

基于 Parzen 窗条件互信息计算的特征选择方法

贺邓超, 张宏军, 郝文宁, 张 睿

(解放军理工大学 指挥信息系统学院 仿真与数据中心, 南京 210007)

摘要: 为解决连续值特征条件互信息计算困难和对多值特征偏倚的问题,提出了一种基于 Parzen 窗条件互信息计算的特征选择方法。该方法通过 Parzen 窗估计出连续值特征的概率密度函数,进而方便准确地计算出条件互信息;同时在评价准则中引入特征离散度作为惩罚因子,克服了条件互信息计算对于多值特征的偏倚,实现了对连续型数据的特征选择。实验证明,该方法能够达到与现有方法相当甚至更好的效果,是一种有效的特征选择方法。

关键词: 特征选择; Parzen 窗; 条件互信息; 特征离散度

中图分类号: TP391.4 **文献标志码:** A **文章编号:** 1001-3695(2015)05-1387-03

doi:10.3969/j.issn.1001-3695.2015.05.027

Feature selection based on conditional mutual information computation with Parzen window

HE Deng-chao, ZHANG Hong-jun, HAO Wen-ning, ZHANG Rui

(Center of Simulation & Data, Institute of Command Information System, PLA University of Science & Technology, Nanjing 210007, China)

Abstract: In order to solve the problems of calculating the conditional mutual information of continuous variables and bias of multi-value features, this paper proposed a novel feature selection method. The method was based on computing conditional mutual information with Parzen window called PCMIFS, which adopted Parzen window to estimate the probability density function and compute conditional mutual information of continuous feature. And introduced a penalty factor, feature dispersion, to overcome the bias of multi-value features. The experiment results show that comparing several existing method, PCMIFS can attain better or comparable performance, and is an effective feature selection method.

Key words: feature selection; Parzen window; conditional mutual information; feature dispersion

0 引言

随着数据量的不断增大,特征选择已成为数据预处理特别是高维数据预处理过程中不可缺少的部分(如文本分类、图像检索、生物信息处理等)。特征选择即从一组数量为 D 的特征中选择出数量为 d ($D > d$) 的一组最优特征^[1],已达到用较少的特征对数据进行有效的表达和减少计算开销的目的,其关键在于如何在选择有用的特征的同时去除冗余或与类标签不相关的特征。目前常用的特征选择方法可以分为嵌入(embedded)法、封装(wrapper)法^[2]和过滤(filter)法^[3]。Embedded 法特征选择被整合进某一特定学习算法的训练过程中,特征选择与训练过程同步进行;Wrapper 法是以某一具体归纳学习算法的性能作为其评价和选择特征的标准,此类算法的不足在于继承了所使用的归纳学习算法对特征的偏置,计算复杂度比较高;Filter 法是基于某一具体的评价准则来选择特征,其过程独立于具体算法。大量实验表明,Filter 方法较 Wrapper 方法效率更高,更适合于大规模维度空间的特征选择任务。常用的 Filter 法的评价准则有卡方检验(χ^2 -test)^[4]、互信息^[5,6]、联合互信息亏损^[7]等。

基于信息论的评价准则能够很好地反映两个特征之间的

相关性,因此常作为特征选择的评价准则。文献[8]采用了条件互信息作为特征的评价准则,不仅考虑了特征与类标签之间的相关性,同时考虑到了特征之间的冗余性。这种方法在处理离散值特征时效果良好,但是对于连续值特征,需要将其进行离散化后再进行计算,不同的数据离散化区间划分会导致不同程度的原始数据信息丢失;另一种解决方案是假设数据服从某种概率密度分布,并在此条件下进行参数估计,然而在实际问题中,假设的概率密度分布很少符合实际情况^[9]。此外,基于信息论的相关性计算存在着偏向于取值较多的特征的缺点^[10]。这是因为随着某一特征的取值不断增多,其信息量也随之不断增大,但是这种特征对于后期分类、数据挖掘等处理的贡献度较小。综上所述,现有的基于信息论的特征选择方法存在以下不足:a)对于连续值特征缺乏有效的计算方法;b)存在偏向于取值较多的特征的现象。

针对上述不足,本文提出了一种基于 Parzen 窗密度估计的条件互信息计算方法,该方法通过 Parzen 窗估计出连续值特征的概率密度函数,进而方便准确地计算出条件互信息,并将该计算方法应用到具体的特征选择算法中,同时在构造特征选择的评价准则时引入特征离散度作为惩罚因子,克服了条件互信息计算对于多值特征的偏倚,实现了对连续型数据的特征

收稿日期: 2014-03-29; 修回日期: 2014-05-12

作者简介: 贺邓超(1989-),男,江苏南京人,博士研究生,主要研究方向为模式识别、数据挖掘(hdchao1989@163.com);张宏军(1963-),男,江苏姜堰人,教授,博导,主要研究方向为部队作战方法理论、军用数据工程等;郝文宁(1971-),男,山西运城人,教授,博士,主要研究方向为海量高维数据归约、作战效能评估;张睿(1977-),男,山东威海人,副教授,博士,主要研究方向为军事建模仿真。

选择。

1 相关理论

1.1 信息熵与条件互信息

信息论中,熵(entropy)是衡量一个随机变量取值的不确定性程度。离散随机变量 X 的熵定义^[11]为

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x) \quad (1)$$

当给定变量 X , 变量 Y 的不确定性程度可以用条件熵^[11]来表示:

$$H(Y|X) = - \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x) \quad (2)$$

为了更好地描述事物之间的普遍联系,信息论引进了互信息(mutual information)的概念。互信息反映了两个随机变量之间的关联程度,定义^[11]为

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \quad (3)$$

如果两个随机变量之间的互信息越大,则说明两个变量是高度相关的。根据互信息和熵的关系,互信息^[11]可由式(4)计算:

$$I(X;Y) = H(Y) - H(Y|X) \quad (4)$$

式(4)表明,当已知 X 时, Y 不确定性地减少量。

对于连续随机变量,熵、互信息可定义^[11]为

$$H(X) = - \int p(x) \log_2 p(x) dx \quad (5)$$

$$I(X;Y) = \int p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} dx dy \quad (6)$$

条件互信息(conditional mutual information)表示在给定离散随机变量 Z 的条件下,两个离散随机变量 X 与 Y 之间的相关程度,定义为 $I(X;Y|Z)$ 。若条件互信息值越大,则说明在给定 Z 条件下 X 和 Y 的关联程度越大。在特征选择中,设 S 为已选特征子集,那么类标签 C 与待选特征 f_i 的条件互信息 $I(C;f_i|S)$ 可用式(7)^[9]表示:

$$I(C;f_i|S) = I(C;S,f_i) - I(C;S) \quad (7)$$

通过式(7)可将条件互信息转换为互信息进行计算。

1.2 Parzen 窗密度估计原理

Parzen 窗估计是一种具有坚实理论基础和优秀性能的非参数函数估计方法。该方法能够较好地描述一维或多维数据的分布状态,利用一组样本对总体概率密度函数进行估计,其基本思想是利用一定范围内各点密度的平均值对总体密度函数进行估计。设 x 为 d 维空间中某一点, N 为样本总数,为了对 x 概率密度 $p(x)$ 进行估计,以 x 为中心做一个边长为 h 的超立方体 V ,则其体积为 $V = h^d$,构造窗函数 $\varphi(u)$ 如下^[12]:

$$\varphi(u) = \begin{cases} 1 & |u_i| \leq \frac{1}{2}, i=1,2,\dots,d \\ 0 & \text{other} \end{cases} \quad (8)$$

$\varphi(u)$ 满足条件 $\varphi(u) \geq 0$, 且 $\int \varphi(u) du = 1$, 则落入体积 V

中的样本数为 $N_v = \sum_{i=1}^N \varphi(\frac{x-x_i}{h})$, 根据落入窗类的样本点,估计出 x 的概率密度为

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{V} \varphi(\frac{x-x_i}{h}) \quad (9)$$

式(9)为 Parzen 窗密度估计的基本公式。

2 基于 Parzen 窗概率密度估计的条件互信息计算

对于现实世界的数据集,类标签通常是离散值,但是特征

变量往往是连续值,此时,通过式(7)将条件互信息转换为互信息 $I(C;S,f_i)$ 、 $I(C;S)$ 进行计算。由式(4)可得

$$I(C;S,f_i) = H(C) - H(C|S,f_i) \quad (10)$$

$$I(C;S) = H(C) - H(C|S) \quad (11)$$

由于类标签是离散值,式中类标签的熵 $H(C)$ 可以由式(1)直接得到;式中条件熵 $H(C|S,f_i)$ 、 $H(C|S)$ 可由式(2)得到(为了表示方便简洁,条件熵统一表示为 $H(C|X)$)^[13]:

$$H(C|X) = - \int p(x) \sum_{i=1}^N P(c_i|x) \log_2 P(c_i|x) dx \quad (12)$$

其中: N 是类标签的数目。由于 x 为连续值,很难估计后验概率 $P(c|x)$, 因此无法直接计算。为了解决这一困难,本文提出一种基于 Parzen 窗概率密度估计的条件互信息计算方法。

由于条件概率的加和为 1, 即 $\sum_{k=1}^N p(c_k|x) = 1$, 根据贝叶斯公式,后验概率可以表示为

$$P(c_i|x) = \frac{P(c_i|x)p(x|c_i)}{\sum_{k=1}^N P(c_k|x)p(x|c_k)} = \frac{P(c_i)p(x|c_i)}{\sum_{k=1}^N P(c_k)p(x|c_k)} \quad (13)$$

由式(13)可以看出,计算的关键在于如何估计连续值特征的概率密度函数 $p(x|c_i)$ 。

本文采用 Parzen 窗对连续值特征进行概率密度估计。根据式(9)可得估计值为

$$\hat{p}(x|c_i) = \frac{1}{n_{c_i}} \sum_{j \in I_{c_i}} \varphi(\frac{x-x_j}{h}) \quad (14)$$

式中: I_{c_i} 是属于类 c_i 的样本集合, $n_{c_i} = |I_{c_i}|$; h 为窗宽参数。将 $p(x|c_i)$ 的估计式(14)代入式(13),便得到后验概率 $P(c_i|x)$ 的估计公式:

$$\hat{P}(c_i|x) = \frac{\sum_{j \in I_{c_i}} \varphi(\frac{x-x_j}{h_{c_i}})}{\sum_{k=1}^N \sum_{j \in I_{c_k}} \varphi(\frac{x-x_j}{h_{c_k}})} \quad (15)$$

其中: h_{c_i} 、 h_{c_k} 为窗宽参数。

在 Parzen 窗密度估计的基本公式中,窗宽 h 是一个非常重要的参数。当样本数 N 有限时, h 对估计的效果有着较大的影响。一般可以选择的窗函数有方窗、正态窗等。本文选择正态窗作为窗函数,主要基于以下考虑:

- 正态函数的平滑性将使得估计函数变化平滑。
- 如果选择完全对称的正态函数,估计函数中只有一个参量变化。

在选择正态窗函数的情形下,式(15)表示为

$$\hat{P}(c_k|x) = \frac{\sum_{j \in I_{c_k}} \exp(-\frac{(x-x_j)^T \Sigma^{-1}(x-x_j)}{2h^2})}{\sum_{k=1}^N \sum_{j \in I_{c_k}} \exp(-\frac{(x-x_j)^T \Sigma^{-1}(x-x_j)}{2h^2})} \quad (16)$$

将式(15)代入式(12)便可得到条件熵的估计公式为

$$\hat{H}(C|X) = - \sum_{j=1}^n \frac{1}{n} \sum_{c_i=1}^N \hat{P}(c_i|x_j) \log_2 \hat{P}(c_i|x_j) \quad (17)$$

式中: n 为样本总数。

综上所述,利用上述方法可较为方便准确地估计连续值特征的概率密度函数,从而实现对连续值特征的条件互信息的计算。

3 基于 Parzen 窗条件互信息计算的特征选择算法

通常,当一个特征含有能够影响类标签分布的信息时,该特征为相关特征,否则该特征为不相关特征或冗余特征。由 2.1 节可知,给定已选特征集 S ,若有 $I(C;f|S) > 0$, 则 f 是一个相关特征,且具有 S 所不具有的关于类标签 C 信息;若 $I(C;f|$

$S)=0$,则特征 f 在给定 S 时不具有关于类标签的任何有用信息,即 f 是一个不相关特征,或者是在给定特征集 S 时的冗余特征。因此,条件互信息 $I(C;f|S)$ 不仅考察了特征与类标签之间的相关性,而且考虑了特征之间的相关性(冗余性判别),因此本文将作为选择特征时的评价准则。

经过大量的实验发现,以互信息为基础的评价准则(包含条件互信息)在计算的过程中,有偏向于取值较多的特征的缺点,但是这些多值特征并不是最优选择。针对这一不足,本文引入特征离散度作为惩罚因子,与条件互信息相结合作为特征选择的评价准则。根据散布矩阵(scatter matrix)的概念,推导出特征离散度的定义。

定义1 若考虑先验概率,定义某一特征 f 的离散度为

$$D_f = \sum_{i=1}^N [P_{c_i} \sum_{k \in I_{c_i}} (d^{(k)})^2] \quad (18)$$

式中: P_{c_i} 表示类 C_i 的先验概率, $d^{(k)}$ 表示样本数据 x_k 与样本均值的距离。

定义2 给定某一样本数据集,定义 $d^{(k)}$,其公式为

$$d^{(k)} = d(x_k, u) \quad (19)$$

式中: $d(\cdot)$ 表示距离函数; x_k, u 分别表示特征 j 下类 C_i 的样本数据和样本均值。对于离散值特征,均值 u 即为其数学期望 $E[X]$ 。

$$\mu = E[X] = \sum_j xP(x) \quad (20)$$

对于不同类型的离散值,可以按照如下方法进行计算:a)如果样本数据为二元或分类变量:如果 $x_k = u, d^{(k)} = 0$,否则 $d^{(k)} = 1$;b)如果样本数据为序数变量:计算秩 r_k 和 $Z_k = \frac{r_k - 1}{M_j - 1}$,并将 Z_k 作为区间标度变量对待。对于连续值特征, $d^{(k)} = x_k - u$ 。

综上所述,由条件互信息的定义,结合式(18),得到特征选择算法的评价准则为

$$J(f_i) = \frac{I(C;f_i|S)}{D_{f_i} + 1} \quad (21)$$

式(21)将特征离散度 $D_{f_i} + 1$ 作为惩罚因子(考虑到样本数据存在单一取值,即 $D_{f_i} = 0$ 的情况),此时如果某一特征样本数据取值越多,则特征离散度 D_{f_i} 越大,评价准则相应地越小,以此克服条件互信息对于多值特征的偏倚。因此,评价准则 $J(f_i)$ 越大,则说明特征 f_i 与类标签的相关性越高,被选择的可能性也越大。给出该特征选择算法的伪代码如下:

```

Input: A training dataset  $U(F, C)$ 
Desired number of features  $N$ .
Output: Selected feature subset  $S$ 
1 Initialize:  $S = \emptyset$ 
2  $\forall f_i \in F$ , compute  $\frac{I(f_i; C)}{D_{f_i} + 1}$ 
3 (Selection of the first feature) find the feature that maximizes  $\frac{I(f_i; C)}{D_{f_i} + 1}$ , set  $F = F - \{f_i\}, S = S + \{f_i\}$ 
4 repeat until desired number of features  $N$  are selected
4.1  $\forall f_i \in F$ , compute  $J(f_i)$ 
4.2 (Selection of the next feature) choose the feature  $f_i \in F$  that maximizes  $J(f_i)$ , and set  $F = F - \{f_i\}, S = S + \{f_i\}$ 
5 Output the preliminary set  $S$  containing the selected features

```

该算法采用一种前向选择的启发式搜索策略,每次选择使 $J(f_i)$ 最大的特征进入最优特征子集。算法通过上文提出的Parzen窗条件互信息计算方法准确计算其条件互信息 $I(C;f_i|S)$,得到评价准则 $J(f_i)$,实现对连续值特征的特征选择,同时

克服了条件互信息计算中对多值特征的偏倚。

4 实验与结果分析

本文将所提出的基于Parzen窗条件互信息计算的特征选择算法应用到实际的数据集,并通过实验结果对比分析该算法的性能。所有实验均在Windows XP SP3操作系统平台上进行,使用MATLAB R2008a实现特征选择算法的具体实现,使用SPSS Clementine 12.0进行分类实验。在算法的实现中设定Parzen窗宽参数 $h = \frac{1}{\log n}$,其中 n 为样本数。为了便于表示,将该算法称为PCMIFS(parzen window & conditional mutual information feature selector)。

4.1 Sonar数据集

Sonar数据集在文献[6,13]中用来测试其特征选择方法的性能。数据集样本总数为208,特征数目为60,类标签数目为2:metal、rock,其中,metal类包含的样本数为111,rock类包含的样本数为97。使用PCMIFS算法对数据集进行特征选择,并选择2~10个特征用来进行分类实验。分类实验参数^[13]为:以神经网络为模型,隐藏层为一层且节点数为3,冲量Alpha设为0.0,学习率Eta取0.2,衰减为5,训练次数为300次。为防止过度训练取各数据集的50%作为样本集,其余50%作为测试集。

图1显示了PCMIFS在Sonar数据集上的特征选择效果,图中柱条表示PCMIFS的特征选择结果,每个柱条上的数字表示所选择的特征编号;横轴下方的数字表明了选择特征的次序,竖轴表示选择特征后对于Sonar数据集的分类正确率。

在图2中比较了PCMIFS、PWFS、MIFS和MIFS-U四种方法对于Sonar数据集的效果。实验采用10次十折交叉验证的方式,所有的分类错误率结果为10次结果的均值,括号中为标准差。从图2可以看出,PCMIFS显示了比其他算法更好的效果。

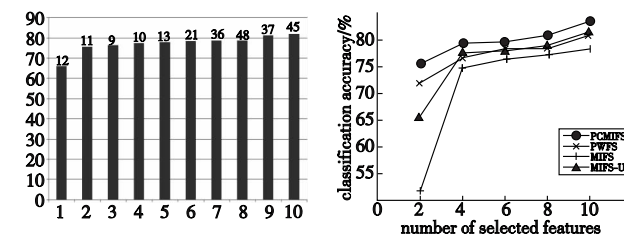


图1 Sonar数据集上特征选择次序与分类准确率的对比

图2 不同特征选择方法对于Sonar数据集的分类准确率/%

4.2 其他UCI数据集

选取UCI库中常用的四个数据集来测试与比较各特征选择算法的性能。数据集的简要描述如表1所示。

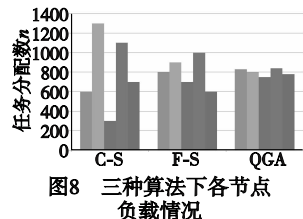
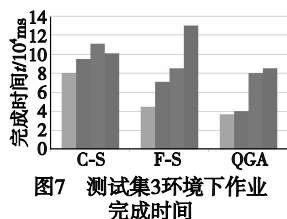
表1 实验数据集描述

Name	#features	#instances	#classes	Name	#features	#instances	#classes
Letter	16	20000	26	Waveform	21	1000	3
Breast Cancer	9	699	2	Vehicle	18	946	4

对于这些数据集,分别按照特征选择算法选择不同的特征进行分类实验,结果如表2所示。分类实验参数同上,其中对于Letter数据集,将数据集的75%作为训练集,剩余25%作为测试集;对于Breast Cancer数据集,将数据集的50%作为训练集,剩余50%作为测试集;对于Waveform数据集,将数据集的30%作为训练集,70%作为测试集;对于Vehicle数据集,将数据集的33%作为训练集,67%作为测试集。为(下转第1398页)

过慢的作业。

3) 在测试集 3 中取 4000 任务分别在三种环境下运行, 图 8 显示了三种环境下各节点任务的分配数, 从图 8 中可以得出, QGA 的负载均衡能力明显强于 Capacity Schedule 和 Fair Schedule。



4 结束语

本文针对 Hadoop 云计算平台资源调度存在的问题提出了一种解决方案, 方案中采用了基于 QoS 的作业调度方法和基于遗传算法的任务调度方法, 该方法不仅提高了 Hadoop 资源调度器的服务水平, 而且实现了负载均衡, 仿真结果表明, 该方案有很好的效果。但在本文的研究中没有考虑多机架集群问题, 不利于 QGA 算法的扩展性, 下一步将重点研究多机架集群的调度方法。

参考文献:

- [1] Hadoop[EB/OL]. [2014-2-01]. <http://hadoop.apache.org/>.
- [2] GHEWAWAT S, GOBIOFF H, LEUNG S T. The google file system [J]. *ACM SIGOPS Operating Systems Review*, 2003, 37(5): 29-43.
- [3] DEAN J, GHEWAWAT S. MapReduce: simplified data processing on large clusters [J]. *Communications of the ACM*, 2008, 51(1): 107-113.
- [4] SHVACHKO K, KUANG H, RADIA S, et al. The Hadoop distributed

file system[C]//Proc of the 26th IEEE Symposium on Mass Storage Systems and Technologies. [S.l.]: IEEE Press, 2010: 1-10.

- [5] ZAHARIA M, BORTHAKUR D, SARMA J S, et al. Job scheduling for multi-user mapreduce clusters[R]. Berkeley: EECS Department, University of California, 2009: 1-16.
- [6] 董西成. Hadoop 技术内幕: 深入解析 MapReduce 架构设计与实现原理[M]. 北京: 机械工业出版社, 2013.
- [7] 夏炜. Hadoop 平台下的作业调度算法研究与改进[D]. 广州: 华南理工大学, 2010.
- [8] TIAN C, ZHOU H, HE Y, et al. A dynamic mapreduce scheduler for heterogeneous workloads[C]//Proc of the 8th International Conference on Grid and Cooperative Computing. Washington DC: IEEE Computer Society, 2009: 218-224.
- [9] POLO J, CARRERA D, BECERRA Y, et al. Performance-driven task co-scheduling for mapreduce environments[C]//Proc of Network Operation and Management Symposium. [S.l.]: IEEE Press, 2010: 373-380.
- [10] KC K, ANYANWU K. Scheduling hadoop jobs to meet deadlines [C]//Proc of the 2nd IEEE International Conference on Cloud Computing Technology and Science (CloudCom). 2010: 388-392.
- [11] SANDHOLM T, LAI K. Dynamic proportional share scheduling in hadoop[C]//Proc of Job Scheduling Strategies for Parallel Processing. Berlin: Springer, 2010: 110-131.
- [12] 陈国良, 王煦法, 庄镇泉, 等. 遗传算法及其应用[M]. 北京: 人民邮电出版社, 1996.
- [13] 张文修, 梁怡. 遗传算法的数学基础[M]. 西安: 西安交通大学出版社, 2003.
- [14] 张钊宁. 数据密集型计算机中任务调度模型的研究[D]. 长沙: 国防科学技术大学, 2009.
- [15] 刘愉, 赵志文, 李小兰, 等. 云计算环境中优化遗传算法的资源调度策略[J]. *北京师范大学学报*, 2012, 48(4): 378-383.
- [16] 杨利宏. 基于遗传算法的资源约束型项目调度问题的优化[D]. 上海: 上海交通大学, 2007.

(上接第 1389 页) 了便于对比结果, 所有数据集均选择特征选择后前四个相关特征。分析表 2 可以看出, PCMIFS 在分类准确率上要优于 PWFS、MIFS 与 MIFS-U。

表 2 特征选择结果对于各实验数据集的分类正确率比较 /%

Datasets	PCMIFS	PWFS	MIFS	MIFS-U
Letter	70.1	67.5	62.4	68.5
Breast Cancer	96.8	96.6	93.7	94.2
Waveform	77.8	75.4	67.6	73.8
Vehicle	64.8	62.5	57.3	59.9

5 结束语

本文提出了一种基于 Parzen 窗条件互信息计算的特征选择方法。该方法通过 Parzen 窗估计出连续值特征的概率密度函数, 实现条件互信息的计算; 同时在特征选择的评价准则中引入特征离散度作为惩罚因子, 克服了条件互信息计算对于多值特征的偏倚。通过实验结果对比分析, 该算法能够较好地估计出连续值特征的条件互信息, 在特征选择中比现有算法 PWFS、MIFS 和 MIFS-U 有着更好的效果。

参考文献:

- [1] 毛勇, 周晓波, 夏铮, 等. 特征选择算法研究综述[J]. *模式识别与人工智能*, 2007, 20(2): 211-221.
- [2] TABAKHI S, MORADI P, AKHLAGHIAN F. An unsupervised feature selection algorithm based on ant colony optimization[J]. *Engi-*

neering Applications of Artificial Intelligence, 2014, 32: 112-123.

- [3] BOLÓN-CANEDO V, PORTO-DÍAZ I, SÁNCHEZ-MARONÓ N, et al. A framework for cost-based feature selection [J]. *Pattern Recognition*, 2014, 47(7): 2481-2489.
- [4] QU G, HARIRI S, YOUSIF M. A new dependency and correlation analysis for features [J]. *IEEE Trans on Knowledge and Data Engineering*, 2005, 17(9): 1199-1207.
- [5] LIU H, SUN J, LIU L, et al. Feature selection with dynamic mutual information [J]. *Pattern Recognition*, 2009, 42(7): 1330-1339.
- [6] BATTITI R. Using mutual information for selecting features in supervised neural net learning [J]. *IEEE Trans on Neural Networks*, 1994, 5(4): 537-550.
- [7] 张逸石, 陈传波. 基于最小联合互信息亏损的最优特征选择算法[J]. *计算机科学*, 2011, 38(12): 200-206.
- [8] SOTOCA J, PLA F. Supervised feature selection by clustering using conditional mutual information based distances[J]. *Pattern Recognition*, 2010, 43(6): 2068-2081.
- [9] 边肇祺, 张学工. 模式识别[M]. 2 版. 北京: 清华大学出版社, 2001.
- [10] HAN Jia-wei, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2007.
- [11] COVER T, THOMAS J. Elements of Information Theory[M]. New-York: Wiley, 1991.
- [12] RICHARD O, DUDA P E, HART D G. 模式分类[M]. 李宏东, 等译. 北京: 机械工业出版社, 2003.
- [13] KWAK N, CHOI C H. Input feature selection by mutual information based on Parzen window [J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2002, 24(12): 1667-1671.