的依赖和特征与特征之间的依赖.UFS-MI同时适用于数值型和非数值型特征.在理论上证明了UFS-MI的有效性,实验结果也表明UFS-MI可以达到与传统的特征选择方法相当甚至更好的性能.

关键词 特征选择;无监督特征选择;互信息;最小冗余-最大相关;无监督最小冗余-最大相关
中图法分类号 TP391

收稿日期:2010-07-07;**修回日期**:2011-06-24
基金项目:国家自然科学基金项目(61073029,90818027,60633010);国家"八六三"高技术研究发展计划基金项目(2009AA01Z147);国家"九七三"重点基础研究发展计划基金项目(2009CB320703)
通信作者:徐宝文(bwxu@nju.edu.cn)

在数据分析中,人们常常需要处理具有很多特征且包含大量实例的数据集.在这类数据集中,有些特征是冗余的甚至是不相关的.冗余特征的存在会降低学习算法的效率,而不相关特征(噪音特征)的存在会有损学习算法的性能.因此,在分析处理这类数据集时,我们有必要对数据进行预处理以去除冗余特征和噪音,即需要进行特征选择.特征选择能给学习算法带来很多好处,例如可降低其计算代价、可使其生成更易理解的结果和更紧凑、泛化能力更强的模型.

人们已经对特征选择方法进行了许多研究.依据数据是否具有类别信息,特征选择可分为有监督和无监督两类.当数据的类别信息已知时,我们可以使用有监督的特征选择方法.然而,当数据的类别信息不可知时,我们只能使用无监督特征选择方法.依据选择过程是否依赖于最终使用被选特征子集的学习算法,特征选择可分为过滤型方法和包装型方法[1].过滤型方法独立于具体的学习算法,而包装型方法将学习算法作为其在特征选择过程中的评价标准.Dash 和 Liu 将特征选择分为 4 个步骤[2],其中的两个主要步骤是搜索和评价,在特征子集空间中的不同搜索方式和对特征子集的不同评价标准产生了各种各样的特征选择方法.最常见的搜索策略就是顺序搜索(前向或后向),还有一些启发式方法,如浮点搜索、柱状搜索、双向搜索和基因搜索等[2].评价标准可以划分为五大类:距离度量[3-4]、依赖性度量[5-8]、信息(不确定性)度量[9-11]、一致性度量[12-13]和分类精确度度量[14].对于有监督特征选择,通常评价函数是用来度量某个特征或特征子集区分不同类别的能力;然而对于无监督特征选择,由于缺乏类别信息,评价函数的定义更加具有挑战性.

无监督特征选择的目的是,在没有类别信息的情况下选择出能够刻画原始特征主要特性的特征.在实际应用中,经常要为没有类别信息的数据设计分类器.如在软件工程中,为一个新项目中的程序模块设计缺陷预测器[15].在这种情况下,我们就需要用无监督特征选择方法来选择与潜在类别变量高度相关的特征.近年来,研究者们在无监督特征选择方面做了很多工作[6-7,10,16-20],其中,Mitra 等人使用最大信息压缩指标来度量特征之间的相似性以检测冗余特征[6].Wei 和 Billings 提出了一种最大化特征全局依赖的前向正交搜索算法,假定特征之间存在线性依赖并以此来检测重要特征[7].Dash 等人提出了一种基于信息熵的标准来度量每个特征的聚类趋势,他们观察到是否具有聚类趋势的数据拥有非常迥异的点对点距离柱状图[10].然而,这些方法都是针对特征全为数值型特征的数据而设计,因而无法处理带有非数值型特征的数据.Dy 和 Brodley 提出了一种对于聚类的特征选择度量,它使用规范聚类可分离度(对于 k-均值聚类算法)和规范似然度(对于期望最大化算法)来评价聚类的质量[16],并以此来对不同的特征子集进行评价.Law 等人将特征选择与期望最大化算法集成在一起,在聚类的过程中同时进行特征选择[17].类似地,刘涛等人也提出了一种通过在不同 K-均值聚类结果上使用有监督特征选择的方法[18].Li 等人采用散度可分离性标准对数据进行局部特征选择,即为每个群集选择代表性特征[19].Modha 和 Spangler 提出基于 k-均值聚类生成的类内和类间散度矩阵为不同的特征子集赋予不同的权重[20].然而,上述方法都是包装型特征选择方法,其在特征搜索和选择的过程中都会加入学习算法的偏置,而且时间复杂度较高.

本文提出了一种基于互信息的无监督特征选择方法 UFS-MI,其搜索方式为前向搜索,评价标准 UmRMR 是综合考虑了特征的相关度和冗余度的信息度量.与先前方法不同的是,UFS-MI 是一种可以同时处理数值型和非数值型数据的过滤型特征选择方法.特别地,UFS-MI 使用互信息作为度量标准,其目的是用来度量两个特征之间的一般依赖.虽然互信息已经广泛应用于有监督特征选择方法[9,11],但据我们所知,其还未曾在无监督特征选择方法中被用作为一种度量标准.在没有给定输入参数的情况下(UFS-MI 唯一可选参数为期望特征个数),UFS-MI 属于无监督过滤型特征排序方法.在进行特征选择时,UFS-MI 首先计算出每个特征的相关度,然后使用前向顺序搜索对特征进行重要性评价,最后输出一个有序特征序列.

1 基于互信息的特征选择

本节首先简明介绍使用互信息度量特征重要性的动机,然后提出一个综合考虑了“相关度”和“冗余度”两个概念的特征评价标准及特征选择算法,最后讨论互信息估计的实现问题.

1.1 动 机

一个学习算法可以看作是一个通过“消费”输入向量中包含的信息来减少初始不确定性的系统[9].输入向量是由用于表示实例的特征所组成,因此学

习算法所“消费”的信息即是特征所包含的信息. 特征选择的目标可认为是从原始特征子集中选取包含所有特征蕴含的全部或绝大部分信息的特征子集. 由于被丢弃的特征几乎是无信息量的,因此学习算法的性能将会很少降低,甚至由于去除掉带有干扰信息的特征而导致算法性能提高. 假设数据集 D 由 n 个特征 $(f_1, f_2, \cdots, f_n)$ 所表示的 N 个实例组成, $P(f_t)$ 是特征 f_t 为不同可能值 f_t 的概率. 特征 f_t 取值的初始不确定性可以由如下信息熵度量:

$$H(f_t) = -\sum_{f_t} P(f_t)\log P(f_t). \tag{1}$$

在已知另一个特征 $f_{t'}$ 的取值之后, f_t 取值的不确定性可以由条件熵来度量:

$$H(f_t \mid f_{t'}) = -\sum_{f_{t'}} P(f_{t'}) \sum_{f_t} P(f_t \mid f_{t'})\log P(f_t \mid f_{t'}). \tag{2}$$

如果特征为连续性变量,则将求和替换为求积分,相应的概率替换为概率密度. 在通常情况下,条件熵小于或等于初始信息熵(当两个变量相互独立时取值相等). 在此基础上,两个特征 f_t 与 $f_{t'}$ 之间的互信息可以定义为

$$I(f_t; f_{t'}) = H(f_t) - H(f_t \mid f_{t'}) = I(f_{t'}; f_t). \tag{3}$$

显然,互信息可以看作是已知特征 $f_{t'}$ 的信息后对于特征 f_t 的不确定性的减少量,即二者共同含有的信息量. 从信息论的角度来看,特征选择的目标就是寻找一个包含原始特征集合带有的绝大部分或者全部信息的特征子集 $S_d = \{g_1, g_2, \cdots, g_d\}$ (该特征子集的存在可以最大程度地降低其他未选特征的不确定性),其中, $g_m = f_{i_m}$, $i_m \in \{1, 2, \cdots, n\}$, $m = 1, 2, \cdots, d (d < n)$. 被选特征子集 S_d 的性能可以通过检测使用 S_d 表征数据集时学习算法的性能来进行评价.

1.2 特征排序和选择

基于以上分析,我们的直觉是特征选择算法应该选择那些与其他特征具有最大互信息的特征,因为它们可以最大程度地降低其他特征的不确定性,也就是说其他特征由于这些特征的存在将会提供微不足道的信息. 本文的特征选择过程从一个空集合 S 开始,并且采用步进的方式,每次选择一个特征. 在第 1 步,令

$$score(f_i) = \frac{1}{n}\sum_{t=1}^{n} I(f_i; f_t), \tag{4}$$

$$\ell_1 = \arg\max_{1 \leqslant i \leqslant n} \{score(f_i)\}. \tag{5}$$

第 1 个重要特征可以选择 $g_1 = f_{\ell_1}$,因为其可以最大程度地降低特征集合中其他特征的不确定性. 也就是说,在只选择一个特征的情况下, $g_1 = f_{\ell_1}$ 对系统提供的信息是最多的.

假设 U 为当前未被选择特征集合, S_{m-1} 为已被选的 $m-1$ 个特征组成的集合,那么第 m 个特征 g_m 怎么选择呢?我们采取的策略类似于有监督特征选择方法中著名的“最小冗余-最大相关”标准(minimum redundancy and maximum relevance, mRMR)[11],在 mRMR 中,第 m 个特征选择的依据是:

$$\ell_m = \arg\max_{1 \leqslant i \leqslant n} \left\{ I(f_i; c) - \frac{1}{m-1}\sum_{f_t \in S_{m-1}} I(f_i; f_t) \mid f_i \in U \right\}. \tag{6}$$

式(6)中的第 1 项是“最大相关”条件,它的存在使得 mRMR 倾向于选择和目标类别变量 c 具有最大依赖的变量. 由于仅根据“最大相关”条件选择的特征可能存在冗余(特征之间依赖度非常大),因而加入“最小冗余”条件以便选择互斥的特征. 因为当两个特征高度依赖时,去除其中的一个对系统的类区分能力的影响可能不大.

然而,在无监督特征选择中,类别信息是未知的. 在我们提出的基于互信息的无监督特征选择方法中,第 m 个特征的选择采取以下形式:相对于 U 中的其他特征, g_m 应该与整个特征集合最大程度的“相关”,同时它应该与 S_{m-1} 中的已选特征最小程度的“冗余”. 为此,我们首先定义特征的“相关度”.

定义 1. 相关度. 一个特征 f_i 的相关度就是其与整个特征集合的平均互信息:

$$Rel(f_i) = \frac{1}{n}\sum_{t=1}^{n} I(f_i; f_t) = \frac{1}{n}\left(H(f_i) + \sum_{1 \leqslant t \leqslant n, t \neq i} I(f_i; f_t)\right). \tag{7}$$

在特征的相关度定义中, $H(f_i)$ 表示特征 f_i 所包含的信息量. $H(f_i)$ 越大,表明特征 f_i 能够给学习算法提供越多的信息. $\sum_{1 \leqslant t \leqslant n, t \neq i} I(f_i; f_t)$ 表示有了特征 f_i 的知识后其他特征包含的信息量的减少量(即为特征 f_i 和其他特征共同含有的信息量),其值越大表示其他特征能够提供给学习算法的“新”的信息越少. 如果选择具有最大 $Rel(f_i)$ 值的特征,那么数据就可以最小程度地丢失信息.

在定义特征的“冗余度”之前,我们首先假设特征 g_t 的相关度与其信息熵 $H(g_t)$ 的值成比例(即对

于某个特征 g_t 其单位信息量提供的相关度 $Rel(g_t)/H(g_t)$ 是一个常数 c_{g_t},虽然每个特征可能具有不同的常数值). 对于 U 中的某个候选特征 f_i,S_{m-1} 中的任一特征 g_t 相对于 f_i 的条件信息量为 $H(g_t|f_i)$. 显然,如果选择 f_i 加入 S_{m-1},则特征 g_t 独家提供的信息量由于 f_i 的加入而变小,因此其相关度应该也变小. 根据上面的假设,下面给出"条件相关度"的定义.

定义 2. 条件相关度. 一个特征 g_t 对特征 f_i 的条件相关度可以定义为

$$Rel(g_t \mid f_i) = \frac{H(g_t \mid f_i)}{H(g_t)} Rel(g_t). \quad (8)$$

显然,条件相关度小于等于相关度(当这两个特征独立时取相等),二者之间的差别即可定义为冗余.

定义 3. 冗余度. 一个特征 f_i 对特征 g_t 的冗余度可以定义为

$$Red(f_i;g_t) = Rel(g_t) - Rel(g_t \mid f_i). \quad (9)$$

在选择第 m 个重要特征时,综合考虑候选特征的相关度以及其对于已选特征的冗余度,得到我们的"无监督最小冗余-最大相关"特征重要性评价标准(UmRMR):

$$UmRMR(f_i) = Rel(f_i) - \max_{g_t \in S_{m-1}} \{Red(f_i;g_t)\}$$

或

$$UmRMR(f_i) = Rel(f_i) - \frac{1}{m-1} \sum_{g_t \in S_{m-1}} Red(f_i;g_t).$$

设

$$\ell_m = \arg\max_{1 \leqslant i \leqslant n} \{UmRMR(f_i) \mid f_i \in U\}, \quad (10)$$

则第 m 个特征可以选择为 $g_m = f_{\ell_m}$,因为这个特征最大程度地降低了其他特征的不确定性,同时它只带来了很少的冗余信息. 在选择后续特征时,我们采用类似的方式逐个进行选择.

特征选择算法的详细信息如算法 1 所示. 如果在步骤 4 没有给定期望特征个数,那么这个数字就默认为 $n-1$. 注意,此时算法输出的不是一个特征子集,而是一个有序特征序列. 在实际应用时,为了给特定的学习任务提供一个良好的特征子集,可以使用以该学习算法作为评价函数的包装型方法来进一步选择特征. 由于被选出的特征继承了该学习算法的偏置,因此有助于提高该学习算法的性能. 与传统意义上包装型方法相比,此处特征选择方法的一个重要特点在于特征已经按照重要性进行了排序,因此在特征子集空间中搜索的时间复杂度是线性的,而不是 $O(2^n)$.

算法 1. 基于互信息的无监督特征选择方法.

① Initialization:
$S=\varnothing$;
/* S is the set of selected features */
$U=\{f_1,f_2,\cdots,f_n\}$.
/* U is the set of unselected features */

② Pre-computation:
For $\forall f_i, f_t \in U$ compute $H(f_i)$, $I(f_i;f_t)$, $Rel(f_i)$.

③ First feature selection:
Find feature f_{ℓ_1} according to formula (5);
$U=U\backslash f_{\ell_1}$, $S=\{f_{\ell_1}\}$.

④ Features selection step:
Repeat until desired num of features are selected
Find feature f_{ℓ_m} according to formula (10);
$U=U\backslash f_{\ell_m}$, $S=S\cup\{f_{\ell_m}\}$.
End Repeat

对 UFS-MI 实现的时间复杂度主要取决于两部分:步骤 2 中的任意两个特征间的互信息的计算和步骤 4 中的特征选择过程. 特征选择过程的最坏时间复杂度为 $O(n^2)$,而互信息计算的时间复杂度为 $O(n^2N)$,因此,UFS-MI 的时间复杂度为 $O(n^2N)$.

1.3 互信息估计

UFS-MI 既可以处理非数值型数据,也可以处理数值型(离散型或连续型)数据. 对于非数值型和离散型特征变量,互信息的计算很直接,因为联合概率分布表和边际概率分布表可以通过清点数据中变量的样本值来计算得到. 然而,当两个特征变量 f_i 和 f_j 中有一个是连续型变量时,它们的互信息 $I(f_i,f_j)$ 很难计算,原因在于从有限数目的样本中估计一个连续型变量的概率分布密度是很困难的. 在这种情况下,我们可以使用密度估计方法(如 Parzen 窗)来近似估计 $I(f_i,f_j)$,也可以采用数据离散化作为预处理步骤. 当采用数据离散化作为预处理步骤时,通常有两种数据离散化方法可供选择:无监督离散化和有监督离散化. 无监督离散化不考虑训练集中实例所带有的类别信息来对每个特征进行离散化,有监督离散化将类别信息考虑进去来对特征进行离散化.

2 与有监督特征选择标准 mRMR 的联系

在有监督特征选择中,Peng 等人[11]用实验证明了由"最大相关"和"最小冗余"标准综合而成的

mRMR 递增选择模式为最大化被选特征与类别变量的依赖提供了一种很好的方式. 然而,对于无监督特征选择,隐藏在数据中的潜在类别信息是未知的,我们能够在没有类别信息的情况下最大化被选特征与潜在类别特征之间的依赖吗?通过研究 UmRMR 和 mRMR 二者之间的联系,我们发现 UmRMR 可以在某种程度上解决这个问题. 下面分别讨论两种标准中相关度和冗余度的联系.

命题 1. 在简单贝叶斯假设的前提下,UmRMR 中特征 f_i 的相关度 $Rel(f_i)$ 是 mRMR 中特征相关度的一个下界,即 $Rel(f_i)\leqslant I(f_i;c)$.

证明. 设 c 是数据集中的潜在类别变量. 我们将 c 看成是特征 $f_1,f_2,\cdots,f_n$ 的一个函数,其中 $f_1,f_2,\cdots,f_n$ 可看作是随机变量,因此,c 也是一个随机变量,于是:

$$H(f_i\mid c)=\sum_c\sum_{f_i}P(f_i,c)\log\frac{1}{P(f_i\mid c)}=$$
$$\sum_c\sum_{f_i}\sum_{f_t}P(f_i,f_t)\frac{P(f_i,f_t,c)}{P(f_i,f_t)}\log\frac{1}{P(f_i\mid c)}=$$
$$\sum_{f_i}\sum_{f_t}P(f_i,f_t)\sum_c\frac{P(f_i,f_t,c)}{P(f_i,f_t)}\log\frac{1}{P(f_i\mid c)}\leqslant$$
$$\sum_{f_i}\sum_{f_t}P(f_i,f_t)\log[\sum_c\frac{P(f_i,f_t,c)}{P(f_i,f_t)}\frac{1}{P(f_i\mid c)}]=\tag{11}$$
$$\sum_{f_i}\sum_{f_t}P(f_i,f_t)\log[\frac{1}{P(f_i,f_t)}\sum_c\frac{P(f_i,f_t,c)}{P(f_i\mid c)}]=$$
$$\sum_{f_i}\sum_{f_t}P(f_i,f_t)\log[\frac{1}{P(f_i\mid f_t)P(f_t)}\sum_c\frac{P(f_i,f_t,c)}{P(f_i\mid c)}]=$$
$$\sum_{f_i}\sum_{f_t}P(f_i,f_t)[\log\frac{1}{P(f_i\mid f_t)}+\log\sum_c\frac{P(f_i,f_t,c)}{P(f_i\mid c)P(f_t)}]\leqslant$$
$$H(f_i\mid f_t)+\log\sum_{f_i}\sum_{f_t}\sum_{f_c}\frac{P(f_i,f_t)P(f_i,f_t,c)}{P(f_t)P(f_i\mid c)}.\tag{12}$$

注意,我们在式(11)和式(12)处应用了詹森不等式[21]. 在简单贝叶斯假设的前提下,$P(f_i,f_t,c)=P(c)P(f_i|c)P(f_t|c)$. 将它带入式(12),可得:

$$H(f_i\mid c)\leqslant H(f_i\mid f_t)+$$
$$\log\sum_{f_i}\sum_{f_t}\sum_c\frac{P(f_i,f_t)P(c)P(f_i\mid c)P(f_t\mid c)}{P(f_t)P(f_i\mid c)}\leqslant$$
$$H(f_i\mid f_t)+\log[\sum_{f_i}\sum_{f_t}P(f_i,f_t)\sum_c\frac{P(f_t,c)}{P(f_t)}]=$$
$$H(f_i\mid f_t)+\log\sum_{f_i}\sum_{f_t}P(f_i,f_t)=$$
$$H(f_i\mid f_t).\tag{13}$$

根据式(13)中得到的结论和式(7)中的相关度定义,我们得到:

$$Rel(f_i)=\frac{1}{n}\sum_{t=1}^{n}I(f_i;f_t)=$$
$$\frac{1}{n}\sum_{t=1}^{n}(H(f_i)-H(f_i\mid f_t))\leqslant$$
$$\frac{1}{n}\sum_{t=1}^{n}(H(f_i)-H(f_i\mid c))=I(f_i;c).$$

证毕.

在数据类别信息未知或丢失的情况下,$I(f_i;c)$ 的值无法计算. 命题 1 表明,最大化 $Rel(f_i)$ 的值可以在某种程度上迫使 $I(f_i;c)$ 取得较大的值. 因此根据 $Rel(f_i)$ 标准可以选择出与潜在类别变量依赖度较大的特征.

在讨论两种标准中冗余度之间的关系前,我们先引入 Wang 等人提出的特征 f_i 对特征 g_t 的熵相关度的定义[22]:$r(f_i;g_t)=I(f_i;g_t)/H(g_t)$. 可以看出,熵相关度表示的是 f_i 和 g_t 所共有的信息占单个特征 g_t 所含信息的比重,即二者信息的冗余率.

命题 2. 特征 f_i 相对于特征 g_t 冗余度等于特征 f_i 对特征 g_t 的熵相关度乘以特征 g_t 的相关度.

证明.

$$Red(f_i;g_t)=Rel(g_t)-Rel(g_t\mid f_i)=$$
$$Rel(g_t)-\frac{H(g_t\mid f_i)}{H(g_t)}Rel(g_t)=$$
$$\frac{I(f_i;g_t)}{H(g_t)}Rel(g_t)=c_{g_t}I(f_i;g_t)=$$
$$r(f_i;g_t)Rel(g_t).$$

证毕.

命题 2 表明,UmRMR 评定两个特征的冗余度时不是简单地考虑二者共同含有的信息(冗余信息),而是考虑二者冗余信息部分所能提供的相关度值,即冗余率乘以相关度值,因为候选特征的加入对

已选特征所产生的冗余就是已选特征相关度的减少值.同时,在命题 2 的证明过程中可以看出,UmRMR 中特征之间的冗余度等于 mRMR 中特征之间的冗余度乘以一个常数.

3 实验结果

我们采用 UCI 机器学习知识库[23]中的数据集来评价 UFS-MI 的性能.在实验中,对每个数据集,我们首先使用 UFS-MI 选择一个特征子集,然后用被选中的特征子集来表征原始数据集以设计模式分类器.虽然所用数据集中类别信息已知,但是在评价特征的重要性时此信息被忽略(我们的算法应用于处理类别信息不可获得或者丢失的情况).UFS-MI 的一个重要特点在于不仅能处理数值型数据,而且能处理非数值型数据.在实验中,我们首先评价 UFS-MI 在非数值型数据上的性能,然后在数值型数据集上比较它与其他特征选择方法的性能.

3.1 非数值型数据集上的实验

在这个实验中,我们试图回答 3 个问题:1) UFS-MI 是否能够识别重要的特征;2) UFS-MI 所选择出的特征是否能有效提高学习算法的性能;3) UFS-MI 生成的有序特征序列是否能够反映特征与潜在类别变量的相关性或者依赖性.

1) UFS-MI 是否能识别重要特征

本实验采用 UCI 中特征类型全为非数值型的 4 个数据集,数据集的部分信息如表 1 所示.其中的 $featNum$,$instNum$ 和 $classNum$ 分别表示数据集中特征的个数、实例的个数和类别的个数:

Table 1 Data Sets with Non-numerical Features

表 1 非数值型特征的数据集

Data Set	*featNum*	*instNum*	*classNum*
audiology	69	226	24
splice	61	3 190	3
lung-cancer	56	32	2
vote	16	435	2

由于 UFS-MI 的输出是一个有序的特征列表(按照重要性降序排列),我们采用如下方式评价其性能.首先,我们根据特征在 UFS-MI 输出列表中的顺序逐个递增地选取以组成一个特征子集,然后用这个特征子集表征原始数据并将其作为简单贝叶斯分类器的输入.在评价分类器的性能时,我们将 10 次 10-fold 交叉验证运行结果的分类准确性的平均值作为最终的分类准确性.

图 1 给出了采用不同个数的特征来表征数据时

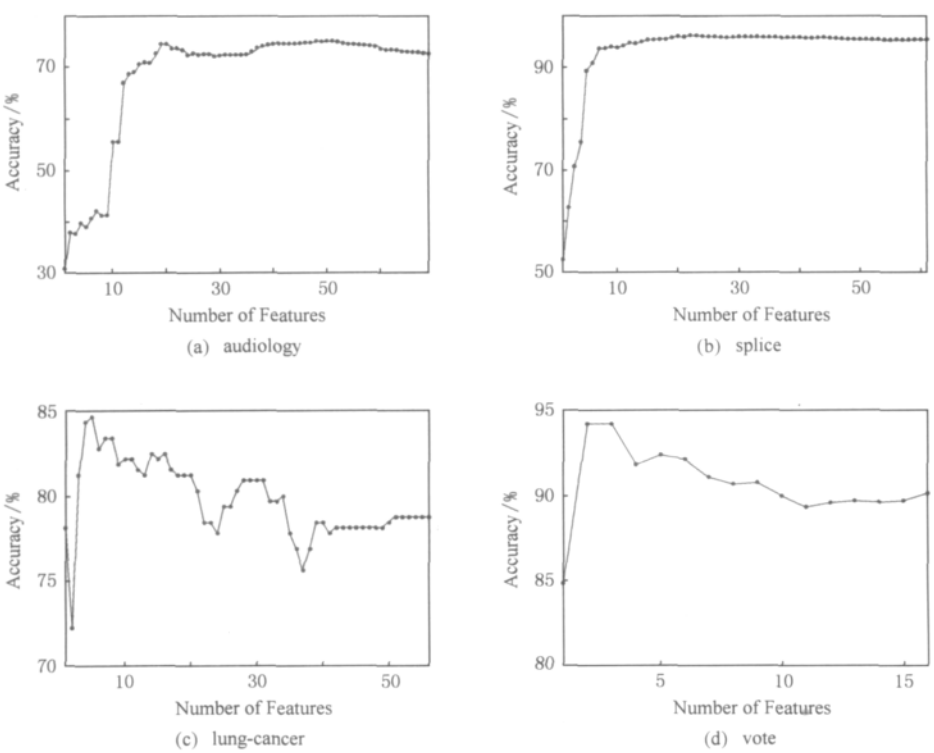

Fig. 1 Performance on data sets with non-numerical features.

图 1 UFS-MI 在非数值型数据集上的性能

分类器的分类准确性变化趋势. 容易看出,对于数据集 audiology 和 splice,分类器可以在拥有少量特征的约简数据上得到与在完全数据上相同的分类准确性. 对于数据集 lung-cancer 和 vote,当使用 UFS-MI 生成的特征序列中的前几个特征表征数据时,分类器的分类准确性高于其在完全数据上的准确性. 这些结果表明,UFS-MI 确实将具有代表性和富含信息的重要特征放在了特征序列的前面. 此外,从图 1 我们还可以看出,随着特征个数的增加,分类准确性起初急速提高,然后保持不变甚至下降. 这是由于越来越多的冗余的或者不相关的特征被包含了进来,它们不能给学习算法提供新的信息甚至会误导学习算法.

2) UFS-MI 的有效性

现在我们在更多的数据集上来检测 UFS-MI 的有效性. 在所使用的数据集中,ionosphere, sonar 和 glass 只含有数值型特征;lung-cancer 和 vote 只含有非数值型特征;zoo, sponge 和 arrhythmia 既含有非数值型特征,也有数值型特征. 对于数值型特征我们采用基于信息熵的使用最小描述长度(MDL)作为停止标准的离散化方法[24]预先对其进行离散化. 所采用的 8 个数据集如表 2 所示.

在表 2 中,lung-cancer 和 vote 数据集前面已被使用,再次用它们目的是为了证明 UFS-MI 对于不同学习算法的有效性. 在这个实验中,我们使用 K-最近邻方法的分类准确性作为特征子集优劣的评价指标. 更具体地说,我们首先根据特征在 UFS-MI 输出列表中的顺序逐个递增地选取以组成一个特征子集,然后用这个特征子集表征的数据作为 K-最近邻方法的输入,最后我们选择出能够达到 K-最近邻方法在完全数据集上所得准确性的最小特征子集和能够达到最高分类准确性的最优特征子集. 分类准确性依然为 10 次 10-fold 交叉验证运行结果的平均值. 至于 K-最近邻方法中 K 值的选择,我们通过用不同的 K 值($1\leqslant K\leqslant\sqrt{N_{tr}}$, N_{tr} 是训练样本集中样本的个数)进行多次实验,最后选定使分类器具有最好性能的 K 值.

Table 2 List of Data Sets with Numerical or Non-numerical Features

表 2 具有数值型和非数值型数据的数据集

Data Set	*featNum*	*instNum*	*classNum*
ionosphere	33	351	2
zoo	17	101	7
sponge	45	76	3
sonar	60	208	2
glass	9	214	7
arrhythmia	279	452	16
lung-cancer	56	32	2
vote	16	435	2

表 3 给出了每个数据集上原始特征集、最小特征子集和最优特征子集中的特征个数(C, M, O)及其对应的分类准确性. 显然,分类器在被选特征表征的数据集上的分类准确性优于其在完全数据集上的结果. 特别地,在数据集 ionosphere 和 arrhythmia 上,UFS-MI 所得结果也明显优于 FOS-MOD[7]所得结果.

Table 3 Classification Accuracy over the Complete Data and Reduced Data

表 3 分类器在完全数据集和约简数据集上的分类准确性

Dataset	No. Features			Accuracy/%		
	C	*M*	*O*	*C*	*M*	*O*
Ionosphere	33	10	10	89.77±0.71{2}	90.57±1.14{2}	90.57±1.14{2}
Zoo	17	12	12	96.14±0.50{1}	98.02±0.00{1}	98.02±0.00{1}
Sponge	45	17	32	92.50±0.66{3}	92.63±0.66{2}	93.68±1.32{2}
Sonar	60	22	54	86.44±1.44{1}	86.88±1.92{1}	88.08±2.64{2}
Glass	9	5	6	70.00±2.10{3}	75.75±2.34{1}	77.57±0.94{1}
Arrhythmia	279	2	48	59.27±0.55{6}	59.60±1.33{4}	67.61±0.55{3}
lung-cancer	56	3	5	79.69±1.56{3}	85.00±1.56{1}	88.44±1.56{2}
Vote	16	2	7	93.15±0.46{4}	95.17±0.00{1}	95.59±0.92{1}

Notes: $C/M/O$: Complete/Minimal/Optimal Data; {}: the values of k used in k-NN rule

3) UFS-MI 是否能够选出与潜在类别变量高度相关的特征

为了检验算法 UFS-MI 选择的特征与潜在类别变量之间的相关度,我们将 UFS-MI 输出的有序特征序列同一种机器学习领域广泛使用的有监督特征选择方法 IG(信息增益)所得到的有序特征序列进行比较. 实验中使用了 7 个数据集,数据集部分信息如表 4 所示. 对于数据集中的数值型特征我们依然采用文献[24]中的方法先对其进行离散化. 2 种方法分别对每个数据集输出一个有序特征序列,实验结果如表 5 所示. 其中粗体数字表示的是在两个序列中具有相同排序的特征的下标或者是两个序列中前若干个特征中相同的特征的下标. 从表 5 中可以看出 IG 所认定的重要特征通常都会在 UFS-MI 输出的特征列表的前端,虽然 UFS-MI 在选择特征的过程中没有使用类别信息(除了在数据的离散化过程中使用了类别信息).

Table 4 List of Data Sets

表 4 数据集

Data Set	*featNum*	*instNum*	*classNum*
Ecoli	7	336	8
Iris	4	150	2
Lymph	18	148	4
dermatology	34	366	6
breast-w	9	699	2
Spambase	57	4 601	2
Haberman	3	306	2

Table 5 Feature Ranking by IG and UFS-MI

表 5 IG 和 UFS-MI 所生成的特征序列

Data Set	Feature Ranking by IG	Feature Ranking by UFS-MI
ecoli	{**6,7,1,2,5,3,4**}	{**6,7,1,2,5,3,4**}
Iris	{**3,4,1,2**}	{**3,4,1,2**}
lymph	{**13**,18,**15**,**14**,2,**10** …}	{**14**,**13**,**10**,12,**15**,5 …}
dermatology	{**21**,**20**,**22**,33,29,**27** …}	{**20**,**27**,**21**,16,**22**,9 …}
breast-w	{**2**,**3**,**6**,**7**,**5** …}	{**2**,**7**,**3**,**5**,**6** …}
spambase	{**52**,**53**,**56**,7,**21** …}	{57,**56**,**53**,**21**,**52** …}
haberman	{**3**,**2**,**1**}	{**3**,**2**,**1**}

3.2 数值型数据集上的实验

现有的无监督特征选择方法都可以处理数值型数据,我们将 UFS-MI 和两种无监督特征选择方法 ENTROPY[7] 和 FOS-MOD[10] 的性能进行了比较. 之所以没有将 UFS-MI 和 Mitra 等人提出的方法[6]进行比较,是因为他们的方法具有一个指定期望特征个数的输出参数 K,并且算法的输出对于不同的 K 很敏感(对于不同的 K,某个特征所在序列中的位置会发生变化). 下面我们对待比较的两种特征选择方法进行一个简要的介绍.

ENTROPY 是一个基于信息熵的特征排序算法,由 Dash 等人提出[10]. 在该方法中,特征的重要性由将其移除后导致的信息熵的减少量来度量. 信息熵的定义如下:

$$E(f_i)=-\sum_{p=1}^{N}\sum_{q=1}^{N}[(S_{p,q}\times\log S_{p,q})+(1-S_{p,q})\times\log(1-S_{p,q})],$$

其中 $S_{p,q}$ 表示实例 p 和 q 之间的相似度. $S_{p,q}=e^{\alpha\times dist_{p,q}}$,$\alpha=-\ln(0.5)/\overline{dist}$,$dist_{p,q}$ 表示去除特征 f_i 后实例 p 和 q 之间的距离,$\overline{dist}$ 表示去除特征 f_i 后所有实例之间的距离的平均值. 信息熵 $E(f_i)$ 的值越小表示特征 f_i 越重要,算法最后的输出是一个有序特征序列.

FOS-MOD 是一个无监督的前向正交搜索算法. 该方法采用步进的方式,一次选择一个特征,其选择特征依据每个候选特征子集在度量空间中表示整个特征集合的能力. 特征之间的依赖由平方相关系数来度量. 特征 x 和 y 的平方相关系数 $sc(x,y)=(x^{\mathrm{T}}y)/[(x^{\mathrm{T}}x)(y^{\mathrm{T}}y)]$. FOS-MOD 中的相关度的定义和式(7)中的定义类似,只不过互信息被换成了平方相关系数. FOS-MOD 中没有显示的冗余度的定义,在正交化的过程中已经隐含了去除冗余的操作(在计算第 m 个重要特征的相关度之前将其与之前选择的 $m-1$ 个特征进行施密特正交化). 本实验中所使用的 6 个数据集的所有特征都是连续型变量(数据集的部分信息如表 6 所示). 在使用 UFS-MI 进行特征选择之前,简单的等宽分箱离散化方法被用来对数据集中的连续型特征进行无监督离散化,其中的分箱个数通过最大似然度估计来自动选择. 此处之所以采用无监督离散化方法,是为了在整个特征选择的过程中(包括预处理的离散化过程)不使用数据中的类别信息,也是为了更彻底地检验 UFS-MI 是否真的可以选择出与潜在类别变量高度相关的特征. 实验过程如下:首先,分别使用 3 种特征选择算法对每个数据集处理一遍,算法的输入都不包含期望的特征个数,因此输出结果就是对于每个数据集分别有 3 个有序特征序列. 然后将特征有序地逐个地选定以组成一个特征子集,使用 K-最近

邻方法算法对不同数据集的不同大小的特征子集表征的数据进行分类，K-值的选择过程如 3.1 节相同，选择使得分类器具有最好性能的 K 值. 图 2 给出了 3 种特征选择算法的 K-最近邻方法分类准确性：

Table 6 Data Sets with Numerical Features

表 6 数值型特征的数据集

Data Set	*featNum*	*instNum*	*classNum*
liver-disorders	6	345	2
glass	9	214	7
segment	19	2 310	7
sonar	60	208	2
vehicle	18	846	4
ionosphere	33	351	2

容易看出，UFS-MI 几乎在所有数据集上性能都优于其他两种方法. 特别地，UFS-MI 可以在选择相对很少的特征表征数据的时候就可以获得完全数据集上所得到性能，然而其他两种方法倾向于选择更多的特征. 更加明显的是，分类器在这 3 种特征选择方法处理过的数据上的分类准确性曲线具有不同的趋势. 对于 UFS-MI，随着特征个数的增加分类准确性先是上升，然后趋近于某一个点或者开始下降. 对于 ENTROPY 和 FOS-MOD，分类准确性几乎是随着特征个数的增加持续上升，这表明在这两种方法中，冗余特征甚至是不相关特征都被赋予了很高的重要性.

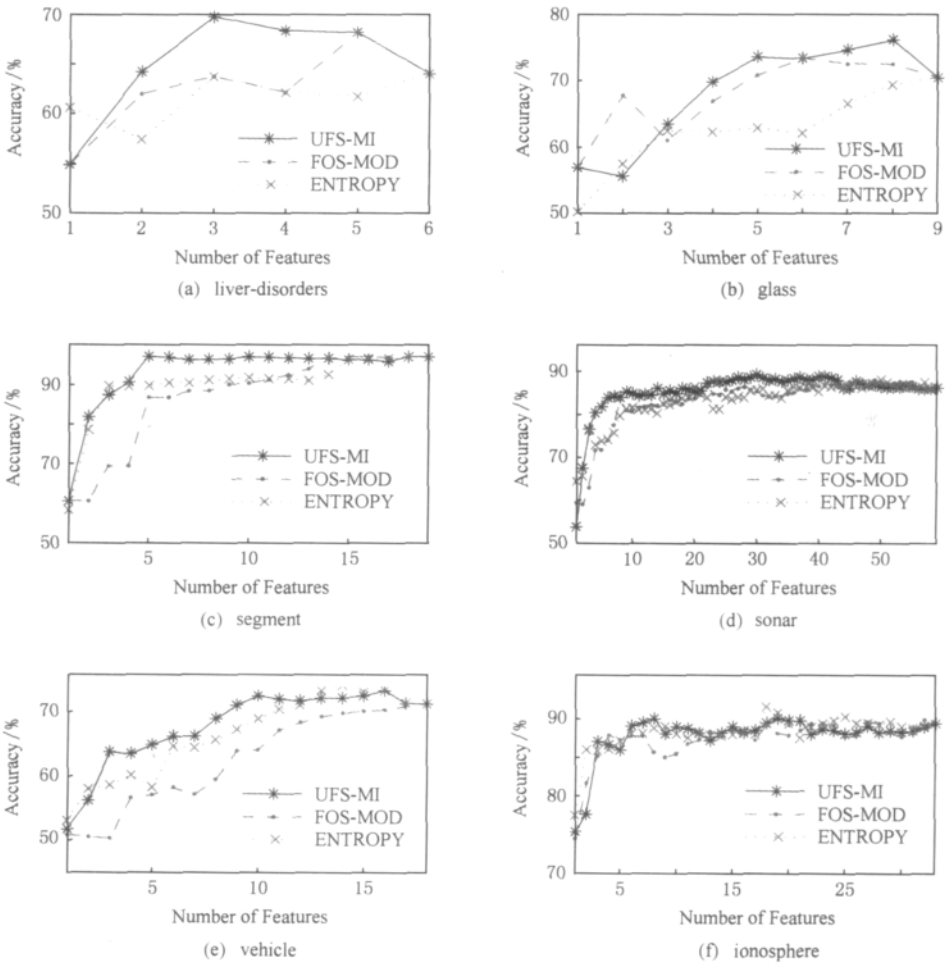

Fig. 2 Performance comparison data sets with numerical features.

图 2 在数值型数据集上各种特征选择方法的性能比较

算法 FOS-MOD 和 ENTROPY 的计算时间复杂度分别为 $O(n^3N)$ 和 $O(n^2N)$. ENTROPY 和 UFS-MI 具有相同量级的时间复杂度，而 FOS-MOD 的时间复杂度较二者高出很多，且 UFS-MI 的性能表现更优. 因此，在实际的特征选择应用中，UFS-MI 相比其他两者更有应用前景.

4 结 论

本文提出了一种基于互信息的无监督特征选择方法 UFS-MI,该方法通过估算每个候选特征降低其他特征的不确定性的能力一次选择一个特征.在特征选择时,UFS-MI 从信息论的角度对特征的相关度和冗余度进行了综合考虑.特别地,与现有的无监督特征选择方法相比,UFS-MI 不要求特征间具有线性关系,因此有更广泛的应用场景.更进一步,我们分析了 UFS-MI 和有监督特征选择之间的联系,从理论上证明了算法的有效性,即其可以选择出与潜在类别变量高度相关的特征.实验结果表明,与传统的无监督特征选择方法相比,UFS-MI 具有更高的效率和性能.

参 考 文 献

[1] Langley P. Selection of relevant features in machine learning [C] //Proc of the AAAI Fall Symposium on Relevance. Menlo Park, CA: AAAI, 1994: 1-5

[2] Dash M, Liu H. Feature selection for classification [J]. International Journal of Intelligent Data Analysis, 1997, 1(3): 131-156

[3] Pudil P, Novovicova J. Novel methods for subset selection with respect to problem knowledge [J]. IEEE Intelligent Systems, 1998, 13(2): 66-74

[4] Robnik-Sikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF [J]. Machine Learning, 2003, 53(1): 23-69

[5] Hall M. Correlation-based feature selection for discrete and numeric class machine learning [C] //Proc of the 7th Int Conf on Machine Learning. San Francisco: Morgan Kaufmann, 2000: 359-366

[6] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312

[7] Wei H L, Billings S A. Feature subset selection and ranking for data dimensionality reduction [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(1): 162-166

[8] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy [J]. Journal of Machine Learning Research, 2004, 5(10): 1205-1224

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

[10] Dash M, Choi K, Scheuermann P, et al. Feature selection for clustering—A filter solution [C] //Proc of the 2nd IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2002: 115-122

[11] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238

[12] Almuallim H, Dietterich T G. Learning Boolean concepts in the presence of many irrelevant features [J]. Artificial Intelligence, 1994, 69(1/2): 279-305

[13] Dash M, Liu H. Consistency-based search in feature selection [J]. Artificial Intelligence, 2003, 151(1/2): 155-176

[14] Kohavi R, John G H. Wrappers for feature subset selection [J]. Artificial Intelligence, 1997, 97(1/2): 273-324

[15] Catal C, Sevim U, Diri B. Software fault prediction of unlabeled program modules [C] //Proc of the World Congress on Engineering. London, UK: IAENG, 2009: 212-217

[16] Dy J G, Brodley C E. Feature selection for unsupervised learning [J]. Journal of Machine Learning Research, 2004, 5(8): 845-889

[17] Law M H C, Figueiredo M A T, Jain A K. Simultaneous feature selection and clustering using mixture models [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2004, 26(9): 1154-1166

[18] Liu Tao, Wu Gongyi, Chen Zheng. An effective unsupervised feature selection method for text clustering [J]. Journal of Computer Research and Development, 2005, 42(3): 729-736 (in Chinese)
(刘涛,吴功宜,陈正.一种高效的用于文本聚类的无监督特征选择算法[J].计算机研究与发展,2003,42(3):729-736)

[19] Li Y H, Dong M, Hua J. Localized feature selection for clustering [J]. Pattern Recognition Letters, 2008, 29(1): 10-18

[20] Modha D S, Spangler W S. Feature weighting in k-means clustering [J]. Machine Learning, 2003, 52(3): 217-237

[21] Jensen J. Sur les fonctions convexes et les inégalités entre les valeurs moyennes [J]. Acta Mathematica, 1906, 30(1): 175-193

[22] Wang H, Bell D, Murtagh F Z. Relevance approach to feature subset selection [G] //Feature Extraction, Construction and Selection: A Data Mining Perspective. Norwell, MA: Kluwer Academic Publishers, 1998: 85-97

[23] Blake C, Merz C. UCI repository of machine learning database [EB/OL]. [2009-03-15]. htpp://www.ics.uci.edu/~mlearn/MLR epository.html

[24] Fayyad U M, Irani K B. Multi-interval discretization of continuous-valued attributes for classification learning [C] //Proc of the 13th Int Joint Conf on Artificial Intelligence. San Francisco: Morgan Kaufmann, 1993: 1022-1027

Xu Junling, born in 1984. PhD candidate. His main research interests include information retrieval, machine learning and pattern recognition.

Chen Lin, born in 1979. PhD. His current research interests include software analysis and software refactoring.

Zhou Yuming, born in 1974. PhD and professor. His main research directions are software metrics, program understanding and software maintenance.

Xu Baowen, born in 1961. Professor and PhD supervisor. His research interests include programming languages, software engineering and information retrieval.