文章编号:1001 - 9081(2010)04 - 0993 - 04

基于改进的 *F*-score 与支持向量机的特征选择方法

谢娟英,王春霞,蒋 帅,张 琰
(陕西师范大学 计算机科学学院,西安 710062)
(xiejuany@ snnu. edu. cn)

摘 要:将传统 *F*-score 度量样本特征在两类之间的辨别能力进行推广,提出了改进的 *F*-score,使其不但能够评价样本特征在两类之间的辨别能力,而且能够度量样本特征在多类之间的辨别能力大小。以改进的 *F*-score 作为特征选择准则,用支持向量机(SVM)评估所选特征子集的有效性,实现有效的特征选择。通过 UCI 机器学习数据库中六组数据集的实验测试,并与 SVM、PCA + SVM 方法进行比较,证明基于改进 *F*-score 与 SVM 的特征选择方法不仅提高了分类精度,并具有很好的泛化能力,且在训练时间上优于 PCA + SVM 方法。

关键词: *F*-score ;支持向量机;特征选择;主成分分析;核函数主成分分析
中图分类号: TP18 **文献标志码**:A

Feature selection method combing improved *F*-score and support vector machine

XIE Juan-ying, WANG Chun-xia, JIANG Shuai, ZHANG Yan
(*School of Computer Science, Shaanxi Normal University, Xi' an Shaanxi* 710062, *China*)

Abstract: The original *F*-score can only measure the discrimination of two sets of real numbers. This paper proposed the improved *F*-score which can not only measure the discrimination of two sets of real numbers, but also the discrimination of more than two sets of real numbers. The improved *F*-score and Support Vector Machines (SVM) were combined in this paper to accomplish the feature selection process where the improved *F*-score was used as the evaluation criterion of feature selection, and SVM to evaluate the features selected via the improved *F*-score . Experiments have been conducted on six different groups from UCI machine learning database. The experimental results show that the feature selection method, based on the improved *F*-score and SVM, has high classification accuracy and good generalization, and spends less training time than that of the Principle Component Analysis (PCA) + SVM method.

Key words: *F*-score ; Support Vector Machine (SVM); feature selection; Principle Component Analysis (PCA); Kernel Principal Component Analysis (KPCA)

0 引言

理论上,特征越多越能提供比较好的识别能力,但是面对实际学习过程时,对于有限的训练数据,过多的特征不仅大大降低学习速度,同时也会导致分类器对训练数据的"过适应"问题[1-3],尤其是那些与类别不相关的特征和冗余特征,会导致参数估计的准确率下降,进而影响分类器的性能。因此,不少学者对降低样本特征的方法进行了研究[4-5],以期得到对分类最有效的特征。

特征提取和特征选择是两种常用的降维方法。特征提取是将原有特征空间进行某种形式的变换,以得到新的特征。主成分分析 (Principle Component Analysis, PCA)和独立成分分析(Independent Components Analysis, ICA)是特征提取中最常用的方法[6],它们对许多学习任务都可以较好地降维,但是特征的理解性很差,因为即使简单的线性组合也会使构造出的特征难以理解,而在很多情况下,特征的理解性是很重要的。另外,由于特征提取的新特征通常由全部原始特征变换得到,从数据收集角度看,并没有减少工作量。

特征选择是从原有的特征集中选择一个最优特征子集,这个特征子集保留了原有特征集的全部或大部分类别信息。通过特征选择,一些无关的或者冗余的特征被剔除,简化后的数据集常会得到更精确的模型,也更容易理解[7]。根据特征选择是否依赖于学习算法,特征选择方法分为 Filter 和 Wrapper 两大类[8]。Filter 特征选择方法的评估准则直接由数据集求得,独立于学习算法,具有计算代价小、效率高等特点。Wrapper 特征选择可能会比 Filter 特征选择的降维效果好,但该算法因为计算代价大而效率较低。

本文结合 Filter 和 Wrapper 的优点,以提出的改进 *F*-score 作为 Filter 评估准则,用 SVM 作 Wrapper 评估方法,除去不相关的冗余特征,得到一组最佳特征子集。利用 UCI 机器学习数据库中的六组数据集进行实验测试,结果表明,基于改进的 *F*-score 与支持向量机的特征选择方法具有很好的泛化能力和较高的分类精度。

1 特征选择方法

1.1 传统的 *F*-score 特征选择方法

特征选择是从众多特征中选择出那些对分类识别最有效的特征,从而实现特征空间维数的压缩。*F*-score 是一种衡量

收稿日期:2009 - 10 - 12;**修回日期**:2009 - 12 - 17。

作者简介:谢娟英(1971 -),女,陕西西安人,副教授,博士,主要研究方向:智能信息处理、模式识别、机器学习; 王春霞(1985 -),女,甘肃秦安人,硕士研究生,主要研究方向:智能信息处理、模式识别; 蒋帅(1985 -),女,吉林松原人,硕士研究生,主要研究方向:智能信息处理、模式识别; 张琰(1982 -),女,山西平顺人,硕士研究生,主要研究方向:智能信息处理。

特征在两类之间分辨能力的方法,通过此方法可以实现最有效特征的选择[9],具体描述如下。

给定训练样本集 $x_k \in \mathbf{R}^m$, $k = 1,2,\cdots,n$,其中正类和负类的样本数分别为 n_+ 和 n_-。则训练样本第 i 个特征的 F-score 定义为:

$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \tag{1}$$

其中:$\bar{x}_i$, $\bar{x}_i^{(+)}$, 和$\bar{x}_i^{(-)}$ 分别为第 i 个特征在整个数据集上的平均值,在正类数据集上的平均值和在负类数据集上的平均值;$x_{k,i}^{(+)}$ 为第 k 个正类样本点的第 i 个特征的特征值;$x_{k,i}^{(-)}$ 为第 k 个负类样本点的第 i 个特征的特征值。F 值越大,此特征的辨别力越强,如图 1 所示。

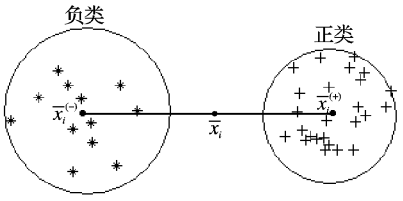

图 1 两类情况

1.2 改进的 F-score 特征选择方法

上述 F-score 是一种简单、有效的特征选择方法,它能够衡量特征在两类之间的辨别力大小。但是,此方法存在一定的局限性,它只适用于两类情况的特征选择,不能直接应用于多类问题中的特征选择。但实际需要解决的问题一般是多类问题,因而将 F-score 进行扩展使它适合于多类情况的特征选择是非常必要的。根据式(1),从计算 F-score 值的表达式开始,将两类情况下 F-score 值的计算公式进行推广,提出了改进的 F-score 特征选择方法。它既能够衡量特征在两类之间的辨别力大小,也能够衡量特征在多类之间的辨别力大小。改进的 F-score 特征选择方法描述如下:

给定训练样本集 $x_k \in \mathbf{R}^m$, $k = 1,2,\cdots,n$;$l(l \geqslant 2)$ 为样本类别数,n_j 为第 j 类的样本个数,其中 $j = 1,2,\cdots,l$。则训练样本第 i 个特征的 F-score 定义为:

$$F_i = \frac{\sum_{j=1}^{l}(\bar{x}_i^{(j)} - \bar{x}_i)^2}{\sum_{j=1}^{l}\frac{1}{n_j - 1}\sum_{k=1}^{n_j}(x_{k,i}^{(j)} - \bar{x}_i^{(j)})^2} \tag{2}$$

其中:$\bar{x}_i$,$\bar{x}_i^{(j)}$ 分别为第 i 个特征在整个数据集上的平均值、在第 j 类数据集上的平均值;$x_{k,i}^{(j)}$ 为第 j 类第 k 个样本点第 i 个特征的特征值。

改进的 F-score 特征选择方法是基于类别可分性准则的,本文采用的是基于类内类间距离的类别可分性评价准则。在式(2)中分子表示的是各类的近似类间距离之和,分母表示总类内样本协方差。F- score 值越大,表明此特征的分类辨别力越强,即类间越疏,类内越密,分类效果就越好,也就是说此特征的辨别力就越强,如图 2 所示。

1.3 支持向量机评估方法

现有研究通常通过直观观察来剔除那些 F-score 值较小的特征[1,6],但这并不能保证所选特征子集是最优的。本文采用支持向量机的分类正确率作为所选特征子集的评估方法,以此决定需要剔除的特征,实现特征选择。

支持向量机是在统计学习的 VC 维和结构风险最小化理论基础上提出[10-11]的。该理论在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势,它根据有限的样本信息在模型复杂性和学习能力之间寻求最佳折中,以获得最好的推广能力。支持向量机的基本思想是将输入空间线性不可分的数据通过核函数映射到一个高维特征空间,通过在特征空间求解一个线性约束的二次规划问题,寻找一个能将数据线性分割的最大间隔分类面。

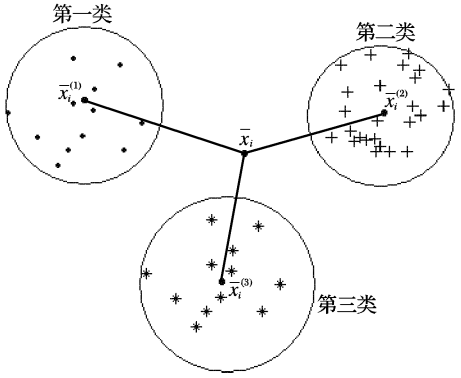

图 2 多类情况

对于线性可分的两类分类问题,支持向量机旨在寻求将两类样本分开且保证分类间隔最大的最优分类面。假设线性可分样本集为 $(x_1,y_1),\cdots,(x_l,y_l)$,其中 $x_i \in \mathbf{R}^n, y_i \in \{-1,+1\}, i = 1,\cdots,l$。线性判别函数的一般形式为 $g(x) = \boldsymbol{w}\cdot\boldsymbol{x} + b$,相应的分类面为 $\boldsymbol{w}\cdot\boldsymbol{x} + b = 0$。寻求最优分类面转化为下面的优化问题:

$$\min \varphi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i) + b] - 1 \geqslant 0;\ i = 1,2,\cdots,l \tag{3}$$

其对偶问题为:

$$\max Q(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle$$
$$\text{s.t. } \sum_{i=1}^{l}\alpha_i y_i = 0 \tag{4}$$
$$\alpha_i \geqslant 0; i = 1,2,\cdots,l$$

通过求解式(4),可得到最优分类函数为:

$$f(x) = \operatorname{sgn}\left\{\sum_{i=1}^{l}\alpha_i y_i\langle \boldsymbol{x}_i,\boldsymbol{x}\rangle + b\right\} \tag{5}$$

对应 $\alpha_i \neq 0$ 的样本称为支持向量。

在线性不可分情况下,引入松弛变量 ξ_i 和惩罚参数 C,式(3)转化为:

$$\min \varphi(\boldsymbol{w},\xi) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t. } y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i) + b] \geqslant 1 - \xi_i;\ i = 1,2,\cdots,l \tag{6}$$
$$\xi_i \geqslant 0; i = 1,2,\cdots,l$$

其对偶问题为:

$$\max Q(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle$$
$$\text{s.t. } \sum_{i=1}^{l}\alpha_i y_i = 0 \tag{7}$$
$$C \geqslant \alpha_i \geqslant 0;\ i = 1,2,\cdots,l$$

对于一般非线性问题,则可通过定义适当的核函数实现非线性变换,将输入空间映射到一个高维特征空间,然后在特征空间求解最优线性分类面。引入核函数后,以上的内积可用核函数代替。此时,SVM 的决策函数为:

$$f(\boldsymbol{x}) = \operatorname{sgn}\left\{\sum_{i=1}^{l}\alpha_i y_i K(\boldsymbol{x}_i,\boldsymbol{x}) + b\right\} \tag{8}$$

不同的内积核函数导致不同的支持向量机算法,目前采用的内积核函数主要有以下四类:

1) 线性核函数 $K(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{x}\cdot\boldsymbol{x}'$;

2) 多项式核函数函数 $K(\boldsymbol{x},\boldsymbol{x}') = (\boldsymbol{x}\cdot\boldsymbol{x}'+1)^d$;

3) 径向基核函数 $K(\boldsymbol{x},\boldsymbol{x}') = \exp(-\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\sigma^2)$;

4) S 型核函数 $K(\boldsymbol{x},\boldsymbol{x}_i) = \tanh(v(\boldsymbol{x}\cdot\boldsymbol{x}_i)+c)$。

2　基于改进 F-score 和 SVM 的特征选择

结合改进 F-score 和 SVM 算法,本文提出了改进的 F-score + SVM 的特征选择模型,以期剔除 F-score 值较低的特征/变量来进行样本特征的选择。然而在剔除 F-score 值较低的特征过程中,阈值的设定至关重要,这直接决定了特征选择的效果。在以往的研究中,对阈值的选取通常是采用直观观察来剔除 F-score 值较低的特征[9]。尽管这种方法达到了对样本特征降维的目的,但是很难实现特征子集的最优选取。针对这一问题,本文提出了利用支持向量机作为评估方法来选取最优的特征子集。由 1.2 节可知,特征的 F-score 值越大,特征的辨别力越强,所以,首先计算每个特征的 F-score 值;然后对每个特征根据其 F-score 值进行降序排序;每次从未被选取的特征中选择一个 F-score 值最大的特征添加到被选特征集合(被选特征集合初始为空集);再应用 SVM 算法对当前选取的特征子集进行评价。每次迭代中,采用 SVM 的分类正确率作为当前被选特征子集的评估,迭代一直进行,直到所有特征都加入被选特征集。根据 SVM 的分类精度选择分类效果最佳的特征子集。如此通过将改进的 F-score 和支持向量机方法相结合,达到了最佳特征子集的选取。该过程的模型描述如图 3 所示。

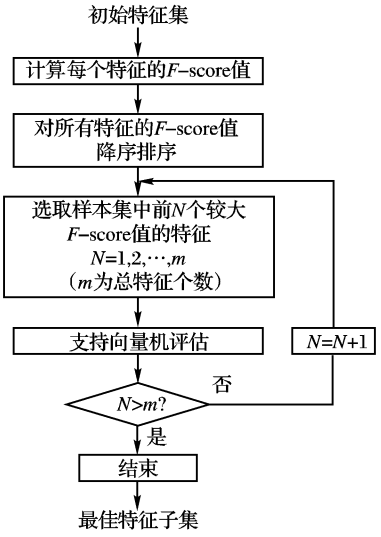

图 3　特征选择模型

3　实验结果与分析

采用 UCI 机器学习数据库中的 6 组数据集(diabetes, newthyroid, cmc, wine, dermatology, satimage, 如表 1 所示)进行测试。训练数据集与测试数据集采用 6∶4的比例进行随机划分。利用训练数据集进行最优特征子集的迭代选择。测试数据集对所选择的最优特征子集的分类有效性进行评价。首先计算每组数据集中每个特征的 F-score 值,结果如图 4 所示;再对其进行降序排序,表 2 为排序后的特征序列;然后每次从未入选的特征集中选择一个 F-score 值最大的特征添加到被选特征集合中,被选特征集合初始为空,每次迭代中,采用支持向量机进行分类,分类结果如图 5 所示;最后根据分类正确率来确定最佳特征子集,即选取的特征数最少,分类正确率最高的特征子集,表 3 为最终选取的最佳特征子集。

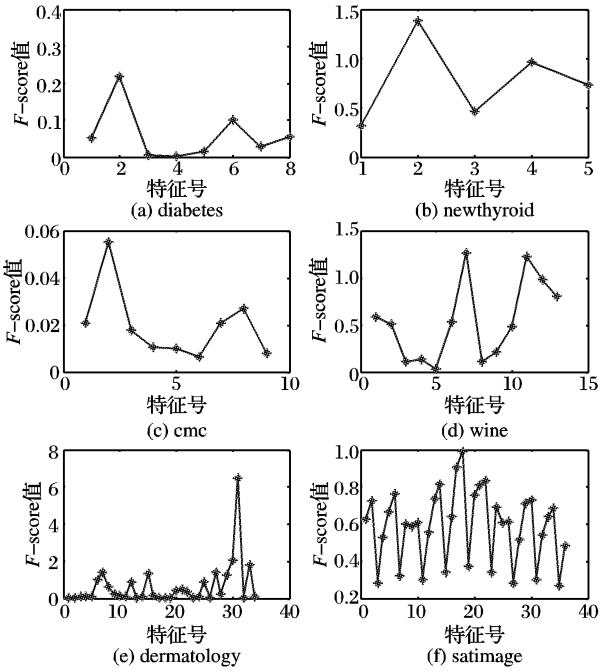

图 4　每个特征的 F-score 值

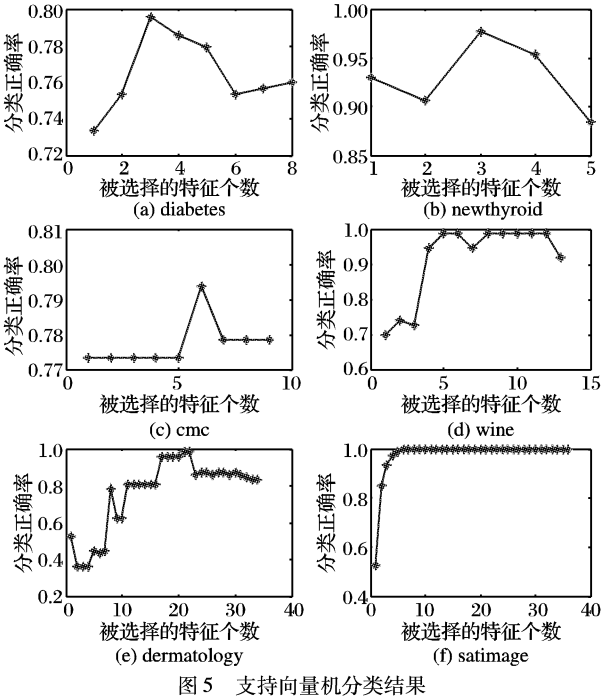

图 5　支持向量机分类结果

表 1　6 组数据集的样本个数和类别数

数据集	样本个数	类别数
diabetes	768	2
newthyroid	215	3
cmc	1 473	3
wine	178	3
dermatology	358	6
satimage	6 435	7

将改进 F-score + SVM 特征选择方法的实验结果与 SVM 方法,以及 PCA + SVM 方法的实验结果进行比较,表 4 为分类正确率的比较,表 5 为训练时间的比较。实验结果表明,基于改进 F-score 与 SVM 的特征选择方法获得了较好的识别结

果且训练时间优于 PCA + SVM 方法。

表 2 *F*-score 值排序结果

数据集	总特征个数	*F*-score 值降序排序后对应的特征号
diabetes	8	2,6,8,1,7,5,3,4
newthyroid	5	2,4,5,3,1
cmc	9	2,8,1,7,3,4,5,9,6
wine	13	7,11,12,13,1,6,2,10,9,4,8,3,5
dermatology	34	31,30,33,7,27,15,29,6,12,25,8,21,20,22,28,9,16,10,14,34,5,3,11,4,24,23,19,2,26,1,17,32,18,13
satimage	36	18,17,22,14,21,6,20,13,30,2,29,24,34,5,16,33,1,26,25,10,8,9,12,32,4,28,36,19,15,23,7,31,11,27,3,35

表 3 最佳特征子集

数据集	选择后的特征个数	被选择的特征
diabetes	3	2,6,8
newthyroid	3	2,4,5
cmc	6	2,8,1,7,3,4
wine	5	7,11,12,13,1
dermatology	21	31,30,33,7,27,15,29,6,12,25,8,21,20,22,28,9,16,10,14,34,5
satimage	10	18,17,22,14,21,6,20,13,30,2

表 4 改进的 *F*-score + SVM、SVM、PCA + SVM 的分类正确率比较

数据集	分类正确率/%		
	SVM	PCA + SVM	Improved *F*-score + SVM
diabetes	75.97	75.32	79.55
newthyroid	88.37	90.70	97.67
cmc	77.83	77.33	79.36
wine	91.78	64.38	98.63
dermatology	83.33	76.71	98.61
satimage	100.00	100.00	100.00

表 5 改进的 *F*-score + SVM、SVM、PCA + SVM 训练时间比较

数据集	训练时间/s		
	SVM	PCA + SVM	Improved *F*-score + SVM
diabetes	2.172	5.485	0.391
newthyroid	0.032	0.046	0.016
cmc	21.469	72.641	20.235
wine	1.250	22.688	2.242
dermatology	0.328	0.547	0.297
satimage	18.937	12.312	12.078

4 结语

本文首先对传统的 *F*-score 进行改进，提出了改进的 *F*-score，使其不但能够衡量样本特征在两类之间的辨别能力，而且能够衡量特征在多类之间的辨别能力。将改进的 *F*-score 与支持向量机结合，实现了最有效特征的选择。实验证明，基于改进的 *F*-score 与支持向量机的特征选择方法能有效提高分类正确率，并具有很好的泛化能力和较高的训练速度。

参考文献：

[1] LEE M C. Using support vector machine with a hybrid feature selection method to the stock trend prediction [J]. Expert Systems with Applications, 2009, 36(8): 10896 - 10904.

[2] MALDONADO S, WEBER R. A wrapper method for feature selection using support machines [J]. Information Sciences, 2009, 179(13): 2208 - 2217.

[3] LIU Y, ZHENG Y F. FS_SFS: A novel feature selection method for support vector machines [J]. Pattern Recognition, 2006, 39(7): 1333 - 1345.

[4] HUA J P, TEMBE W D, DOUGHERTY E R. Performance of feature-selection methods in the classification of high-dimension data [J]. Pattern Recognition, 2009, 42(3): 409 - 424.

[5] GUNAL S, GEREK O N, ECE D G, *et al*. The search for optimal feature set in power quality event classification [J]. Expert Systems with Applications, 2009, 36(7): 10266 - 10273.

[6] WIDODO A, YANG B S. Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors [J]. Expert Systems with Applications, 2007, 33(1): 241 - 250.

[7] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. Machine Learning Research, 2003, 3: 1157 - 1182.

[8] TALAVERA L. An evaluation of filter and wrapper methods for feature selection in categorical clustering [C]// Proceedings of 6th International Symposium on Intelligent Data Analysis. Madrid: Springer, 2005: 440 - 451.

[9] CHEN Y W, LIN C J. Combining SVMs with various feature selection strategies[EB/OL]. [2009 - 08 - 10]. http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf.

[10] VAPNIK V N. The nature of statistical learning theory [M]. New York: Springer, 1995.

[11] BURGES C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121 - 167.

(上接第 992 页)

参考文献：

[1] TAHANI H, KELLER J M. Information fusion in computer vision using the fuzzy integral [J]. IEEE Transactions on Systems, Man and Cybemetics, 1990, 20(3): 733 - 741.

[2] 章新华，林良骥，王骥程. 目标识别中信息融合的准则和方法[J]. 软件学报，1997，8(4)：303 - 307.

[3] KELLER J M, GADER P, TAHANI H, *et al*. Advances in fuzzy integration for pattern recognition [J]. Fuzzy Sets and Systems, 1994, 65(3): 273 - 283.

[4] 刘汝杰，袁保宗，唐晓芳. 用遗传算法实现模糊测度赋值的一种多分类器融合算法[J]. 电子学报，2002，30(1)：145 - 147.

[5] ZHAN Y, YE J, NIU D, *et al*. Facial expression recognition based on Gabor wavelet transformation and elastic templates matching [J]. International Journal of Image and Graphics, 2006, 6(1): 125 - 138.

[6] 成科扬，文传军，詹永照. 模糊深隐马尔可夫模型研究[J]. 计算机科学，2008，35(6)：163 - 167.

[7] BILMES J A. Buried Markov models for speech recognition [C]// Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Phoenix: IEEE Press, 1999: 713 - 716.