DOI:10.11992/tis.201609008
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.tp.20170407.1758.016.html

基于特征相关的谱特征选择算法

胡敏杰,林耀进,杨红和,郑荔平,傅为
(闽南师范大学 计算机学院,福建 漳州 363000)

摘 要:针对传统的谱特征选择算法只考虑单特征的重要性,将特征之间的统计相关性引入到传统谱分析中,构造了基于特征相关的谱特征选择模型。首先利用 Laplacian Score 找出最核心的一个特征作为已选特征,然后设计了新的特征组区分能力目标函数,采用前向贪心搜索策略依次评价候选特征,并选中使目标函数最小的候选特征加入到已选特征。该算法不仅考虑了特征重要性,而且充分考虑了特征之间的关联性,最后在 2 个不同分类器和 8 个 UCI 数据集上的实验结果表明:该算法不仅提高了特征子集的分类性能,而且获得较高的分类精度下所需特征子集的数量较少。

关键词:特征选择;谱特征选择;谱图理论;特征关联;区分能力;索搜策略;拉普拉斯;分类精度
中图分类号:TP18 **文献标志码**:A **文章编号**:1673-4785(2017)04-0519-07

中文引用格式:胡敏杰,林耀进,杨红和,等.基于特征相关的谱特征选择算法[J]. 智能系统学报, 2017, 12(4): 519-525.
英文引用格式:HU Minjie, LIN Yaojin, YANG Honghe, et al. Spectral feature selection based on feature correlation[J]. CAAI transactions on intelligent systems, 2017, 12(4): 519-525.

Spectral feature selection based on feature correlation

HU Minjie, LIN Yaojin, YANG Honghe, ZHENG Liping, FU Wei
(School of Computer Science, Minnan Normal University, Zhangzhou 363000, China)

Abstract: In the traditional spectrum feature selection algorithm, only the importance of single features are considered. In this paper, we introduce the statistical correlation between features into traditional spectrum analysis and construct a spectral feature selection model based on feature correlation. First, the proposed model utilizes the Laplacian Score to identify the most central feature as the selected feature, then designs a new feature group discernibility objective function, and applies the forward greedy search strategy to sequentially evaluate the candidate features. Then, the candidate feature with the minimum objective function is added to the selected features. The algorithm considers both the importance of feature as well as the correlations between features. We conducted experiments on two different classifiers and eight UCI datasets, the results of which show that the algorithm effectively improves the classification performance of the feature subset and also obtains a small number of feature subsets with high classification precision.

Keywords: feature selection; spectral feature selection; spectral graph theory; feature relevance; discernibility; search strategy; Laplacian score; classification performance

特征选择是指在原始特征空间中选择能让给定任务的评价准则达到最优的特征子集的过程,是模式识别、机器学习等领域中数据预处理的关键步骤之一[1-3]。其主要目标是在不显著降低分类精度和不显著改变类分布情况下选择一个重要特征子集并且移除不相关或不重要的特征,使留下的特征具有更强的分辨率[4]。其中评价准则是特征选择算法中的关键步骤,国内外研究者已设计了多种评价准则,包括距离度量[5]、信息度量[6]和谱图理论[7-8]等方法。由于基于谱图理论的特征选择模型的可理解性及其完备的数学理论,受到了广泛的关注[8-9]。

收稿日期:2016-09-08. **网络出版日期**:2017-04-07.
基金项目:国家自然科学基金项目(61303131,61379021);福建省教育厅科技项目(JA14192).
通信作者:胡敏杰.E-mail:zzhuminjie@sina.com.

根据数据是否带有标记,基于谱图理论的特征选择可分为有监督特征选择和无监督特征选择[8-12]。无监督特征选择算法在构造相似性矩阵时不考虑类信息,通常对给出的样本值采用核函数构造相似性矩阵。有监督特征选择算法将类信息引入相似性矩阵中,常根据类别个数来构造对应的相似性矩阵。利用谱图理论进行特征选择的主要思想是对邻接图 Laplacian 矩阵进行谱分解,其特征向量反映了样本的类属关系[11]。基于该思想,Zhao 等[8]设计了一个谱特征选择框架,框架根据相似性矩阵是否考虑类标记信息分别应用于有监督和无监督算法,而选择特征子集过程与具体学习器无关,利用特征对样本分布的影响对特征进行排序。He 等[10]结合谱图理论和特征的局部保持能力提出了基于 Laplacian 得分的特征选择算法。Zhao[8] 和 He[10] 等基于谱图理论的特征选择均仅考虑每个单独的特征按一定的可分性或统计判据进行排队以形成特征序列,并取靠前的特征子集进行学习。该策略仅在各个特征间统计独立且类别正态分布时较优,但特征间具有这种关系仅仅是极少数[13],实际上特征空间中特征之间存在较为紧密的关联性。

针对已有的基于谱图理论有监督特征选择算法存在的上述问题,我们提出融合特征相关的谱特征选择算法,在原始的整个特征空间中不仅考虑每一个特征的区分力,还考虑特征组的区分性能,迭代地寻找对保持数据的局部结构比上一组特征更强的特征组合。由此,提出了基于特征相关的谱特征选择算法(spectral feature selection based on feature correlation,SPFC)。实验结果表明,该算法不仅提高了特征选择的分类性能,而且获得了较高的分类精度下所需特征子集的数量较少。

1 谱特征选择算法

谱特征选择算法的主要理论是谱图理论,本文研究的算法是以 Laplacian Score 特征选择算法为基础,因此本节只介绍图 Laplacian 矩阵谱分析。

假设训练样本集 $\boldsymbol{X}=[\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_m]^{\mathrm{T}}$,类标记 $\boldsymbol{y}=[\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_m]^{\mathrm{T}}$。用标量 $F^i(1\leqslant i\leqslant n)$ 记为第 i 个特征,向量 $\boldsymbol{f}^i$ 表示所有样本在第 i 个特征上的取值,第 j 个样本可以表示成 $\boldsymbol{x}_j=(f_j^1,f_j^2,\cdots,f_j^n)$。样本分布的结构由邻接图 $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$ 表示,其中 $\boldsymbol{V}$ 为图的点集,图的第 j 个结点 $v_j\in\boldsymbol{V}$ 对应于训练样本中的第 j 个样本 $\boldsymbol{x}_j$,$\boldsymbol{E}$ 为图的边集,边 $e_{ji}\in\boldsymbol{E}$ 的权重 W_{ij} 对应第 j 个样本和第 i 个样本的相似度 $S_{ij}(1\leqslant i,j\leqslant m)$。本文研究有监督谱特征选择,即构建相似性矩阵用到类标记信息,那么对于给定的图 $\boldsymbol{G}$,其相似性矩阵 $\boldsymbol{S}$ 为

$$\boldsymbol{S}_{ij}=\begin{cases}\dfrac{1}{n_k}, & y_i=y_j=k\\ 0, & \text{其他}\end{cases}$$

式中 n_k 为类别为 k 的样本个数。

令 $\boldsymbol{G}$ 为一无向有权图,则邻接矩阵 $\boldsymbol{W}_{ij}=\boldsymbol{S}_{ij}(1\leqslant i,j\leqslant m)$,且 $\boldsymbol{W}$ 为对称矩阵。令度矩阵 $\boldsymbol{D}$ 为

$$\boldsymbol{D}=\begin{cases}\sum_{j=1}^{m}\boldsymbol{W}_{ji}, & j=i\\ 0, & j\neq i\end{cases} \tag{1}$$

由式(1)可以看出度矩阵 $\boldsymbol{D}$ 是一个对角矩阵,对角线上的每个元素是邻接矩阵 $\boldsymbol{W}$ 每一行或每一列的和。度矩阵可以解释为每个样本周围围绕的其他样本的密集度,度矩阵中的元素越大,意味着有更多的样本靠近这个元素代表的样本。

由邻接矩阵和度矩阵得到相应的 Laplacian 矩阵 $\boldsymbol{L}$ 和正则化的 Laplacian 矩阵 $\mathscr{L}$

$$\mathscr{L}=\boldsymbol{D}-\boldsymbol{S},\mathscr{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}$$

根据 Laplacian 矩阵的性质[5],给出下面定义:

定义 1 Laplacian 矩阵的最小特征值为 0,对应特征向量为单位向量

$$\text{let } \boldsymbol{I}=[1\ 1\ \cdots\ 1]^{\mathrm{T}},\boldsymbol{L}*\boldsymbol{I}=0$$

定义 2 对于任意一个 n 维向量 $\boldsymbol{x}$(数据集中的特征列),都满足下式成立:

$$\forall\boldsymbol{x}\in\boldsymbol{R}^n,\boldsymbol{x}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{x}=\frac{1}{2}\sum\ \boldsymbol{w}_{ij}\ (\boldsymbol{x}_i-\boldsymbol{x}_j)^2$$

定义 3 对于任意一个 n 维向量 $\boldsymbol{x}$(数据集中的特征列),任意一个实数,都有(特征列中的每个元素减去一个相同的值得到的新特征列仍然保持结果不变):

$$\forall\boldsymbol{x}\in\boldsymbol{R}^n,\forall t\in R,(\boldsymbol{x}-\boldsymbol{t}*\boldsymbol{e})^{\mathrm{T}}L(\boldsymbol{x}-\boldsymbol{t}*\boldsymbol{e})=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{x}$$

谱图理论说明,Laplacian 矩阵的特征值与特征向量包含着数据集的样本分布结构。谱特征选择在选取有强识别度的特征时,以特征取值的分布是否与样本分布的结构保持一致作为特征选择的评价标准。例如在图 1 中,每个图形(三角和星)表示一个样本,形状不同意味样本在同一特征上的取值不同,各圆形分别为类 1 和类 2 的区域,即同一区域内的样本属于同一类别。在图 1 左侧中属于同一类的样本在特征 F^1 上取值近似相同,而不属于同一类的样本在特征 F^1 上取值不同,因此特征 F^1 对类 1 和类 2 就有很好的识别能力,此时称特征 F^1 的取值分布与样本分布一致。在图 1 右侧中特征 F^2 的取值分布则与样本分布不一致,F^2 对类 1 和类 2 不具

有很好的识别能力。因此在谱特征选择算法中会选取 F^1 而不选 F^2。

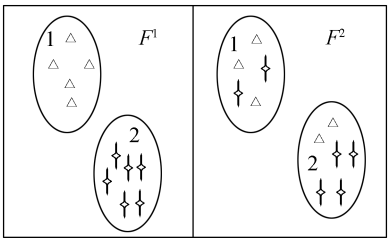


图 1　特征与样本分布一致性示意图

Fig.1　The characteristics and the sample distribution consistency schematic diagram

而谱特征选择算法将选择那些与样本分布结构一致的特征,即选择那些使得式(2)取较小值的特征[7]:

$$\varphi(F^r)=\frac{\sum_{i,j}S_{ij}(f_{ri}-f_{rj})^2}{V(\boldsymbol{f}_r)} \tag{2}$$

式中:F^r 表示第 r 个特征,$\boldsymbol{f}_r$ 是第 r 个特征向量,f_{ri} 和 f_{rj} 表示第 r 个特征上的 $i,j(1\leqslant i,j\leqslant m)$ 个样本的取值,$V(\boldsymbol{f}_r)$ 表示第 r 个特征 $\boldsymbol{f}_r$ 的方差。

一个随机变量 x 的方差定义为[7]:

$$V(x)=\int_M (x-u)^2\mathrm{d}P(x)$$

式中:M 是数据的流行结构,u 表示 x 期望值,$\mathrm{d}P$ 是一个概率度量。根据谱图理论[7],$\mathrm{d}P$ 可以用对角矩阵 $\boldsymbol{D}$ 估计出来,因此特征 $\boldsymbol{f}_r$ 的方差 $V(\boldsymbol{f}_r)$ 为

$$V(\boldsymbol{f}_r)=\sum_i (f_{ri}-u_r)^2D_{ii}$$

式中 u_r 表示第 r 个特征 $\boldsymbol{f}_r$ 的期望值,定义为

$$u_r=\sum_i\left(f_{ri}\frac{D_{ii}}{\sum_i D_{ii}}\right)=\frac{\sum_i f_{ri}D_{ii}}{\sum_i D_{ii}}$$

$\sum_{i,j}S_{ij}(f_{ri}-f_{rj})^2$ 越小表示在样本分布结构图里近邻的样本在该特征上差异很小,即一个识别能力强的特征会使得 S_{ij} 大而 $(f_{ri}-f_{rj})$ 小,因而式(2)趋小。$\sum_i (f_{ri}-u_r)^2D_{ii}$ 越大表示该特征在各样本上的取值方差越大,一个区分能力强的特征应该会赋予同类样本近似的值而不同类样本差异大的值,即具有较大方差的特征具有较强的识别能力。因此式(2)通过谱图理论结合特征的局部信息保持能力和方差来进行特征选择。

2　基于特征相关的谱特征选择模型

传统的谱特征选择算法采用单独最优特征组合的启发式搜索策略,用式(2)对每个特征单独度量其重要度,该策略并未考虑特征间的冗余度和交互性,因此需要考虑候选特征与已选特征之间的冗余性和交互性。本文在式(2)的基础上定义了特征组的重要度公式如式(3)。为了度量每个候选特征对已选特征的贡献程度,同时定义了式(4)来计算候选特征的重要度。模型思想是:首先利用传统谱特征选择算法选出使目标函数式(2)最好的一个特征,然后以这个特征为核心据点作为已选特征,依次逐个评价候选特征与已选特征的相关性,即依次根据目标函数(式(3))评价特征组合后的图的保持能力,然后根据式(4)选出保持能力优于未组合时的最强一个特征,并将该特征加入到已选特征中形成新的组合,接着对余下候选特征进行下一轮的迭代。该算法不仅考虑了特征间的相关效应,而且通过式(4)避免了特征间的冗余。

定义特征组相关的目标函数为

$$\varphi(\boldsymbol{F}_S)=\frac{\mathrm{sqrt}\left(\sum_{r=1}^{k}\sum_{i,j}(S_{ij}(f_{ri}-f_{rj})^2\right)}{\frac{\sum_{r=1}^{k}\sum_i(f_{ri}-u_r)^2D_{ii}}{k}} \tag{3}$$

式中:$\boldsymbol{F}_S$ 表示已选出的特征子集,k 表示已选出的特征子集中的特征个数,$\varphi(\boldsymbol{F}_s)$ 表示已选出的特征子集对数据的识别能力。$\mathrm{sqrt}(\sum_{r=1}^{k}\sum_{i,j}S_{ij}(f_{ri}-f_{rj})^2)$ 评估已选的特征子集对样本空间的局部保持能力,通过欧式距离计算各特征间的相关性,特征子集对样本局部保持能力越强,则 $\mathrm{sqrt}(\sum_{r=1}^{k}\sum_{i,j}S_{ij}(f_{ri}-f_{rj})^2)$ 越小。$\frac{\sum_{r=1}^{k}\sum_i(f_{ri}-u_r)^2D_{ii}}{k}$ 通过各特征的均方差来衡量各特征对样本数据的区分能力,已选特征子集区分能力越强,则 $\frac{\sum_{r=1}^{k}\sum_i(f_{ri}-u_r)^2D_{ii}}{k}$ 越大。因此式(3)越小则说明已选子集能力越强。

在式(3)的基础上通过式(4)评估候选特征中能提升已有特征子集的区分能力的特征,其目标函数定义为

$$\mathrm{argminRed}(\boldsymbol{f}_i)_{\boldsymbol{f}_i\in\boldsymbol{F}_U}=\varphi(\boldsymbol{F}_S\cup\boldsymbol{F}_i)-\varphi(\boldsymbol{F}_S) \tag{4}$$

式中:$\boldsymbol{F}_U$ 表示候选特征集合,$\boldsymbol{f}_i\in\boldsymbol{F}_U$,通过评估一个新的特征 $\boldsymbol{f}_i$ 能否使得同类样本距离小而不同类样本距离大来衡量是否加入已选 $\boldsymbol{F}_s$。又在候选特征集合里可能有多个 $\boldsymbol{f}_i$ 能提升已选子集的能力,由式(3)知新加入的 $\boldsymbol{f}_i$ 使得 $\varphi(\boldsymbol{F}_S)$ 越小越好,因此在多个具有提升子集能力的候选特征中选择使

$\varphi(\boldsymbol{F}_S \cup \boldsymbol{f}_i) - \varphi(\boldsymbol{F}_S)$ 最小的一个特征。

根据式(4),可提出基于特征相关的谱特征选择算法(SPFC)的伪代码如算法 1 所示。

算法 1 基于特征相关的谱特征选择算法(SPFC)

输入 样本集 $\boldsymbol{X}$,类标记 $\boldsymbol{Y}$;

输出 $\boldsymbol{F}_S$ 特征相关后的特征序列。

1) $\boldsymbol{F}_S = \varnothing$, $\boldsymbol{F}_U = [F^1\ F^2 \cdots\ F^n]$;

2) 依据 X、Y 计算每两个样本间的相似度矩阵 $\boldsymbol{S}_{ij}(1 \leqslant i,j \leqslant m)$;

3) 依据相似度构建 Laplacian 图 $\boldsymbol{G}$,并计算 $\boldsymbol{W}$、$\boldsymbol{D}$、$\boldsymbol{L}$;

4) 根据传统谱特征选择算法求出最具有识别力的一个特征 $\boldsymbol{f}_i$

$\boldsymbol{F}_S = \boldsymbol{F}_S \cup \boldsymbol{f}_i$, $\boldsymbol{F}_U = \boldsymbol{F}_U - \{\boldsymbol{f}_i\}$;

5) while $\boldsymbol{F}_U$ 不为空

6) 根据式(3)计算 $\varphi(\boldsymbol{F}_S)$;

7) for $i = 1$ to length($\boldsymbol{F}_U$)

if $(\varphi(\boldsymbol{F}_S \cup \boldsymbol{f}_i) - \varphi(\boldsymbol{F}_S)) > 0$ then

$L(j) = \varphi(\boldsymbol{F}_S \cup \boldsymbol{f}_i) - \varphi(\boldsymbol{F}_S)$;

$j = j + 1$;

end if

8) end for

9) 将 L 按升序排序

$\boldsymbol{F}_S = \boldsymbol{F}_S \cup \boldsymbol{f}_{L(1)}$, $\boldsymbol{F}_U = \boldsymbol{F}_U - \{\boldsymbol{f}_{L(1)}\}$;

10) end while

3 实验设计与对比

3.1 实验数据

为了进一步验证 SPFC 算法的有效性,本文从 UCI (http://www.ics.uci.edu)中选择 8 个数据集,各数据集相应的描述信息见表 1,在表 1~3 中 australian _ credit 数据集简写为 AC,VeteranLungCancer 数据集简写为 VE。

表 1 实验数据描述

Table 1 Experimental data description

数据集	样本数	特征数	类别数
AC	690	14	2
crx	690	15	2
heart	270	13	2
ICU	200	20	3
rice	104	5	2
Ve	137	7	2
wpbc	198	33	2
zoo	101	17	7

3.2 实验结果与分析

为了验证 SPFC 算法的性能,实验分为两部分。第 1 组实验与 CFS[14]、ChiSquare[15]、FCBF[16]、Laplacian[10]、NRS[17] 以及 Relief[18] 算法进行比较由特征子集诱导出来的分类精度。另一组实验采用 Friedman test[19] 和 Bonferroni-Dunn test[20] 在统计上对比 SPFC 与其他算法在 8 个数据集上的实验结果。由于 ChiSquare、Laplacian、FCFS、Relief 这 4 种算法得到的是一个特征序列,而 CFS、FCBF、NRS 3 种算法得到的是子集约简,因此,ChiSquare、Fisher、FCFS、Relief 这 4 种算法得到的特征序列取前 k 个特征作特征子集,其中 k 为 CFS、FCBF、NRS 算法中得到的 3 个约简子集数量的最小值。此外,NRS 算法中的邻域参数值 δ 为 0.10。在实验中,采用十折交叉验证法进行评价特征子集的优劣,用 KNN(K=10)、CART 2 个不同的基分类器来评价分类精度。

实验 1 为了比较特征选择后的分类精度,在表 2~4 中,分别采用 KNN(K=10)、CART 这 2 个不同的基分类器进行特征子集分类精度的评价。此外,为了更加直观地比较不同方法得到的特征子集的性能,表 2、3 中加粗的数值表示最高分类进度,下划线表示精度次优,最后一行表示不同算法得到的特征子集的平均分类精度,最后一行中加下划线的数值表示平均分类精度最高的值。另外,括号里面的数值表示数据的标准差,括号外面的数值表示分类精度。

表 2 不同特征选择算法在 KNN 分类器下的分类精度比较

Table 2 Classification accuracies of feature selection with different algorithms based on KNN %

数据集	SPFC	Laplacian	CFS	ChiSquare	FCBF	NRS	Relief
AC	86.1(3.0)	84.7(4.4)	84.7(4.4)	86.6(3.1)	84.7(4.4)	85.2(4.8)	84.7(4.1)
crx	83.8(1.7)	84.0(1.6)	83.4(1.7)	83.1(1.7)	83.4(1.7)	84.3(1.7)	**84.4(1.8)**
heart	82.6(6.4)	**84.0(4.2)**	82.5(5.8)	84.0(4.3)	82.5(3.9)	80.0(8.7)	75.9(7.5)
ICU	**92.1(2.3)**	91.5(3.3)	91.5(3.3)	91.5(3.3)	91.5(3.3)	89.9(6.1)	81.3(3.0)
rice	**82.7(10.0)**	76.9(11.0)	76.9(11.0)	76.9(11.0)	76.9(11.0)	81.6(10.3)	78.9(11.4)
Ve	**75.8(12.4)**	**75.8(12.4)**	**75.8(12.4)**	**75.8(12.4)**	**75.8(12.4)**	72.1(3.7)	70.7(1.6)
wpbc	**74.6(9.2)**	**74.6(9.2)**	70.5(11.11)	70.5(11.11)	70.5(11.11)	73.6(9.3)	67.4(13.6)
zoo	**91.4(7.5)**	**91.4(7.5)**	90.5(10.5)	86.1(8.7)	89.6(8.5)	74.3(10.0)	87.2(10.0)
平均值	**83.64**	82.86	81.98	81.81	81.86	80.13	78.81

表 3　不同特征选择算法在 CART 分类器下的分类精度比较

Table 3　Classification accuracies of feature selection with different algorithms based on CART　%

数据集	SPFC	Laplacian	CFS	ChiSquare	FCBF	NRS	Relief
AC	**84.2**(**4.5**)	83.6(3.3)	82.6(7.3)	81.1(4.5)	82.4(7.5)	83.4(4.7)	80.5(4.0)
crx	**83.9**(**16.9**)	82.0(13.4)	80.1(14.4)	79.5(13.6)	79.9(14)	82.4(15.4)	80.1(12.8)
heart	**81.1**(**7.9**)	76.3(6.8)	77.0(6.9)	76.6(7.4)	77.7(7.4)	75.1(9.2)	77.7(7.8)
ICU	92.1(2.3)	90.5(7.9)	90.5(7.9)	90.5(7.9)	90.5(7.9)	85.3(17.5)	**92.6**(**2.3**)
rice	**83.9**(**9.4**)	**83.9**(**9.4**)	83.0(10.0)	83.0(10.0)	83.0(10.0)	82.0(11.7)	80.8(7.5)
Ve	70.6(10.7)	70.6(10.7)	70.6(10.7)	70.6(10.7)	70.6(10.7)	68.7(11.5)	**70.7**(**1.5**)
wpbc	**73.2**(**8.8**)	69.0(12.9)	72.6(9.5)	72.6(9.5)	72.6(9.5)	67.0(8.4)	64.4(20.3)
zoo	**96.9**(**9.8**)	**96.9**(**9.8**)	93.6(10.1)	88.6(10.2)	89.6(8.5)	84.3(6.9)	87.7(10.5)
平均值	**83.24**	81.6	81.25	80.31	80.79	78.53	79.31

结合表 2、3 的实验结果可知:

1)从总体上看,SPFC 算法相比 CFS、ChiSquare、FCBF、Laplacian、NRS 以及 Relief 算法在 KNN、CART 基分类器下表现稳定,且均获得最高平均分类精度。相比考虑类信息的传统谱特征选择算法 Laplacian, SPFC 算法优于 Laplacian。

2)相比 ChiSquare、Laplacian、Relief 这 3 种同样获得特征系列的算法,SPFC 算法以相同的前 k 个特征在不同的基分类器下获得的平均分精度明显较高,相比子集约简的算法 CFS、FCBF、NRS,SPFC 取它们子集约简数量的最小值在两个基分类器下分类精度明显要高于 NRS,在 CART 基分类器下 SPFC 的分类精度高于 CFS、FCBF 达两个百分点以上,而在 KNN 基分类器下也显著高于 CFS、FCBF。

3)每一种算法均会在某一个分类器上某个数据集上获得最高分类精度,但只有 SPFC 能在两个基分类器上多个数据集上获得最高分类精度。SPFC 算法在数据集 ICU、rice、zoo 上性能提升更为明显,在两个分类器上均达到最高。而 ICU 为混合型数据,rice 为连续型数据、zoo 为离散型数据。说明 SPFC 可以处理多类型数据集,在大部分各类型数据集上 SPFC 均能达到较好的稳定表现。

实验 2　为了进一步研究比较 SPFC 算法与其他算法在两个分类器下的分类性能是否明显不同,我们采用 Friedman test 和 Bonferroni-Dunn 在统计上进行验证。Friedman 统计值定义为

$$x_F^2 = \frac{12N}{k(k+1)}\left[\sum_{i=1}^{k} R_i^2 - \frac{k(k+1)^2}{4}\right]$$

$$F_F = \frac{(N-1)x_F^2}{N(k-1)-x_F^2} \tag{5}$$

式中:k 代表对比算法个数,N 表示数据集个数,R_i 表示第 i 个算法在 8 个数据集上的排序均值(见表 4)。由表 4 结合式(5)算出 KNN 分类器下 F_F 的值为 2.18,cart 分类器下 F_F 的值为 3.05,又当显著性水平 $\alpha=0.1$ 时 $F(6,42)=1.87$,因此在两个分类器下都拒绝了零假设(所有算法性能相等),这时还需要结合特定的 post-hoc test 来进一步分析各个算法性能的差异。本文采用显著性水平为 0.1 的 Bonferroni-Dunn test。在这里定义两个算法的不同用下面的临界差:

$$\mathrm{CD}_\alpha = q_\alpha\sqrt{\frac{k(k+1)}{6N}}$$

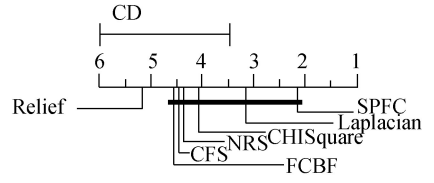
在 Bonferroni-Dunn test 里显著性水平为 0.1 且 7 个算法对比时 $q_\alpha=2.394$,因此 CD=2.58($k=7$,$N=8$)。如果两个算法在所有数据集上的平均排序的差不低于临界差 CD,则认为它们有显著性差异。图 2 给出了在两个分类器下 SPFC 算法与其他算法的比较,其中,每个子图中最上行为临界值,坐标轴画出了各种算法的平均排序且最左(右)边的平均排序最高(低)。用一根加粗的线连接性能没有显著差异的算法组。

从图 2 可以直观看出在 KNN 分类器下,SPEC 算法显著优于 Relief 算法,虽然与其他算法没有显著差别,但可以看出 SPFC 算法性能要高于其他算法;在 CART 分类器下 SPFC 算法性能显著优于算法 NRS、ChiSquare、Relief,而与算法 Laplacain、CFS、FCBF 性能相当,但性能相当的同一组里 SPFC 算法要远远优于算法 Laplacain、CFS、FCBF。

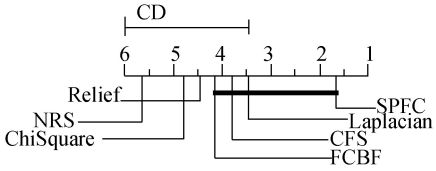
表 4 不同算法在两个分类器下的排序均值表

Table 4 Different algorithms under the two classifiers the mean of sorting table

数据集	SPFC	Laplacian	CFS	ChiSquare	FCBF	NRS	Relief
KNN	2.13	3.13	4.44	4.06	4.56	4.38	5.31
CART	1.63	3.44	3.81	4.81	4.13	5.63	4.44



(a) KNN 分类器



(b) CART 分类器

图 2 在 KNN 和 CART 分类器下 SPEC 与其他算法的比较

Fig.2 SPEC compared with other algorithms Under the CART and KNN classifier

4 结论

本文针对传统的谱特征选择只考虑特征的单独最优组合问题进行改进,提出基于谱图理论的特征相关的特征选择算法,本文研究发现:1)引入特征之间的统计相关性到谱特征选择中,能有效地解决有用特征可能是冗余的问题;2)在公开的 UCI 数据集上的实验结果表明,本文算法能够选择较少的特征,获得较好的分类精度;3)由表 2~4 中的数据亦看出考虑特征间的相关性算法(SPEC)比不考虑特征间相关性算法(Laplacian)能显著提高特征子集的分类性能。但由于本文实验采用欧式距离统计特征间的相关性,而欧式距离对于高维特征的计算差值变化不大,因此对于高维特征间的相关性的设计有待进一步研究。

参考文献:

[1] LIN Yaojin, Li Jinjin, LIN Peirong, et al. Feature selection via neighborhood multi-granulation fusion[J]. Knowledge-based systems, 2014, 67: 162-168.

[2] MANORANJAN D, LIU Huan. Consistency-based search in feature selection[J]. Artificial intelligence, 2003,151(1): 155-176.

[3] ZHANG C, ARUN K, CHRISTOPHER R. Materialization optimizations for feature selection workloads[J]. ACM transactions on database systems, 2016, 41(1): 2.

[4] 曹晋, 张莉, 李凡长. 一种基于支持向量数据描述的特征选择算法[J]. 智能系统学报, 2015, 10(2): 215-220.
CAO Jin , ZHANG li, LI Fanchang . A feature selection algorithm based on support vector data description[J]. CAAI transactions on intelligent systems, 2015, 10(2): 215-220 .

[5] MANORANJAN D, LIU Huan. Feature selection for classification[J]. Intelligent data analysis, 1997, 1(3): 131-156.

[6] SUN Yujing, WANG Fei, WANG Bo, et al. Correlation feature selection and mutual information theory based quantitative research on meteorological impact factors of module temperature for solar photovoltaic systems[J]. Energies, 2016, 10(1): 7.

[7] CVETKOVIC D M, ROWLINSON P. Spectral graph theory[J]. Topics in algebraic graph theory, 2004: 88-112.

[8] ZHAO Zheng, LIU Huan. Spectral feature selection for supervised and unsupervised learning[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 1151-1157.

[9] ZHAO Zhou, HE Xiaofei, CAI Deng, et al. Graph regularized feature selection with data reconstruction[J]. IEEE transactions on knowledge and data engineering, 2016, 28(3): 689-700.

[10] HE Xiaofei, CAI Deng, NIYONGI P. Laplacian score for feature selection[M].Cambridge: MIT Press, MA, 2005, 17: 507-514.

[11] BELABBAS M A, WOLFE P J. Spectral method in machine learning and new strategies for very large datasets[J]. Proceedings of the national academy of sciences, 2009, 106(2): 369-374.

[12] WANG Xiaodong, ZHANG Xu, ZENG Zhiqiang, et al. Unsupervised spectral feature selection with l 1-norm graph[J]. Neurocomputing, 2016, 200: 47-54.

[13] 边肇祺,张学工.模式识别[M]. 2 版. 北京: 清华大学出版社, 2000.

[14] HALL M A. Correlation-based feature selection for discrete

and numeric class machine learning [C]//the 17th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2000: 359-366.

[15] ANDREAS W, ANDREAS P. Attacks on stegan ographic systems[M]. Heidelberg, Berlin: Springer-Verlag, 2000: 61-76.

[16] YU Lei, LIU Huan. Efficient feature selection via analysis of relevance and redundancy [J]. Journal of machine learning research, 2004, 5(1): 1205-1224.

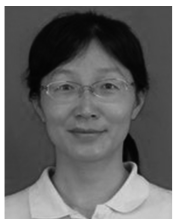
[17] HU Qinghua, YU Daren, LIU Jinfu, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information sciences, 2008, 178 (18): 3577-3594.

[18] CRAMMER K, GILAD-BACHRACH R, NAVOT A. Margin analysis of the lvq algorithm [C]//Advances in Neural Information Processing Systems. 2002, 14: 462-469.

[19] FRIEDMAN M, A comparison of alternative tests of significance for the problem of m rankings[J]. The annals of mathematical statistics, 1940, 11(1): 86-92.

[20] DUNN O J.Multiple comparisons among means[J]. Journal of the american statistical association, 1961, 56(293): 52-64.

作者简介:



胡敏杰,女,1979 年生,讲师,主要研究方向为数据挖掘。

林耀进,男,1980 年生,主要研究方向为数据挖掘、粒计算。主持国家自然科学基金 2 项。发表学术论文 60 余篇。

杨红和,男,1969 生,高级实验师,主要研究方向为数字校园。

2017 **Workshop on SAR in Big Data Era: Models, Methods and Applications**

During the last decade a series of SAR satellites has been launched, including Chinese Gaofen-3, providing great amount of SAR data with varied modes to meet the varieties of applications. It becomes a challenge to retrieve information from these big data. The main objective of this workshop is to share models, methods and applications of SAR data exploration in the big data era.

The workshop includes different subjects, such as big SAR data modeling, large-scale intelligent SAR processing, SAR applications in big data frameworks. It will feature keynote presentations by distinguished researchers on this topics, most of them are IEEE GRSS members.

Website:http://www.radi.ac.cn/BIGSARDATA2017/