基于相关系数的有效特征光谱筛选方法

申金媛[1]，　李　航[1]，　刘润杰[1]，　孔银亮[2]，　程仲记[2]

(1. 郑州大学 信息工程学院　河南 郑州 450001;
2. 河南省烟草公司 平顶山分公司　河南 平顶山 467000)

摘要: 为降低数据采集时间、分级模型的计算复杂度及提高烟叶分级速度,提出了一个基于半监督学习的有效特征光谱筛选方法. 首先定义判别特征好坏的鉴别函数 R,并根据 R 值基于半监督方法删除不好特征;然后利用不同特征间的相关系数,基于有监督学习方法去除相关度高光谱,进一步减少有效特征光谱的数目;最后利用全光谱和两次筛选后的特征光谱建立了 13 个等级的 SVM 分级模型. 实验结果表明所构造的光谱特征筛选模型,可从原始数据中筛选出有效特征光谱,从而极大地减少原始光谱采集量,在保证正确分级率的前提下,极大地提高了烟叶分级速度.

关键词: 离散度; 相关系数; 支持向量机; 烟叶分级; 相关性

中图分类号: TN219　　**文献标志码**: A　　**文章编号**: 1671-6841(2017)03-0028-04

DOI: 10. 13705/j. issn. 1671-6841. 2016274

0　引言

烟叶智能分级具有快速且准确率高的特点,可以避免人工分级的主观性. 目前智能分级主要依据烟叶的图像信息[1]或者光谱信息进行分级. 光谱信息可以很好地反映与烟叶等级密切相关的厚度、含油分、叶片结构等因素,光谱分析技术广泛应用于烟草行业中[2-3].

采集的光谱特征具有维数高、冗余度大的特点,分等级时需要降维处理. 第一类方法,利用主成分分析法[4-5]、小波分解法[6-7]、独立成分分析法[8-9]、连续投影法[10-11]、间隔最小二乘法[12-13]等方法对原始数据进行降维处理,提取特征. 这些方法可以有效地减少分类器的输入维数,从而降低分级模型的计算复杂度,但不可以减少原始光谱数据的采集时间,因此极大地影响了烟叶的整个分级速度. 第二类方法,直接从原始光谱中筛选出有效特征光谱,筛选特征光谱的方法主要有聚类算法[14]、粒子群算法[15]和遗传算法[16]. 这样采集数据时只需采集筛选后的特征光谱即可,不仅可以降低分级模型的计算复杂度,而且可以降低光谱数据采集量. 基于第二类算法思想,本文构造基于半监督学习的有效光谱特征选择模型,将筛选的特征采用 SVM 分类器进行验证,对 13 个等级的烟叶进行分级.

1　特征筛选原理及分类器

1.1　基于离散度的初筛选

假设训练集有 C 个类别,共计 N 个 P 维矢量. k 类中有 C_k 个样本,则 k 类中第 i 个样本表示为: $\boldsymbol{X}_i^{(k)} = (x_{i1}^{(k)}, x_{i2}^{(k)}, \cdots, x_{ij}^{(k)}, \cdots, x_{iP}^{(k)})'$,$(1 \leqslant i \leqslant C_k, 1 \leqslant j \leqslant P, 1 \leqslant k \leqslant C)$; k 类中第 j 个特征的平均值表示为: $\bar{\boldsymbol{X}}_j^{(k)} = \frac{1}{C_k}\sum_i x_{ij}^{(k)}$; C 个类别中第 j 个特征的平均值表示为: $\boldsymbol{Y}_j = (\bar{\boldsymbol{X}}_j^{(1)}, \bar{\boldsymbol{X}}_j^{(2)}, \cdots, \bar{\boldsymbol{X}}_j^{(k)}, \cdots, \bar{\boldsymbol{X}}_j^{(C)})$.

对于采集的烟叶的光谱特征,由聚类思想可知:相同特征在同一类别中的离散度越小越好;相同特征在不同类别中的离散度越大越好. 采集的原始光谱特征中某些特征不能更好地反映聚类思想,本文同时考虑相

收稿日期:2016-10-17
基金项目:河南烟草公司科技计划项目(M201335).
作者简介:申金媛(1966—),女,山西晋中人,教授,主要从事模式识别、光谱分析研究,E-mail: jyshen@ zzu. edu. cn.

同特征的类内离散度和类间离散度,实现方法如下:

1) 特征的类内离散度值表征为该类别的聚集性,基于聚类算法的思想,其值越小越有利于分级. 第 j 个特征的类内离散度函数为 $\alpha_j = \frac{1}{C}\sum_k \left(\frac{1}{C_k - 1}\sum_i \left(x_{ij}^{(k)} - \bar{\boldsymbol{X}}_j^{(k)}\right)\right)^2$.

2) 类间离散度值表征为相同特征在不同类别当中的差异性,在分级时,其值越大越有利于分级. 第 j 个特征的类间离散度函数为 $\beta_j = \frac{1}{C-1}\sum_k \left(\bar{\boldsymbol{X}}_j^{(k)} - \frac{1}{C}\sum_k \bar{\boldsymbol{X}}_j^{(k)}\right)^2$.

3) 定义判别特征好坏的鉴别函数 R,即相同特征的类内离散度与类间离散度的比值为

$$R_j = \frac{\alpha_j}{\beta_j}. \tag{1}$$

根据式(1)计算所有特征的鉴别函数值,将鉴别值按由小到大进行排序,并根据 R 值基于半监督的方法删除拐点右侧的不好特征. 删除不好特征后,余下的特征之间可能存在很强的相关性,在保证分级准确率的前提下,为获得更少的有效光谱特征数目和加快分级速度,需要进行特征深度筛选.

1.2　基于相关系数的深筛选

相关系数分析可以有效地进行特征的筛选[17],主要思想是:在众多相关性特征中,筛选出一个代表特征,用它来表示这些相关性大的特征,去除其余特征. 这样可以选取更少的特征变量,减少光谱数据采集量和分级模型的计算复杂度. 特征 x 和特征 y 的相关系数计算公式为

$$C_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2},$$

其中:$\bar{x} = \sum_{i=1}^{n} x_i/n$;$\bar{y} = \sum_{i=1}^{n} y_i/n$;$n$ 为训练集个数.

依据相关系数法进行特征深度筛选的方法为:假设初筛选后余下 m 个特征,它们的鉴别函数集合为 $u = \{u_1, u_2, \cdots, u_m\}$,设定合适的阈值,选取 u 中值最小的特征作为初选特征. 在 k 类中计算该特征与其余特征的相关系数,将相关系数大于阈值的特征记为 C_k,求取 C 个类别的特征交集,用初选特征代表所有级别中的交集特征,并在 u 中删除交集的特征. 在删除交集特征后的集合中,选取值最小的特征为第二个被选特征,同样的方法求取所有类别中大于阈值的特征的交集,用它代表所有级别中的交集特征,再在 u 中删除交集的特征. 按照同样规则选取特征,直至 u 为空集.

1.3　SVM 分类器

支持向量机模型同时考虑经验风险和结构风险最小,对小样本、高维数据的分类具有良好的推广能力. 本文通过线性核函数实现数据由低维向高维的映射,判决函数为 $g(x) = \mathrm{sgn}\left[\sum_{i=1}^{n} \alpha_i d_i K(\boldsymbol{x}_i, x) + b\right]$. 其中:$K(\boldsymbol{x}_i, x)$ 为核函数;$\boldsymbol{x}_i$ 为训练样本的支持向量;x 为测试样本;b 为由 $\boldsymbol{x}_i$ 确定的阈值;d_i 为训练样本的标签;α_i 为拉格朗日乘子. 由于 SVM 是两分类问题,本文采用并行投票式进行分类,分类器个数为 $N(N-1)/2$,其中 N 为类别的数目.

2　实验及结果分析

2.1　实验数据及预处理

实验样本为郑州市烟草局提供的 13 个等级的烟叶,包含有 *B2F*、*B3F*、*B4F*、*C2F*、*C2L*、*C3F*、*C3L*、*X2F*、*X2L*、*X3F*、*X3L*、*X4F*、*X4L*,采用日本岛津公司生产的 UV3600 型号的光谱仪,采集每片烟叶的反射光谱,光谱范围为 1 500 ~ 2 400 nm,采样间隔为 2 nm,共有 642 条反射光谱. 随机选取三分之一的样本为训练集,其余样本作为测试集验证模型的推广能力. 为消除光谱仪带来的基线漂移和噪声,对采集的光谱数据进行以下预处理:

$$y_i = (m_i - \min(m_i)) / (\max(m_i) - \min(m_i)).$$

其中:m_i 为未预处理的原始光谱;y_i 为归一化后的光谱;$\max(m_i)$ 和 $\min(m_i)$ 分别为 m_i 的最大值和最小值.

2.2　特征的初筛选

依据公式(1)进行光谱数据的预处理,计算特征的类内离散度与类间离散度的比值,按由小到大进行排

序,得到的拐点和删除拐点右侧特征后识别率的结果如图1所示.

以原始451个光谱特征作为SVM的输入,训练集、测试集正确率分别为100%、90.89%.在离散度比值由小到大排序后的10个拐点中,删除拐点右侧特征后识别率由图1可知. 第6个拐点下的训练集和测试集的正确率分别为100%、94.39%,识别率明显高于其余拐点和全光谱,此时余下326个特征,相比全光谱下的451特征有一定的减少.根据鉴别函数半监督的学习方法去掉部分离散度大的特征,不仅提高了分级正确率,实现特征的初步筛选,而且为下一步进行特征深度筛选模型降低了输入维数.

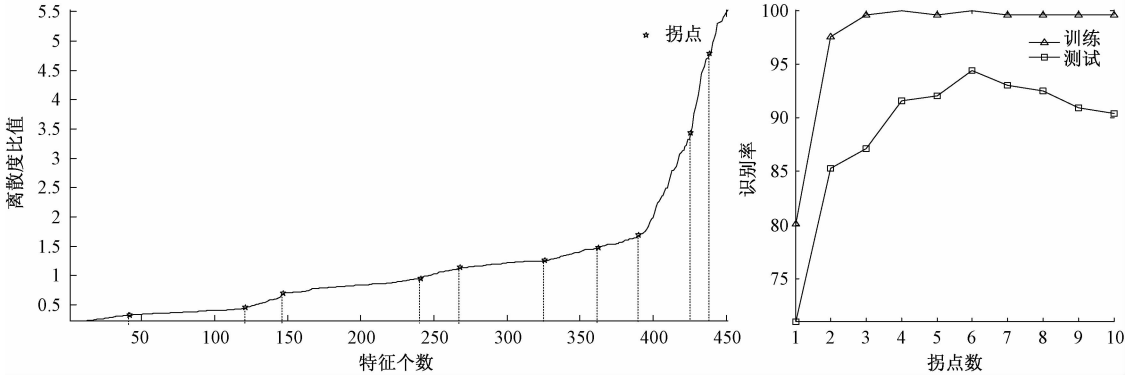

图1 排序后的拐点和各拐点下的识别率

Fig. 1 The inflection point after sorted and recognition rate of the turning point

2.3 去相关特征

对初筛选后余下的326个特征进行相关系数分析,进一步去除相关性大的特征,进行特征深度筛选.设定不同阈值,将余下特征作为SVM的输入、分级识别率、分级时间、特征数目随阈值变化结果如图2所示.随着阈值的减小,余下特征数目和分级时间越来越少,准确率呈现先减小后增大的趋势,阈值为0.995时取得最大值95.21%.说明去除部分相关性的特征可以提高准确率,小于一定阈值后特征数目过少,分级准确率会降低.为寻找更好的阈值,在保证分级准确率不低于全光谱特征条件下的准确率,细化阈值范围(0.99~1)得到结果如图3所示. 综合图2和图3,最少的特征数目为155个,相比原来451个,减少了65%,可以极大的减少光谱的采集量和提高分级速度.在特征数目没有限制下,分级准确率最高可以达到95.21%,特征数目为207个,分级时间比全光谱有所下降,减少一半的光谱采集量,从而加快了整个系统的分级速度.

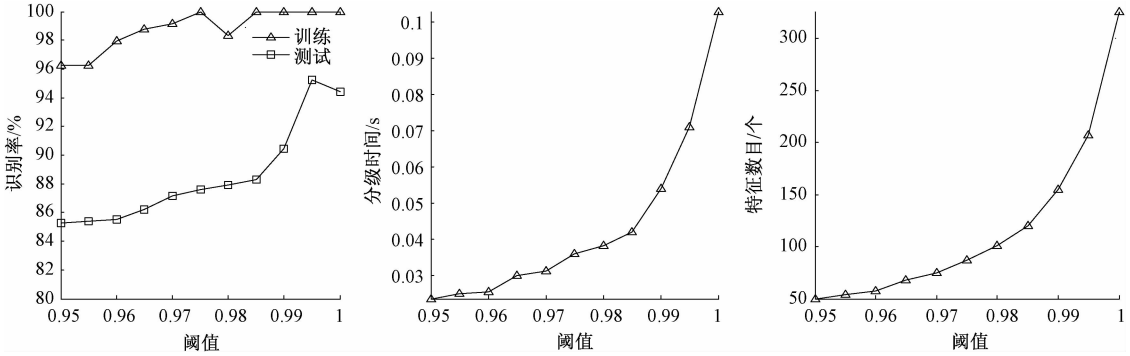

图2 识别率、分级时间、特征数目随阈值的变化

Fig. 2 Identification and classification time, number of features along with the change of threshold value

3 结论

通过以上工作得出以下结论:1)可以依据烟叶的光谱特征实现烟叶的智能分级.2)投票式的支持向量机可以作为实现烟叶分级的分类器.3)利用同一特征的类内离散度与类间离散度比值可删减部分对分级不好的特征,特征间的相关系数分析可以删减相关性特征.

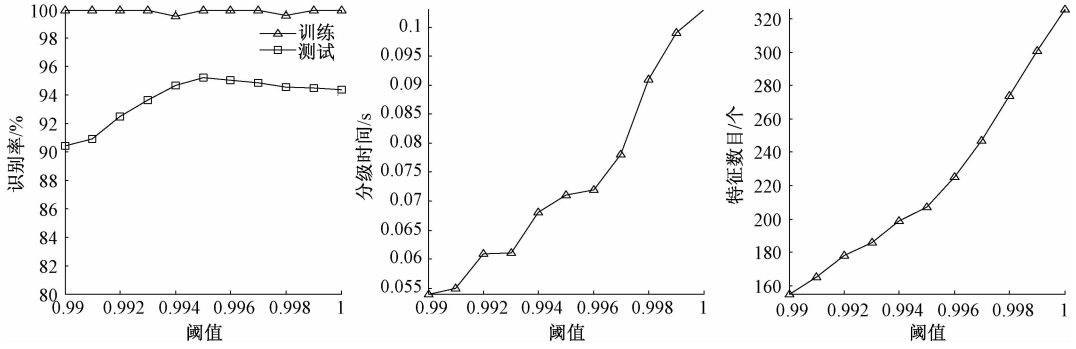

图 3　细化阈值下的结果

Fig. 3　Detailed results of threshold

如果分级系统为串行式的,减少光谱数据的采集时间可极大地提高烟叶的分级速度,使烟叶收购阶段的实时分级成为了可能. 烟叶的图像特征对分级也有一定的影响,将图像信息和光谱信息相结合是今后改进的方向.

参考文献:

[1] 王夏,贺立源. 烤烟烟叶反射和透射图像的同步图像分割[J]. 武汉大学学报(信息科学版),2014,39(8):998 – 1002.

[2] 田旷达,邱凯贤,李祖红,等. 近红外光谱法结合最小二乘支持向量机测定烟叶中钙、镁元素[J]. 光谱学与光谱分析,2014, 34(12):3262 – 3266.

[3] 任晓,劳彩莲,徐照丽,等. 估测田间烟叶色素含量的光谱模型研究[J]. 光谱学与光谱分析,2015, 35(6):1654 – 1659.

[4] 王毅,马翔,温亚东,等. 近红外光谱与多元统计方法用于生产过程实时分析[J]. 光谱学与光谱分析,2013, 33(5):1226 – 1229.

[5] 秦玉华,丁香乾,宫会丽. 高维特征选择方法在近红外光谱分类中的应用[J]. 红外与激光工程,2013, 33 (5):1355 – 1359.

[6] 彭丹青,申金媛,刘剑君,等. 基于径向基网络的烟叶光谱分级[J]. 农机化研究,2009, 53(10):15 – 18.

[7] 罗霞,洪添胜,罗阔,等. 小波变换和连续投影算法在火龙果总酸无损检测中的应用[J]. 光谱学与光谱分析,2016, 36(5):1345 – 1351.

[8] 侯振雨,王伟,蔡文生,等. 基于独立成分的局部建模方法及其在近红外光谱分析中的应用研究[J]. 计算机与应用化学, 2006, 23(3):224 – 226.

[9] 王功明,刘志勇. 基于光谱表示和独立成分分析的混合颜料成分分析方法[J]. 光谱学与光谱分析,2015, 35 (6):1682 – 1689.

[10] 杨凯,蔡嘉月,张朝平,等. 应用近红外光谱投影模型法分析烟叶的部位特征[J]. 光谱学与光谱分析,2014, 34(12):3277 – 3280.

[11] 熊雅婷,李宗朋,王健,等. 基于连续投影算法的黄酒成分检测模型[J]. 食品与发酵工业,2015, 41 (3):185 – 190.

[12] 章海亮,孙旭东,刘燕德,等. 近红外光谱检测苹果可溶性固形物[J]. 农业工程学报,2009,25(S2):340 – 344.

[13] 於海明,李石,吴威,等. 稻谷千粒质量近红外光谱预测模型的波长选择方法[J]. 农业机械学报,2015, 46(11):275 – 279.

[14] 赵海东,申金媛,刘润杰,等. 基于聚类的烟叶近红外光谱有效特征的筛选方法[J]. 红外技术,2013, 35 (10):659 – 664.

[15] 李航,赵海东,申金媛,等. 基于 BPSO 和 SVM 的烟叶近红外有用特征光谱选择[J]. 物理实验,2015, 35(6):8 – 12.

[16] 王徽蓉,李卫军,刘杨阳,等. 基于遗传算法与线性鉴别的近红外光谱玉米品种鉴别研究[J]. 光谱学与光谱分析,2011, 31(3): 669 – 672.

[17] 周金治,唐肖芳. 基于相关系数分析的脑电信号特征选择[J]. 生物医学工程学杂志,2015, 32(4):735 – 739.

(下转第 38 页)

[13] TZORTZIS G, LIKAS A. Kernel-based weighted multi-view clustering[C]//Proceedings of the IEEE 12th International Conference on Data Mining. Washington,2012: 675 - 684.
[14] 陈宝林. 最优化理论与算法[M]. 北京:清华大学出版社,1989:411 - 432.
[15] ZANGWILL W I, MOND B. Nonlinear programming: a unified approach[M]. New Jersey: Prentice-Hall Englewood Cliffs, 1969.
[16] HATHAWAY R, BEZDEK J, TUCKER W. An improved convergence theory for the fuzzy isodata clustering algorithms[J]. Analysis of fuzzy information, 1987, 3: 123 - 132.
[17] 张志华, 郑南宁, 史罡. 极大熵聚类算法及其全局收敛性分析[J]. 中国科学:技术科学, 2001,31(1):59 - 70.

A Convergence Proof of Multi-view Kernel *K*-means Clustering Algorithm

QIU Baozhi[1], HE Yanfang[2]
(1. *School of Information Engineering, Zhengzhou University, Zhengzhou* 450001, *China*;
2. *Department of Information Engineering, Guangdong Polytechnic College, Zhaoqing* 526100, *China*)

Abstract: The Zangwill convergence theorem was utilized to analyze the convergence of the multi-view kernel *K*-means (MVKKM). The study results showed that, under certain conditions, the iterative sequence generated by MVKKM converges, or there existed at least one subsequence converged to a local minimum or a saddle point of the objective function of the algorithm. And in Matlab, the convergence of the algorithm under different views and different index weight was verified.

Key words: multi-view clustering; *K*-means; kernel functions; convergence

(责任编辑:王浩毅)

(上接第31页)

Screening the Effective Spectrum Features Based on Correlation Coefficient

SHEN Jinyuan[1], LI Hang[1], LIU Runjie[1], KONG Yinliang[2], CHENG Zhongji[2]
(1. *School of Information Engineering, Zhengzhou University, Zhengzhou* 450001, *China*;
(2. *Pingdingshan Branch of Henan Provincial Tobacco Company, Pingdingshan* 467000, *China*)

Abstract: To increase the tobacco classification speed, it was necessary to reduce the data acquisition time and the computational complexity of the classification mode. An effective spectral filter method based on a semi-supervised learning was constructed to reduce the amount imput data. The discriminant function of *R* that determined an input spectrum good or bad was defined. The bad spectra were pruned based on the value of *R* and semi-supervised learning method. In order to further reduce the spectral data, the correlation coefficient between different spectra was employed to remove those spectra with higher correlation based on the supervised method. The training samples with original spectra and the characteristic spectra after twice screening were used to construct SVM tobacco leaf classifiers of 13 grades respectively. The results of experiments showed that the proposed feature screening method was effective. It could greatly reduce the grading time while of guaranteeing the correct classification rate.

Key words: discreteness; correlation coefficient; SVM; tobacco grade; correlation

(责任编辑:方惠敏)