文章编号: 1001-0920 (2015) 05-0831-08 DOI: 10.13195/j.kzyjc.2014.0341

基于 ISE 的核密度估计和随机置换的单一或协同特征的选择方法

张景祥[a,b], 王士同[b], 蒋亦樟[b], 倪彤光[b]

(江南大学 a. 理学院，b. 数字媒体学院，江苏 无锡 214122)

摘 要: 针对数据的特征存在单一和协同特征的选择问题, 基于平方误差标准核密度估计和随机置换理论, 首先提出一种针对单一特征的特征选择方法 (FSKDE-RP); 然后, 针对协同特征的情况, 通过拓展随机置换理论, 提出多维协同特征选择算法 (SFSKDE-MRP), 并利用核神经网络 (KNN) 分类器的分类精度选择最优特征子集. 在模拟数据和真实数据集上的实验结果表明了所提出算法的有效性.

关键词: 特征选择；核密度估计；平方误差；随机置换

中图分类号: TP391.4 **文献标志码:** A

Feature selection method based on the ISE kernel density estimation and random permutation for single or synthetic features

ZHANG Jing-xiang[a,b], *WANG Shi-tong*[b], *JIANG Yi-zhang*[b], *NI Tong-guang*[b]

(a. School of Science，b. School of Digital Media，Jiangnan University，Wuxi 214122，China. Correspondent: ZHANG Jing-xiang，E-mail：zjx145@163.com)

Abstract: For the single or synthetic features selection problems, based on the integrated square error(ISE) criterion and random permutation, a supervised feature ranking criterion of a single feature is proposed firstly. Then, such a random permutation is extended to synergetic features, and accordingly the synergetic feature selection method is developed. Finally, the optimal feature subset is determined by the classification accuracy obtained by the kernel neural network(KNN) method. Experimental results on synthetic and real datasets show the effectiveness of the proposed algorithm.

Keywords: feature selection；kernel density estimation；integrated squared error；random permutation learning

0 引 言

特征选择是机器学习领域的重要研究问题之一. 特征选择是从一组原始特征集合中挑选出一些最具有代表性的特征子集使得系统的特定指标最优化. 特征选择作为数据预处理方法, 能有效降低特征空间的维数, 减小样本中不相关或冗余特征对算法的负面影响, 有效地提高应用算法的空间和时间效率[1-2], 并且在一定程度上避免“维数灾难”问题.

根据与学习算法的结合方式, 特征选择算法一般可分为封装式 (Wrapper) 和过滤式 (Filter). 过滤式方法独立于具体的学习算法, 而封装式方法将学习算法作为其在特征选择过程中的评价标准. Wrapper 方法具有代表性的工作有: Kira 等[3]提出的 Relief 算法及 Kononenko[4]提出的可以解决多类问题的 Relief-F 算法; Hsu[5]利用遗传算法寻找最小的特征子集; Huang 等[6]使用混合的遗传算法与分类器一起获取特征子集; Krzysztof 等[7]基于相互关系的双重策略提出特征选择方法; Song 等[8]以 HSIC 度量特征与类标记之间的依赖性, 提出一种特征选择框架. Filter 方法代表性的工作有: Jain 等[9]利用类别和特征之间的互信息进行特征排序, 取少量特征作为所选择的特征子集; Geng 等[10]利用小波变换, 通过核密度估计的方法提出特征选择策略; He 等[11]利用拉普拉斯变换, 通过比较每个特征的得分 LS (Laplacian score) 来进行特征排序; Peng 等[12]在提出的 mRMR 选择算法中使用互信息估算候选特征与分类类别和已选特征之间的相关性和冗余性.

上述方法虽然取得了一定的效果, 但也存在一些

收稿日期: 2014-03-12；**修回日期:** 2014-06-26.

基金项目: 国家自然科学基金项目(61202311, 61272210)；江苏省研究生培养创新工程项目(CXLX13748).

作者简介: 张景祥(1977−), 男, 博士生, 从事模式识别、人工智能的研究；王士同(1964−), 男, 教授, 博士生导师, 从事模式识别、人工智能等研究.

不足之处: 1) 文献 [3, 5-7] 的算法一般都要通过模型参数优化进而选出特征子集, 该优化步骤往往时间复杂度较高且泛化能力较差, 并要求数据规模较小; 2) 文献 [8-11] 的算法利用搜索方式进行特征选择, 评价标准以及参数依赖性较大, 进而使得特征选择的效率较低, 并与后续学习算法的学习性能偏差较大.

针对上述问题, 基于ISE的核密度估计和随机置换理论, 本文提出一种新的过滤式特征选择算法, 即随机置换的核密度估计特征选择方法(FSKDE-RP). 该算法把原始特征集的特征逐个进行随机置换而其他特征保持不变, 以特征随机置换前后所得之再生核Hilbert空间(RKHS)概率密度之间的积分平方误差(ISE)距离作为度量标准来确定特征的重要性, ISE的数值越大说明此特征越重要, 最终根据特征重要性的排序选取特征子集. ISE方法虽然已被广泛应用于特征选择, 但大都需要利用特征加权的方法对特征进行变换[9-12], 且所得的特征子集严重依赖于加权方法和权系数, 参数较多, 优化过程繁琐, 时间复杂度较高. FSKDE-RP方法基于随机置换理论对特征进行变换, 以变换前后的积分平方误差(ISE)距离作为评价标准, 其优势在于积分平方误差距离在再生核Hilbert空间上可以实现完全核化, 计算效率高, 不受特征加权技术的限制, 通用性好. 另外, 本文针对实际数据集协同特征的现象, 将FSKDE-RP方法推广到多维协同特征选择的问题中, 得到了多维协同特征同时选择的新算法(SFSKDE-MRP).

本文以KNN算法作为检测所选特征好坏的评价分类器, 将所提出的FSKDE-RP和SFSKDE-MRP方法与Relief、Relief-F、mRMR和mREL等经典特征选择算法进行比较. 与现有相关方法相比, 本文方法的特色之处在于: 1) 实现了在再生核Hilbert空间中积分平方误差(ISE)距离的完全核化, 使ISE作为一种特征排序标准, 几何结构意义更加清晰, 适应性更强; 2) 基于随机置换理论的特征变换方法, 参数较少, 无需优化; 3) 针对实际数据集协同特征现象, 把单一特征选择拓展应用到多维协同特征同时选择的问题中, 有效地提高了特征选择的效率.

1 基于核密度估计的特征选择

1.1 核密度估计

核密度估计(KDE)是一种非参数密度估计方法, 在统计学理论和其他相关的应用领域受到了高度的重视.

定义 1 Parzen窗方法核密度估计函数可写成

$$p(x)=\sum_{i=1}^{n}\alpha_i K_\sigma(x,x_i);$$

$$\text{s.t.}\ \sum_{i=1}^{n}\alpha_i=1,\ \alpha_i\geqslant 0,\ i=1,2,\cdots,n. \tag{1}$$

其中: $K_\sigma(x,x_i)$是窗宽为σ的核函数, $\alpha_i\,(i=1,2,\cdots,n)$为系数, x_i为训练样本, x为预测样本. 核函数$K_\sigma(x,x_i)$必须满足以下条件: 1) $K_\sigma(t)\geqslant 0$; 2) $\int K_\sigma(t)\mathrm{d}t=1$.

常用的核函数有: Gauss函数、Epanechnikov函数、Biweight函数等. 在再生核Hilbert空间(RKHS)中, Gauss函数等价于一个线性函数, 并且Gauss函数的定义域是$(-\infty,+\infty)$, 其具有单峰、对称、计算方便等优良特性, 因此本文选择Gauss函数作为核函数.

1.2 基于随机置换核密度估计的特征排序准则

特征选择就是要根据评估标准, 选择出特征数量较少、评估效果较好的特征子集. 积分平方误差(ISE)作为特征选择常用的方法, 其定义形式为

$$\mathrm{ISE}(x)=\int(p(x)-\hat{p}(x))^2\mathrm{d}x.$$

其中: $p(x)$表示原数据集的概率密度, $\hat{p}(x)$表示特征经过变换后数据集的概率密度. 由定义1可知, 通过某种非线性映射, 把数据从原始的非线性空间映射到维数更高的线性空间, 基于核密度估计的ISE可作为一种度量特征选择前后数据集概率密度之间的差异的标准.

近几年, 随机森林理论(RF)[13]被用于进行特征变换, 取得了一定效果. 文献[14]通过分析证明了数据集的特征向量的第k维特征上的N个样本点完全随机置换, 其他特征保持不变时, 根据后验概率原理得到的新数据集的属性不改变, 可以写成如下定义.

定义 2(随机置换)[14] 设$\varsigma_i\sim U(0,1)$, 即在区间$(0,1)$上均匀产生$\varsigma_1,\varsigma_2,\cdots,\varsigma_{N-1}$, $m=\lfloor\varsigma\rfloor$表示向下取整, $m=\lfloor N*\varsigma_k\rfloor+1, k=1,2,\cdots,N-1$. 若交换$x_i^k$和$x_i^m$, 则称将第$k$维特征上数据完全随机置换.

引理 1[15] 若对数据集$\boldsymbol{X}$上某一特征进行完全随机置换, 则等价于删除该维特征.

1.2.1 单一特征随机置换核密度估计

由定义2和引理1, 本文提出一种基于核密度估计的特征排序准则, 即: 设数据集$\boldsymbol{X}\in\boldsymbol{R}^d$, $p(x)$表示数据集的概率分布函数, $\hat{p}_{(k)}(x)$表示数据集特征向量的第k维特征随机置换后的概率密度函数, 定义距离$\Re(k)$如下:

$$\Re(k)=\int(p(x)-\hat{p}_{(k)}(x))^2\mathrm{d}x. \tag{2}$$

由ISE的定义形式可知, 式(2)表示数据集的第k维特征随机置换前后数据集概率分布之间的差异, $k=1,2,\cdots,d$. 若数据集是二分类数据, 则$p(x)$和$\hat{p}_{(k)}(x)$可以分解为正负类概率之和, 根据统计学中的全概率定理, 分别表示为

$$p(x)=p(c^+)p(x|c^+)+p(c^-)p(x|c^-)=\Big(\frac{N^+}{N^++N^-}\Big)p(x|c^+)+\Big(\frac{N^-}{N^++N^-}\Big)p(x|c^-),\tag{3}$$

$$\hat{p}_{(k)}(x)=p(c^+)\hat{p}_{(k)}(x|c^+)+p(c^-)\hat{p}_{(k)}(x|c^-)=\Big(\frac{N^+}{N^++N^-}\Big)\hat{p}_{(k)}(x|c^+)+\Big(\frac{N^-}{N^++N^-}\Big)\hat{p}_{(k)}(x|c^-).\tag{4}$$

其中: N^+ 和 N^- 分别表示数据集中正类和负类的样本容量, c^+ 和 c^- 分别表示数据的类标. 若令

$$\alpha_1=\Big(\frac{N^+}{N^++N^-}\Big),\ \alpha_2=\Big(\frac{N^-}{N^++N^-}\Big),$$

则置换前后概率密度之间的 $\Re(k)$ 距离为

$$\Re(k)=\int(p(x)-\hat{p}_{(k)}(x))^2\mathrm{d}x=\alpha_1^2\int(p^+(x)-\hat{p}^+_{(k)}(x))^2\mathrm{d}x+\alpha_2^2\int(p^-(x)-\hat{p}^-_{(k)}(x))^2\mathrm{d}x+2\alpha_1\alpha_2\int(p^+(x)-\hat{p}^+_{(k)}(x))(p^-(x)-\hat{p}^-_{(k)}(x))\mathrm{d}x.\tag{5}$$

利用 Parzon window 估计方法, 将 $p(x)$ 和 $\hat{p}_{(k)}(x)$ 映射到RKHS 中, 则 $\Re(k)$ 可表示为

$$\begin{aligned}\Re(k)=\alpha_1^2\Big(&\sum_{i=1}^N\sum_{j=1}^N\boldsymbol{K}_{\sigma_+}(x_i^+,x_j^+)-2\sum_{i=1}^N\sum_{j=1}^N\boldsymbol{K}_{\sigma_+}(x_i^+,x^+_{(k),j})+\sum_{i=1}^N\sum_{j=1}^N\boldsymbol{K}_{\sigma_+}(x^+_{(k),i},x^+_{(k),j})\Big)+\\&\alpha_2^2\Big(\sum_{i=1}^N\sum_{j=1}^N\boldsymbol{K}_{\sigma_-}(x_i^-,x_j^-)-2\sum_{i=1}^N\sum_{j=1}^N\boldsymbol{K}_{\sigma_-}(x_i^-,x^-_{(k),j})+\sum_{i=1}^N\sum_{j=1}^N\boldsymbol{K}_{\sigma_-}(x^-_{(k),i},x^-_{(k),j})\Big)+\\&2\alpha_1\alpha_2\Big(\sum_{i=1}^N\sum_{j=1}^N\boldsymbol{K}(x_i^+,x_j^-)-\sum_{i=1}^N\sum_{j=1}^N\boldsymbol{K}(x_i^+,x^-_{(k),j})-\sum_{i=1}^N\sum_{j=1}^N\boldsymbol{K}_{2\sigma}(x^+_{(k),i},x_j^-)+\sum_{i=1}^N\sum_{j=1}^N\boldsymbol{K}(x^+_{(k),i},x^-_{(k),j})\Big).\end{aligned}\tag{6}$$

计算整个特征空间上每一个特征在RKHS 中的 $\Re(k)$, $\Re(k)$ 的数值越大, 其对应特征的重要性越高, 进而每个特征可以按 $\Re(k)$ 数值的大小进行排序.

定义 3　形如式 (6), 以再生核 Hilbert 空间中密度估计差的大小作为特征排序评价函数, 本文称之为基于随机置换的核密度估计特征排序准则.

综上所述, 可以得到如下 FSKDE-RP 算法.

算法 1　FSKDE-MRP 算法.

输入: 带标签数据集 $\boldsymbol{S}$, $g=1$;

输出: $\tilde{\Re}=\{\Re(1),\cdots,\Re(d)\}$.

1) 给定数据集

$$\boldsymbol{S}=\{(\boldsymbol{x}_n,\boldsymbol{y}_n)\in\boldsymbol{X}'\times\boldsymbol{Y}:1\leqslant n\leqslant N\}.$$

其中: $\boldsymbol{X}'\in\boldsymbol{R}^d$ 表示样本的特征空间, $X^i=(x_1^i,x_2^i,\cdots,x_N^i)$ 表示数据 $\boldsymbol{X}'$ 上第 i 个特征上的数据, $\boldsymbol{Y}$ 表示样本的标签空间, (x_i,y_i) 为数据集中的一个样本.

2) 在区间 $(0,1)$ 上均匀产生 $\varsigma_1,\varsigma_2,\cdots,\varsigma_{N-1}$, $m=\lfloor\varsigma\rfloor$ 表示向下取整, $m=\lfloor N*\varsigma_k\rfloor+1$, $k=1,2,\cdots,N-1$. 交换 x_k^i 和 x_m^i, 将得到的随机序列 $X^i=(x_1^i,x_2^i,x_k^i,x_m^i,\cdots,x_N^i)$ 上的数据完全随机置换, 得到 $X^i=(x_1^i,x_m^i,x_k^i,\cdots,x_N^i)$, $i=1,2,\cdots,d$.

3) 按式 (6) 计算每一特征置换前后数据在RKHS 中密度估计差 $\Re$, 返回2). 令 $g\leftarrow g+1$, 直到 $g=d$ 得到全部特征的 $\Re$ 数值.

4) 按照 $\Re$ 的大小进行特征排序, 输出特征序列子集 $\tilde{\Re}$.

5) 根据 KNN 分类器的分类精度选择特征维数.

1.2.2　基于协同性的多维随机置换核密度估计

在数据挖掘过程中很多实际数据具有协同性. 例如: 血压检查包括收缩压和舒张压, 血常规检查中有高密度胆固醇和低密度胆固醇等指标. 它们都是特征向量中具有不同角度的分量特征. 正是由于数据特征向量具有这样的协同性, 增加了数据的特征空间维数, 降低了学习算法的效率和性能. 对此, 本文提出一种协同多维随机置换的核密度估计准则.

定义 4　协同特征选择是指数据集的特征空间中某些特征若共同出现, 则按相同的随机置换关系同时进行置换, 并按特征重要性进行特征选择.

本文提到的协同特征与冗余特征是两个不同的概念. 冗余特征是指某些特征分量对分类没有影响或者有些特征之间有很大的关联性; 而协同特征是指实际应用中数据特征统计过程中同时出现的特征, 与它们之间是否相关无关. 已有的算法中大都是从降低冗余和减少不相关的角度提出很多算法, 而本文基于协同特征现象, 以全新视角进行多维特征同时降维, 可以提高应用算法的性能. 例如可以降低算法的计算代

价、模型更易理解且适用性更强. 因此, 给出如下定理.

定理 1 若数据集 $\boldsymbol{X} \in \boldsymbol{R}^d$ 存在 m 维协同特征, 其余 $d-m$ 维相互独立, 则将 m 维以同样规则进行随机置换, 等价于删除此 m 维.

证明 设 $\boldsymbol{X} \in \boldsymbol{R}^d$, d 表示数据集的特征向量维数, 若已知其中 m 维是协同特征, 其余 $d-m$ 维相互独立, 则: $\boldsymbol{X}_m$ 表示 m 维协同特征的数据, $\boldsymbol{X}^k = \{\boldsymbol{X}_m^k\}|_{k=1}^N$ 表示 m 维协同特征的全部数值, $\boldsymbol{X}_{-m} \in \boldsymbol{R}^{d-m}$ 表示移出 m 维协同特征后的数据集, $\boldsymbol{X}_{(m)} \in \boldsymbol{R}^d$ 表示随机置换了 m 维协同特征后的新数据集, $\boldsymbol{X}_{(m)}^* \in R^m$ 表示随机置换了 m 维协同特征后的数据, w_c 表示数据的类别. 由统计学中相互独立的性质可知

$$p(\boldsymbol{X}_{(m)}) = p(\boldsymbol{X}_{(m)}^*, \boldsymbol{X}_{-m}) = p(\boldsymbol{X}_{(m)}^*)p(\boldsymbol{X}_{-m}). \quad (7)$$

置换 m 维协同特征后的数据集的概率分布为

$$p(\boldsymbol{X}_{(m)}, w_c) = p(\boldsymbol{X}_{(m)}^*, \boldsymbol{X}_{-m}, w_c) =$$

$$p(\boldsymbol{X}_{(m)}^*)p(\boldsymbol{X}_{-m}, w_c) = p(\boldsymbol{X}_m)p(\boldsymbol{X}_{-m}, w_c), \quad (8)$$

由置换后数据集属于 w_c 的条件概率可得

$$p(w_c|\boldsymbol{X}_{(m)}) = \frac{p(\boldsymbol{X}_{(m)}, w_c)}{p(\boldsymbol{X}_{(m)})} = \frac{p(\boldsymbol{X}_m)p(\boldsymbol{X}_{-m}, w_c)}{p(\boldsymbol{X}_m)p(\boldsymbol{X}_{-m})} =$$

$$p(w_c|\boldsymbol{X}_{-m}). \quad (9)$$

由随机置换定义可知, 置换前后数据集特征向量各维的期望和方差不变, 各维之间的相关系数在置换前后不变, 则由式 (9) 可知, 随机置换了 m 维协同特征后数据属性一致. □

正如文献 [16] 分析指出, 通过构造适当的量来度量各个特征之间的关系是特征选择算法的基础, 也为后续分类[17]和聚类等问题[18]提供了信息支撑. SFSKDE-MRP 算法基于协同特征的划分需要先验知识, 作为一种特征关系的度量是合理的. 其他自动化的确定方法仍然是值得进一步研究的开放性课题.

综上所述, 多维协同特征同时选择的算法如下.

算法 2 SFSKDE-MRP 算法.

输入: 带标签数据集 $\boldsymbol{S}$, $g = 1$;

输出: $\tilde{\Re}_s = \{\Re_s(1), \Re_s(2), \cdots, \Re_s(d)\}$.

1) 给定数据集 $\boldsymbol{S} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n) \in \boldsymbol{X}' \times \boldsymbol{Y} : 1 \leqslant n \leqslant N\}$. 其中: $\boldsymbol{X}' \in \boldsymbol{R}^d$ 表示样本的特征空间, $X^i = (x_1^i, x_2^i, \cdots, x_N^i)$ 表示数据 $\boldsymbol{X}'$ 上第 i 个特征上的数据, $\boldsymbol{Y}$ 表示样本的标签空间, (x_i, y_i) 表示数据集中的一个样本.

2) 根据先验知识得到 m 维协同特征 $\boldsymbol{X}_m \in \boldsymbol{R}^m$, 组成协同特征集合, 其余特征作为非协同特征集合.

3) 在区间 $(0, 1)$ 上均匀产生 $\varsigma_1, \varsigma_2, \cdots, \varsigma_{N-1}$, $m = \lfloor \varsigma \rfloor$ 表示向下取整, $m = \lfloor N * \varsigma_k \rfloor + 1$, $k = 1, 2, \cdots, N-1$, 交换 x_k^i 和 x_m^i, 将得到的随机序列 $X^i = (x_1^i, x_2^i, x_k^i, x_m^i, \cdots, x_N^i)$ 上的数据完全随机置换, 得到 $X^i = (x_1^i, x_m^i, x_k^i, \cdots, x_N^i)$, $i = 1, 2, \cdots, d$.

4) 按式 (6) 从协同特征集合选择特征进行置换, 其他特征保持不变, 计算每一次置换前后数据在再生核 Hilbert 空间中密度估计差 $\tilde{\Re}_S$, 返回 2). 令 $g \leftarrow g + 1$, 直到 $g = d$ 得到全部特征的 $\tilde{\Re}_S$ 数值.

5) 按 $\tilde{\Re}_S$ 大小进行特征排序, 其中协同特征相邻输出, 得到特征序列子集 $\Re_S$.

6) 按照 KNN 分类器的分类精度进行特征维数的选择.

2 实验分析

为了验证 FSKDE-RP 算法的可行性、有效性和可比性, 选用模拟数据集和 UCI 数据集进行实验, 数据信息如表 1 所示. 其中 Feature number、Sample number 和 Class number 分别表示数据的特征个数、实例个数和类别个数. 实验前先将各数据集归一化; 实验中将 FSKDE-RP 与相关的方法进行比较, 以 KNN 方法的分类精度作为测试特征选择方法的性能指标. 采用的对比算法有 Relief[3]、Relief-F[4]、mRMR[12] 和 mREL[12]. 对比算法的参数均根据具体应用通过交叉验证获得.

FSKDE-RP 算法采用高斯函数作为核函数, 窗宽参数 σ 值是通过网格搜索选取的, $\sigma \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. 实验环境为: WinXP, CPU Intel T4300 2.1 GHz, RAM 3 G, Matlab R2010a 等.

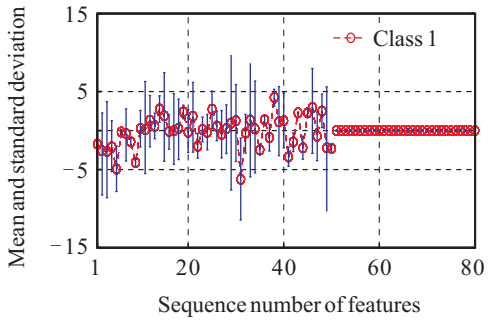
表 1 数据集

Data set	Feature number	Sample number	Class number
Heart	13	270	2
Waveform21	21	5 000	3
Wine	13	178	3
Breast-Cancer	10	699	2
Iris	4	150	3
Glass	9	214	7
Australian	14	690	2
Ionosphere	33	351	2
Liver-disorder	6	345	2
Pima-Diabetes	8	768	2
Haberman	3	306	2
Ecoli	7	336	8
Sonar	60	205	2

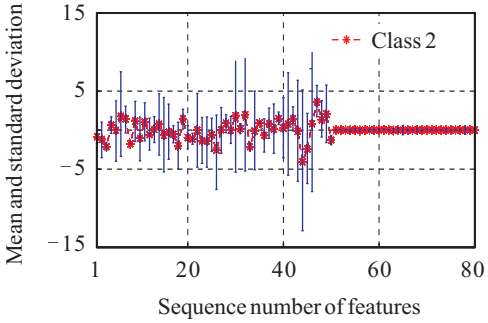
2.1 测试 FSKDE-RP 算法识别重要特征的能力 (可行性验证)

首先使用的数据集是参照文献 [19] 由高斯分布函数模拟生成的二分类数据集, 样本容量为 1 000, 数

据集共有特征 80 维, 每一维的期望和方差不同, 如图 1 所示. 其中前面的 50 维特征是有用的特征, 后面的 30 维是无用的特征. 这里无用特征是指该特征的期望和方差都很小, 对数据分类仅贡献很小的特征.



(a) 正类特征序列

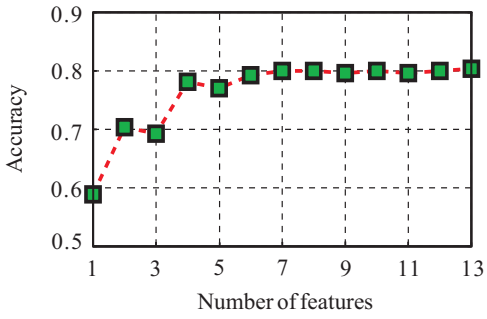


(b) 负类特征序列

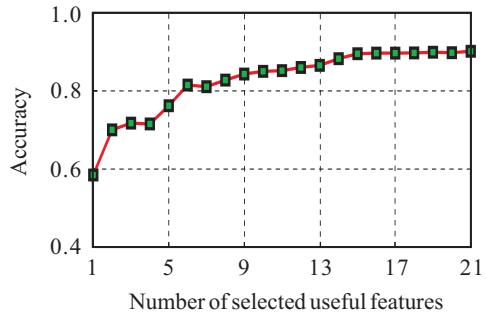
图 1　模拟数据的特征序列

为了说明 FSKDE-RP 算法的可行性, 本文定义了一种衡量算法效率的指标, 称之为排序正确率, 即 $\gamma = n_t/N$. 其中: n_t 表示算法找出的有用特征的维数排在前 50 位的个数, N 表示数据集有用特征的维数. 利用 FSKDE-RP 算法得到两类的排序正确率分别为 $\gamma_1 = 48/50 = 96\%$ 和 $\gamma_2 = 100\%$, 说明模拟数据集 FSKDE-RP 算法找出了数据集有用的特征, 且准确率较高. 针对 UCI 数据集, FSKDE-RP 算法的输出是按特征的重要性排列的一个有序序列, 所以, 采用如下方式评价其性能. 首先, 根据特征在 FSKDE-RP 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 KNN 分类器的输入. 在评价分类器的性能时, 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 2 给出了采用不同个数的特征来表征数据时分类器的分类准确性变化趋势.

从图 2 的曲线变化趋势可以看出: 对于数据集 Heart 和 Waveform21 经过 FSKDE-RP 方法特征提取之后, KNN 分类器可以在数据的少量特征时得到与在全部特征时几乎相同的分类准确性; 随着特征个数的增加, 分类精度逐渐升高, 之后保持不变甚至略微下降. 这是由于越来越多的冗余的或者不相关的特征被包含进来, 它们不能给学习算法提供有益信息甚至误导学习算法所致. 实验结果表明: 通过 FSKDE-RP 方法得到的特征子集确实可以将具有代表性和携带重要信息的特征提取出来, 并按其重要程度放在了特征子集序列的前面; FSKDE-RP 方法用于特征提取是可行的, 可以作为特征的预筛选器使用.



(a) Heart



(b) Waveform 21

图 2　分类性能随特征个数的变化趋势

2.2　测试 FSKDE-RP 算法所获得的新特征空间数据集分类精度 (有效性验证)

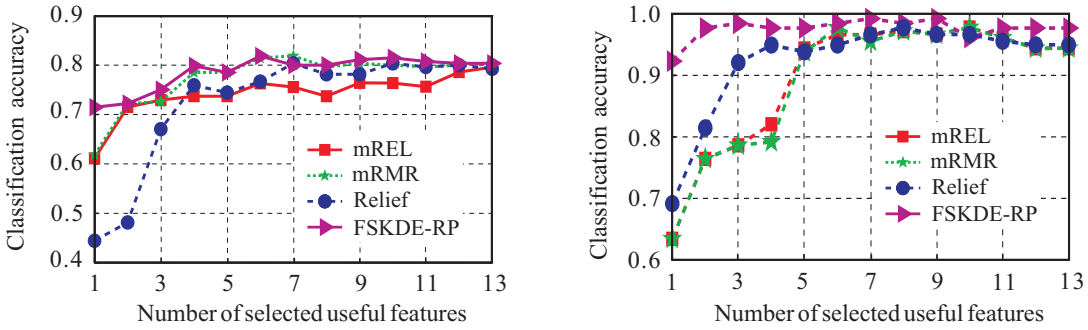
为了测试 FSKDE-RP 算法的有效性, 实验中选用 UCI 数据库中包含典型数值型特征的 Glass、Sonar、Australian 和 ionoshere 等数据集从两个角度进行实验. 一是对于 KNN 分类方法而言, K 的取值对分类性能有较为显著的影响, 因此, K-最近邻方法中 K 值的选择方法为, 通过用不同的 K 值 ($1 \leqslant K \leqslant \sqrt{N_{\rm tr}}$, $N_{\rm tr}$ 是训练样本集中样本的个数) 进行多次实验, 最终选定使分类器具有最好性能的 K 值. 表 2 给出了 FSKDE-RP 算法在各数据集全部特征和最优特征子集上的分类精度、核窗宽和近邻数 K 值. 二是进一步选择 Relief、mRMR 和 mREL 等方法进行性能的比较. 图 3 给出了不同方法的分类精度随特征个数的变化.

从图 3 和表 2 的实验结果可以给出如下观察.

1) 从图 3 可以看出, 随着特征维数的增加, 各种方法的分类性能都有所提高: mREL、mRMR 和 Relief 方法对于不同的数据集分类精度变化较大; FSKDE-RP 方法的分类精度优于或近似于其他 3 种方法, 且稳定性较高.

表 2 数据集分类精度

Datasets	Complete features	Optimal features	Complete accuracy	Optimal accuracy	Width	K
Heart	13	6	80.37±2.102 5	81.85±0.051 4	1	5
Breast	9	7	92.70±1.051 7	97.14±0.004 0	0.1	5
Iris	4	2	94.67±0.018 9	97.33±0.000 2	1	2
Glass	9	5	75.95±0.098 8	82.16±0.075 4	10	6
Australian	14	4	84.92±0.026 8	86.38±0.021 4	0.1	2
Ionoshere	33	7	84.35±0.060 2	90.33±0.036 4	0.1	2



(a) Heart dataset (b) Wine dataset

图 3 不同方法的分类精度随特征个数的变化

2) 从表 2 可以看出, 用 FSKDE-RP 算法得到的特征子集的数据作为 KNN 方法的输入, 通过 10 重交叉验证得到各数据集的分类精度都比原始全部特征数据的分类性能高, 从而说明各数据集都存在冗余或者不相关的特征, 利用 FSKDE-RP 算法可以准确地从原始特征集里挑出对分类有益的重要特征子集.

2.3 测试 FSKDE-RP 算法与其他算法选出特征排序的一致性 (可比性验证)

为了检验 FSKDE-RP 算法与其他算法选出特征排序的一致性, 将 FSKDE-RP 输出的特征序列与 Relief 和 Relief-F 方法所得到的特征序列进行比较. 数据集部分信息如表 3 所示.

表 3 Relief、Relief-F 和 FSKDE-RP 所生成的特征序列

Datasets	Relief	Relief-F	FSKDE-RP
Liver-disorder	3, 6, 4, 1, 2, 5	3, 6, 2, 4, 1, 5	3, 6, 5, 2, 4, 1
Pima	2, 1, 6, 8, 4, 3, · · ·	4, 6, 2, 5, 1, 8, · · ·	7, 6, 2, 4, 5, 1, · · ·
Wine	13, 1, 5, 6, 10, 12, · · ·	13, 5, 6, 12, 1, · · ·	5, 13, 1, 6, 12, 9, · · ·
Iris	4, 3, 1, 2	4, 3, 1, 2	4, 3, 2, 1
Haberman	1, 3, 2	1, 3, 2	3, 1, 2
Ecoli	6, 7, 1, 2, 5, 3, 4	6, 1, 7, 2, 5, 3, 4	7, 6, 1, 2, 5, 3, 4

从表 3 可以看出, 通过 Relief 和 Relief-F 两种方法提取的重要特征与 FSKDE-RP 方法得到的重要特征以及特征序列大致相同. Relief 和 Relief-F 方法确定的重要特征, 通常都会在 FSKDE-RP 方法输出的特征序列的前端, 进一步说明 FSKDE-RP 方法与其他算法具有较好的可比性.

2.4 协同特征实验分析 —— 随机置换概率密度多维协同特征选择算法 (SFSKDE-MRP)

SFSKDE-MRP 算法仍采用高斯函数作为核函数, 其窗宽参数通过网格搜索选取. 该实验从协同特征同时被选择的角度验证本文算法的有效性, 实验数据集的部分信息见表 4. 由定义 4, 根据先验知识定义 Iris 数据集中的花萼长和宽, 将花蕊长和宽作为协同特征; Heart 数据集中的 ST depression 和 ST slope 作为协同特征; Water treatment 数据集中的 SS-E 和 SSV-E 以及 SS-P 和 SSV-P 作为协同特征; Wine-quality 数据集中的固态酸 (fixed acidity) 和挥发酸 (volatile acidity) 以及游离二氧化硫 (free sulfur dioxide) 和总二氧化硫 (total sulfur dioxide) 作为协同特征.

利用 SFSKDE-MRP 算法对上述 4 个数据集进行

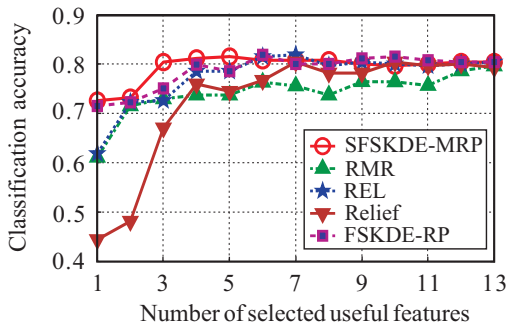
实验, 在数据集中随机抽取 70% 用于训练, 剩余的 30% 进行测试. 分别从运行时间和分类精度上与已有方法进行比较, 以验证 SFSKDE-MRP 算法的有效性. 实验结果见表 5 和图 4.

表 4　数据集以及协同信息

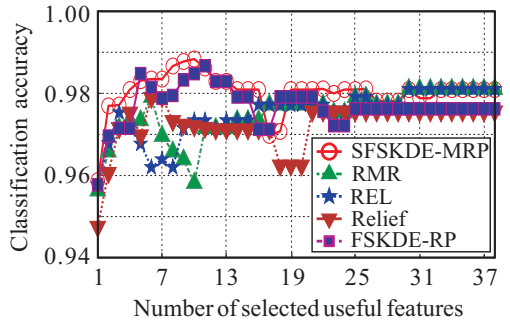
Datasets	Feature number	Sample number	Class number	Synergetic features
Iris	4	150	3	sepal length, sepal width,petal length ,petal width
Heart	13	270	2	ST depression, ST slope
Water treatment	38	527	2	SS-E, SSV-E,SS-P,SSV-P
Wine Quality	12	6 497	2	fixed acidity , volatile acidity, free sulfur dioxide , total sulfur dioxide

表 5　各种方法的运行时间　　s

Datasets	mRMR	mREl	Relief	FSKDE-PR	SFSKDE-MPR
Iris	5.40	4.90	0.78	4.06	0.55
Heart	135.53	22.37	22.24	39.69	15.33
Water treatment	324.65	298.51	152.72	188.26	62.37
Wine Quality	828.27	729.64	635.98	367.89	228.76



(a)　Heart



(b)　Water treatment

图 4　不同方法的分类准确率随特征个数的变化趋势

从表 5 和图 4 描述的实验结果给出如下观察.

1) 由表 5 可以看出: 5 种方法的运行时间都随样本容量增大而增加; mRMR 和 mREL 算法所需运行时间较大; 对于样本容量较小的 Iris、Heart 和 Water treatment 数据集而言, Relief 算法的运行时间优于 FSKDE-PR 算法, 但相比较而言, 由于 SFSKDE-MPR 算法本身参数较少且兼顾了协同特征、多维协同特征同时选择, SFSKDE-MPR 算法的运行时间优于其他 4 种方法.

2) 从图 4 可以看出: 随着特征维数的增加, 各种方法的分类精度都呈现先上升后下降的趋势, 其中 mREL、mRMR 和 Relief 方法随着特征数目增加波动明显; 相比于其他 3 种方法, SFSKDE-MRP 和 FSKDE-RP 算法的分类精度较高且稳定性较好; 在分类精度上, SFSKDE-MRP 算法略高于 FSKDE-RP 算法, 从而也说明针对协同特征同时出现的问题, 通过先验知识找出协同特征后, 再利用本文方法可以更有效地找出主要特征且分类准确率也有所提高.

3　结　论

针对已有特征选择算法依赖特征加权技术以及优化过程时间复杂度高的问题, 基于 ISE 的核密度估计和随机置换, 本文提出了单一或协同特征的选择方法. 相对于已有的方法, 本文提出的基于平方误差标准 (ISE) 和核密度估计的 FSKDE-RP 方法可以在再生核 Hilbert 空间 (RKHS) 中实现完全核化, 具有参数少、便于理解和易于实现等特点, 通过拓展随机置换理论使之适合协同特征的情形; 同时提出了多维协同特征选择方法 SFSKDE-MPR, 更加符合实际的应用问题. 实验分析验证了所提出的方法具有良好的适应性. 虽然本文方法展现出了较好的有效性, 但其依然面临一些需要进一步探讨的问题. 例如将本文算法应用到高维和复杂的图像或视频中, 并进一步设计更好的自动识别协同特征的方法等问题, 将是一个非常有意义的工作.

参考文献(References)

[1] Langley P. Selection of relevant features in machine learning[C]. Proc of the AAAI Fall Symposium on

Relevance. Menlo Park: AAAI, 1994: 1-5.

[2] 姜慧研, 柴天佑. 基于可信间隔的特征选择方法研究[J]. 控制与决策, 2011, 26(8): 1229-1232.
(Jiang H Y, Chai T Y. Method for feature selection based on authentic distance[J]. Control and Decision, 2011, 26(8): 1229-1232.)

[3] Kira K, Rendell L A. A practical approach to feature selection[C]. Proc of the 9th Int Workshop on Machine Leaning. San Francisco: Morgan Kaufmann, 1992: 249-256.

[4] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF[C]. Proc of ECML. New York: Springer-Verlag, 1994: 171-182.

[5] Hsu W H. Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning[J]. Information Sciences, 2004, 163(17): 103-122.

[6] Huang J, Cai Y, Xu X. A hybrid genetic algorithm for feature selection wrapper based on mutual information[J]. Pattern Recognition Letters, 2007, 28(13): 1825-1844.

[7] Krzysztof M, Halina K. Correlation-based feature selection strategy in classification problems[J]. Int J of Application Mathematics Computer Science, 2006, 16(4): 503-511.

[8] Song L, Smola A, Gretton A, et al. Supervised feature selection via dependence estimation[C]. Prco of the 24th Int Conf on Machine Learning. Corvallis: ACM, 2007: 823-830.

[9] Jain A K, Robert P W, Mao J C. Statistical pattern recognition: A review[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-37.

[10] Geng X L, Hu G H. Unsupervised feature selection by kernel density estimation in wavelet-based spike sorting[J]. Biomedical Signal Processing and Control, 2012, 7(2): 112-117.

[11] He X F, Cai D, Niyogi P. Laplacian scores for feature selection[J]. The Neural Information Processing Systems. 2005, 18(9): 507-514.

[12] 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.

[13] Kim J, Scott C D. L_2 kernel classification[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(10): 1822-1831.

[14] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[15] Yang J B, Shen K Q, Ong C J, et al. Feature selection for MLP neural network: The use of random permutation of probabilistic outputs[J]. IEEE Trans on Neural Networks, 2009, 20(12): 1911-1922.

[16] Hu W J, Choi K S, Gu Y G, et al. Minimum-maximum local structure information for feature selection[J]. Pattern Recognition Letters, 2013, 34(5): 527-535.

[17] 顾鑫, 王士同, 许敏. 领域自适应的最小包含球设计方法[J]. 控制与决策, 2013, 28 (2): 177-182.
(Gu X, Wang S T, Xu M. Minimum enclosing ball for domain adaptation[J]. Control and Decision, 2013, 28(2): 177-182.)

[18] 王骏, 王士同, 邓赵红. 聚类分析研究中的若干问题[J]. 控制与决策, 2012, 27(3): 321-328.
(Wang J, Wang S T, Deng Z H. Survey on challenges in clustering analysis research[J]. Control and Decision, 2012, 27(3): 321-328.)

[19] Deng Z H, Chung F L, Wang S T. Robust relief-feature weighing, margin maximization, and fuzzy optimization[J]. IEEE Trans on Fuzzy Systems, 2010, 18(4): 726-744.

(责任编辑: 李君玲)