广义邻域粗集下的集成特征选择及其选择性集成算法

马超，陈西宏，徐宇亮，王光明
（空军工程大学导弹学院，713800，陕西三原）

摘要：针对实际模式识别系统中样本特征常具有的连续值属性、高维性、强相关性和冗余性等影响分类效果的问题，在广义邻域粗集模型下提出一种集成特征选择及其选择性集成算法. 该算法先提取样本特征并利用所提出的马氏距离分布熵评估其重要度，再基于特征重要度构建广义邻域粗集模型，并在此模型上以特征重要度为启发式信息设计基于蚁群算法的属性约简算法，然后通过改变广义邻域粗集模型参数的方式获得更多具有更大差异性的基分类器，最后利用主成分分析法对产生的基分类器进行选择性集成. 模拟电路故障诊断结果表明，该算法比 AdaBoost 等算法取得的分类精度至少提高了 2.6%.

关键词：集成特征选择；广义邻域粗集；马氏距离分布熵；选择性集成；模拟电路故障诊断
中图分类号：TP181 **文献标志码**：A **文章编号**：0253-987X(2011)06-0034-06

Ensemble Feature Selection Based on Generalized Neighborhood Rough Model and Its Selective Integration

MA Chao, CHEN Xihong, XU Yuliang, WANG Guangming
(Missile Institute, Air Force Engineering University, Sanyuan, Shaanxi 713800, China)

Abstract: A new ensemble feature selection method under the model of generalized neighborhood rough set is presented to improve the classification accuracy in actual pattern recognition systems. The importance degrees of sample features are evaluated by the distribution entropy of Mahalanobis distance (DEMD), and the generalized neighborhood rough model is constructed based on resulting degrees. Then a fast attribute reduction algorithm that takes features importance degrees as heuristic information is designed to produce multiple reduction results for training basic classifiers. More basic classifiers with higher diversity are obtained through changing parameter values in the model, and these basic classifiers are then selectively integrated using the principal component analysis method. The fault diagnosis results of an analog circuit show that the proposed method increases classification accuracy by at least 2.6% compared with other methods such as Adaboost.

Keywords: ensemble feature selection; generalized neighborhood roughset; distribution entropy of Mahalanobis distance; selective integration; analog circuit fault diagnosis

个体分类器的分类精度和个体分类器之间的差异性是影响集成学习器泛化能力的2个关键因素，其中不同个体分类器的差异性对提高集成学习器的分类精度以及泛化能力至关重要[1]. 实际模式识别

收稿日期：2010-11-04. 作者简介：马超(1983—)，男，博士生；陈西宏(联系人)，男，教授，博士生导师. 基金项目：国家自然科学基金资助项目(60971118).
网络出版时间：2011-03-11 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20110311.1506.001.html

系统中的原始样本数据往往是随机采集的,如轴承的振动信号、电机的电流信号、模拟电路的电压信号等. 为了获取更多的信息,通常情况下会按照一定的特征提取方法将特征维数设计得比较高,然后在众多特征中选择对分类有利的特征,因此初始特征之间会存在强相关或者大量冗余,故适合采用扰动样本特征空间的方式产生具有差异性的基分类器,其中特征子空间的选择,即集成特征选择是关键.

粗糙集理论的属性约简是以最小概率描述为准则来选择特征,每一个约简结果生成一个特征子空间,每一个特征子空间包含的属性是不同的,但是都可以完成对整个系统的描述,差别在于描述的角度和分类的能力不同,彼此具有一定差异性,因此利用多个约简结果训练的基分类器是准确而有差异的,满足了分类器集成的条件. 目前,应用粗糙集进行集成特征选择面临的主要问题有:①实际模式识别系统的特征值大多数情况下是连续值,利用经典粗集理论进行分析时先要进行离散化,而离散化会造成信息损失[2-3];②扩展粗集模型中常用的属性约简算法没有启发式信息引导,其效率往往不高[2-3];③不能保证不同约简结果训练的基分类器之间误差不相关[4]. 针对以上问题,本文在根据所提出的马氏距离分布熵对特征进行评估的基础上,通过广义邻域粗集模型和属性约简算法实现了集成特征选择,并利用主成分分析法进行选择性集成,最后通过模拟电路故障诊断实例对算法进行验证.

1　基于马氏距离分布熵的特征评估

针对连续值特征的特点,采用马氏距离计算样本之间的距离.

1.1　马氏距离分布熵的定义

定义 1　样本集中各点相对于空间某点的马氏距离直方图的熵为马氏距离分布熵(distribution entropy of mahalanobis distance,DEMD).

定义 2　某类样本集各点相对于该类样本集均值的马氏距离分布熵为自马氏距离分布熵(ADEMD),相对于异类样本集均值的马氏距离分布熵为互马氏距离分布熵(CDEMD).

定义 3　定义各类样本的互马氏距离分布熵之和与各类样本的自马氏距离分布熵之和的比值 $B=\sum_{i=1}^{D}C_i/\sum_{i=1}^{D}A_i$,其中 C_i 为各类样本的 CDEMD,A_i 为各类样本的 ADEMD,D 为类别数.

由熵的性质可知:当样本空间的样本点均匀分布时,马氏距离分布熵最大;当样本空间的样本点分布趋于集中时,马氏距离分布熵会变小. 所以,当 ADEMD 较小时,此类样本集在空间中聚集度高,而当 ADEMD 较大时,此类样本集在空间中聚集度低;当 CDEMD 较小时,样本间可分性差,当 CDEMD 较大时,样本间可分性好. 由熵的可加性原理可知,针对多类问题,B 越大,其样本的分类性能越好.

1.2　马氏距离分布熵的计算

马氏距离的计算公式为

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})=D_i^2 \tag{1}$$

式中:$\boldsymbol{x}$ 为任意样本;D_i^2 为样本 $\boldsymbol{x}$ 的马氏距离;$\boldsymbol{\mu}$ 为样本均值;$\boldsymbol{S}$ 为样本协方差矩阵,$\boldsymbol{S}^{-1}$ 为 $\boldsymbol{S}$ 的逆矩阵.

令 $D_{\max}^2=\max(D_i^2)$,$D_{\min}^2=\min(D_i^2)$,则对于任意样本 $\boldsymbol{x}_i$,其 $D_i^2\in[D_{\min}^2,D_{\max}^2]$. 设 $M(M\ll m,m$ 为样本个数)为常数,取 $\Delta D^2=(D_{\max}^2-D_{\min}^2)/M$,得到 M 个连续区间 $\Omega_i(i=1,2,\cdots,M)$. 在马氏距离 D_i^2 中,属于区间 Ω_i 的数目设为 $p_i(i=1,2,\cdots,M)$. 马氏距离 D_i^2 属于区间 Ω_i 的概率 $P_i=p_i/m(i=1,2,\cdots,M)$,由此可得马氏距离分布直方图. 对马氏距离统计直方图求熵,可得马氏距离分布熵 $E(P)=-\sum_{i=1}^{M}P_i\ln P_i$. 当 $\boldsymbol{\mu}$ 为同类样本均值时,即可按照上述过程求得自马氏距离分布熵,当 $\boldsymbol{\mu}$ 为异类样本均值时,可求得互马氏距离分布熵.

定义 4　B_i' 为特征 i 的重要度,$B_i'=B_i/\sum_{i=1}^{N}B_i$,其中 N 为样本特征个数.

2　广义邻域粗集模型及属性约简算法

近年来,许多学者为在应用粗糙集理论时避免离散化,引入了模糊粗糙集模型、相似关系粗糙集模型[5]和邻域关系模型[2-3]等. 邻域模型比较直观、易于理解,本文采用邻域关系模型来处理连续值样本特征. 胡清华等在文献[2]中对邻域粗糙集模型及其性质进行了详细的论述,计算以欧式距离为度量的邻域时,认为每个样本在空间中各个特征的重要性是相等的. 文献[3]认为仅从普通欧式距离的角度度量样本间的相似性,对分类方差不等,或呈非球形分布的样本来说,分类结果不会很好,并提出一种特征广义重要度对每一维特征进行加权来解决这个问题,但在计算特征广义重要度的时候只用到每类样本的边界样本,边界样本提供的信息显然不能反映类内样本的分布情况. 因此,本文利用马氏距离分布熵来估计特征的重要性,在文献[2]的基础上构建了

广义可变精度邻域粗集模型.

2.1 广义可变精度邻域粗集模型

定义5 基于马氏距离分布熵的广义欧式距离为

$$\Delta_B(x_1,x_2)=\left(\sum_{i=1}^{N}B_i'\mid f(x_1,a_i)-f(x_2,a_i)\mid^2\right)^{1/2} \tag{2}$$

式中:$\Delta_B(x_1,x_2)$为属性集合B下样本x_1、x_2之间的广义欧式距离.

定义6 给定实数空间上的非空有限集合$U=\{x_1,x_2,\cdots,x_n\}$,对于U上的任意对象x_i,以欧式距离作为实数空间上的度量,定义其δ广义邻域为

$$\delta_B(x_i)=\left\{x\Big/\left(\sum_{i=1}^{N}B_i'\mid f(x_i,a_i)-f(x,a_i)\mid^2\right)^{1/2}\leqslant\delta\right\} \tag{3}$$

式中:$x\in U$;$\delta\geqslant 0$;$\delta_B(x_i)$为属性集合B下x_i生成的δ广义邻域信息粒子.

定义7 给定一个广义邻域决策系统$T=\langle U,C,D\rangle$,U为样本集合,C为连续值条件属性,D为决策属性,D将U划分为d个等价类$X_1,X_2,\cdots,X_d$,$\forall A\subseteq C$,定义决策属性D关于A的下近似和上近似为

$$N_{A-}D=\bigcup_{i=1}^{N}N_{A-}X_i;\ N_A^-D=\bigcup_{i=1}^{N}N_A^-X_i \tag{4}$$

式中:$N_{A-}X=\{x_i/x_i\in U,\delta_B(x_i)\subseteq X\}$;$N_A^-X=\{x_i/x_i\in U,\delta_B(x_i)\cap X\neq\varnothing\}$.

由于实际工程应用中采集的数据不可避免地会存在一些噪声,在定义7关于下近似的定义中要求广义邻域信息粒子严格地被X所包含,不存在噪声容忍能力.为此,放松对严格包含关系的要求,定义程度包含关系,并对定义7进行改进.

定义8 设A、B是2个集合,定义A包含于B的程度为

$$I(A,B)=\mathrm{card}(A\cap B)/\mathrm{card}(A) \tag{5}$$

式中:$A,B\neq\varnothing$;card表示集合的势.

定义广义邻域模型可变精度阈值k的下近似和上近似分别为

$$N^k_{-}X=\{x_i/x_i\in U,I(\delta_B(x_i),X)\geqslant k\};\quad N^{k-}X=\{x_i/x_i\in U,I(\delta_B(x_i),X)\geqslant 1-k\} \tag{6}$$

定义9 给定广义邻域决策系统$T=\langle U,C,D\rangle$,如果$A\subseteq C$是C的一个约简,则A必须满足

$$\left.\begin{aligned}&\forall a\in A,P_{A-a}(D)<P_A(D)\\&P_A(D)=P_C(D)\end{aligned}\right\} \tag{7}$$

式中:$P_A(D)$为决策属性D关于集合A的下近似;$A-a$表示从集合A中删去任意属性a后剩余属性组成的集合.

定理1 给定欧式距离度量和邻域阈值δ,如果$B_1\subseteq B_2$,$x_i\in P_{B_1}(D)$,则$x_i\in P_{B_2}(D)$成立.

由定理1可知,对于任意$B\subseteq C$,$a\in(C-B)$,若样本$\boldsymbol{x}$为属性集合B上的正域样本,则$\boldsymbol{x}$也是属性集$(B+a)$上的正域样本,在计算$(B+a)$关于决策属性D的下近似时,只需判断原来负域中的样本即可,可以大大减少样本判断次数.

2.2 基于蚁群算法的属性约简算法

根据广义邻域粗集模型的特点,提出一种基于蚁群算法的快速属性约简算法,能够得到多个属性约简结果,而且以特征评估结果作为启发式信息,使得算法在初期就以较高概率选择重要度高的特征加入属性约简结果,并在计算样本邻域时,利用定理1的性质来减少样本判断次数,有效提高了属性约简算法的计算效率.

2.2.1 *基于特征重要度的启发式信息* 设$\tau_i(t)$为特征i的信息素,α为信息素重要度,$\eta_i(t)$为特征i的启发式信息,β为启发式信息重要度,W_n为蚂蚁n的可允许集合,l为可允许集合中的元素.蚁群算法按照如下转移概率公式寻找下一个特征

$$P_{ij}^{(n)}=\begin{cases}\tau_j^{\alpha}(t)\eta_j^{\beta}(t)\Big/\sum\limits_{l\in W_n}\tau_l^{\alpha}(t)\eta_l^{\beta}(t), & j\in W_n\\ 0, & \text{其他}\end{cases} \tag{8}$$

由式(8)可知,若无启发式信息,即$\beta=0$时,蚁群算法为纯粹正反馈启发式算法,算法初期相当于随机搜索导致算法收敛速度较慢.属性约简的目的是在分类能力不变的前提下,用最少特征代替全部特征完成对样本空间的分类,由定义4可知,特征重要度B_i'越高,特征对样本分类贡献越大,所以特征重要度高的特征应该优先被选入属性约简结果,以达到用最少特征实现全部特征所具有的分类能力的目的.因此,本文将特征重要度B_i'设为各特征的启发式信息$\eta_i(t)$,即令$\eta_i(t)=B_i'$,算法在初期就会以较大概率寻找特征重要度高的特征,以加速属性约简结果形成,从而提高了整个算法的效率.

2.2.2 *算法描述* 广义邻域粗集模型下基于蚁群算法的属性约简算法,通过输入决策表$T=\langle U,C,D\rangle$得到多个属性约简结果,其算法描述如图1所示.

算法说明:信息素更新策略为

$$\tau_j(t+T_j)=\rho\tau_j(t)+\Delta\tau_j \tag{9}$$

$$\Delta\tau_j=\sum_{n=1}^{m}\Delta\tau_j^n \tag{10}$$

$$\Delta\tau_j^n=\begin{cases}Q/|S|, & \text{若蚂蚁 } n \text{ 经过特征 } a_j\\ 0, & \text{其他}\end{cases} \tag{11}$$

式中:$|S|$表示第 n 只蚂蚁寻找到的属性约简含的特征个数;Q 为常数.

（1）初始化. 设迭代次数 $N_C=0$，最大迭代次数为 $N_{C,\max}$，蚂蚁个数 $m=N$（为条件属性个数），初始信息素均设为 τ_0；蚂蚁的禁忌表 $t_i(i=1,2,\cdots,m)$ 清零；
（2）将 m 只蚂蚁 $m_i(i=1,2,\cdots,m)$，随机分配给条件属性 a_i，计算每个蚂蚁的启发信息 $\eta_i(t)=B_i'$.
（3）当 $N_C\leqslant N_{C,\max}$ 时，对每只蚂蚁执行如下运算：
  $R=a_i$；待测样本集 $S=U$；$P_R(D)=\varnothing$；
  while $S=\varnothing$
  蚂蚁 m_i 按转移概率公式寻找下一个特征 a_j；
    $B=R\cup a_j$；$T_j=<U,B,D>$；
    for each $x_i\in S$
    计算 x_i 的邻域 $\delta_B(x_i)$；
    if $T/\mathrm{card}(\delta_B(x_i))\geqslant k//T$ 为决策属性取值相同的样本数
      $P_B(D)=P_R(D)\cup x_i$；
    end if
    end for
    if $P_B(D)<P_R(D)$
      $R=R\cup a_j$；
      $S=U$;
      $S=S-P_R(D)$；//下近似中样本下次不再计算
      $t_n=t_n\cup a_j$；
    else
      退出while循环；
    end if
record R；//记录蚂蚁寻找到的属性约简
$\tau_j(t+T_j)=\rho\tau_j(t)+\Delta\tau_j$；//信息素更新
$t_n=\varnothing$，将蚂蚁放回原来位置；
（4）record R_{lbest}；//记录每次迭代的最优属性约简 R_{lbest}；
（5）record R_{gbest}；//记录全局最优属性约简 R_{gbest}；
（6）如 R_{gbest} 的数量小于20，寻找与 R_{gbest} 所含属性数量最接近的 R_{lbest}，直到寻找到的属性约简数量等于20;
（7）输出寻找到的属性约简.

图 1　基于蚁群算法的属性约简算法描述

3　基于主成分分析法的选择性集成

通过实验发现,当邻域阈值 δ 和可变精度阈值 k 固定时,属性约简结果训练的基分类器分类精度差别不大,当改变参数 δ 和 k 时,属性约简结果训练的基分类器分类精度差别显著增大,这说明随着参数 δ、k 的改变,训练所得基分类器的差异性变大了.因此,本文在设置基分类器分类精度阈值的基础上,通过改变参数 δ、k 的值来增加基分类器之间的差异性,以达到提高集成分类器的分类精度的目的.

选择性集成可以消除基分类器之间的相关性,其思想是:在独立训练生成若干基分类器后,先剔除部分相关性大的基分类器,然后用剩下的基分类器进行集成,能够取得比全部基分类器集成更好的泛化性能[6-7].本文提出一种基于主成分分析法的选择性集成新算法,其思路是:每一个基分类器都可以看成是影响集成分类器性能的指标,利用主成分分析法将基分类器重新组合成一组新的相互无关的几个综合分类器(基分类器的线性组合)来代替原来基分类器,为避免集成学习器过拟合,可以根据实际情况从中选择较少的综合分类器尽可能多地反映原来所有基分类器能表达的信息,将选择出来的综合分类器按照一定方式进行集成,即可得到集成分类器.具体算法描述如图 2 所示.

（1）初始化：设邻域阈值 δ 的取值范围为[0, 1]，其初值为0，可变精度阈值 k 的取值范围为[0.5, 1]，其初值为0.5, 设 δ 和 k 的变化步长为 Δl，设基分类器精度阈值为 Q，综合分类器累计贡献阈值为 W，属性约简集合 $V=\varnothing$；
（2）随机选择样本 U 中1/2的样本作为训练集 U_1，剩余部分作为测试集 U_2；
（3）while $k\leqslant 1$
    while $\delta\leqslant 1$
      利用算法2求得属性约简集合 ΔV；
      $V=V\cup\Delta V$;
      $\delta=\delta+\Delta l$;
    end while
    $k=k+\Delta l$；
  end while
（4）将 U_1 中的样本按照集合 V 中属性约简结果进行变换后训练基分类器；
（5）将 U_2 中的样本按照集合 V 中属性约简结果进行变换后计算各个基分类器的分类精度,保留分类精度大于阈值 T 的基分类器组成集合 M，将集合 M 中基分类器的测试结果矩阵记为 $\boldsymbol{M}_{ij}=1$；
（6）求相关矩阵 $\boldsymbol{R}_M$ 及其特征值 λ 和特征向量 $\boldsymbol{L}$；
（7）求综合分类器 Z_k 的贡献率 $W_k=\lambda_k/\Sigma\lambda$，其中 $Z_k=L_{k1}m_1+L_{k2}m_2+L_{k3}m_3+\cdots+L_{kp}m_p$；
（8）求出累计贡献率大于 W 的前 k 个主成分并输出集成学习器 $Z=W_1Z_1+W_2Z_2+\cdots+W_kZ_p$.

图 2　基于主成分分析法的选择性集成算法描述

4　实例验证

选择 ITC'97 国际标准电路[8] 中的 CTSV (Continuous-Time State-Variable)滤波器来验证本文算法的可行性和有效性.电路如图 3 所示.元件标称值分别为 $R_1=R_2=R_3=R_4=R_5=10\ \mathrm{k\Omega}$，$R_6=3\ \mathrm{k\Omega}$，$R_7=7\ \mathrm{k\Omega}$，$C_1=C_2=20\ \mathrm{nF}$，各元件容差均为 5%.

考虑待测电路软故障:$F_1(R_1+50\%)$、$F_2(R_1-50\%)$、$F_3(C_1+50\%)$、$F_4(C_1-50\%)$、$F_5(R_1+$

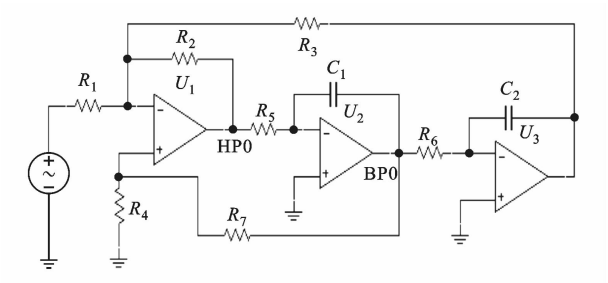

图 3 CTSV 电路图

50%，C_1 +50%)、F_6(R_1 −50%，C_1 −50%)，其中“+”或“−”代表元件偏离其标称值，再加上无故障状态共有 7 种状态模式. 电路仿真分析使用 Multisim7 平台.

提取电路幅频响应曲线上的离散点为故障特征，对位于区间[100，10 000] Hz 的响应电压值上的 20 个点进行采样作为原始故障特征值. 对各种故障模式和无故障模式进行 100 次 Monte Carlo 仿真，获得 700 个样本，每个样本含有 20 个属性，随机选取一半样本作为训练样本，剩余部分作为测试样本.

构造广义邻域决策表 $T=\langle U,C,D\rangle$，其中 U 为模拟电路故障采集样本集合，C 为条件属性集合，由频率响应曲线上各采样点组成，D 为决策属性，即待测电路各故障类别. 利用属性约简算法对所构造决策表进行属性约简，参数设置如下：最大迭代次数 $N_{C,\max}=200$，蚂蚁个数 $m=20$，各蚂蚁初始信息素均设为 $\tau_0=0.1$，启发信息 $\eta_i=B'_i$，信息素重要度 $\alpha=1.5$，启发信息重要度 $\beta=1.9$，挥发系数 $\rho=0.8$，邻域阈值 $\delta=0.4$，可变精度阈值 $k=0.9$. 共有 20 个属性约简结果.

将有启发式信息和无启发式信息的属性约简算法进行比较，收敛曲线如图 4 所示.

由图 4 可知，两种方式都可以找到最优解(包含 5 个特征)，比较而言，当有启发式信息时，算法运行初期收敛速度比较快，迭代 58 次即可找到最优解，运行时间为 3.1 s；没有启发式信息引导时，算法需要迭代 95 次才能找到最优解，运行时间为 4.9 s. 没有启发式信息引导，蚁群算法在计算初期类似于随机搜索，所以速度比较慢.

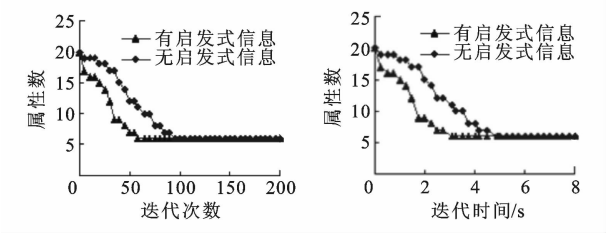

图 4 算法收敛曲线比较

以 SVM 为基分类器，图 5 显示了训练基分类器的平均分类精度随邻域阈值和可变精度阈值的变化情况，其中，δ 的取值以 0.05 为步长从 0 变化到 1，k 的取值以 0.05 为步长从 0.5 变化到 1. 由图 5 可知，平均分类精度从小于 70%变化到大于 90%，说明随着 δ 和 k 的变化，基分类器之间的差异性会变大，有利于集成分类器分类精度的提高.

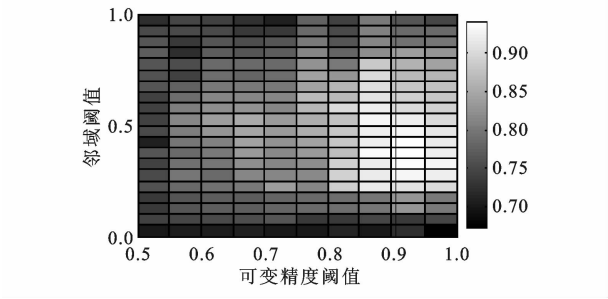

图 5 平均分类精度随阈值变化的情况

设置基分类器精度阈值 $T=80\%$，综合分类器累计贡献阈值 $W=95\%$，利用主成分分析法对满足属性约简结果的基分类器进行选择性集成，得到最终的集成分类器. 比较下面各集成分类器：满足精度阈值的所有基分类器进行集成得到集成分类器 A；具有最高平均分类精度(当 $\delta=0.45$，$k=0.85$ 时获得最高平均分类精度 94.25%)的属性约简集合进行集成，得到集成分类器 B；基于模糊 C 均值离散化和克隆选择算法产生的属性约简集合进行集成，得到集成分类器[9] C；基于 AdaBoost 的神经网络集成学习器 D[10]；对满足精度阈值的基分类器经主成分分析法选择性集成得到集成分类器 E. 比较结果如表 1 所示.

表 1 中除算法分类器 D 外，其余算法的基分类器均为 SVM，前 4 种集成分类器的集成算法均为多数投票法.

表 1 集成分类器性能比较

分类器	分类精度%	最高精度%	平均精度%	差异性
A	97.14	95	90.75	0.475
B	95.14	94	92.85	0.164
C	93.16	91.14	89	0.178
D	90.28	88.57	86.84	0.103
E	99.71			

注：最高精度指组成集成分类器的所有基分类器分类精度的最高值；平均精度指组成集成分类器的所有基分类器分类精度的平均值；差异性指组成集成分类器的基分类器之间的差异性，由不一致度进行度量.

5　实验结果分析

(1)由集成分类器 A、C 的比较可知，离散化后获得的基分类器无论是最高精度、平均精度还是集成后的分类精度都略微降低，说明离散化造成了信息的损失.

(2)集成分类器 A 在平均精度方面小于 B，但 A 具有较明显的差异性优势，取得了更高的分类精度，相比较，差异性方面，集成分类器 B 略小于 C，但由于 B 在平均精度方面明显优于 C，所以取得了更高的分类精度. 上述现象说明两点：差异性在提高集成分类精度的过程中起关键作用；在差异性相当时，较高的平均精度也能有效提高集成的分类精度.

(3)集成分类器 E 的分类精度明显高于 A，说明基于主成分分析法的选择性集成能有效提高集成分类器的分类性能.

(4)集成分类器 E 的分类精度明显高于 D，原因有两点：①样本个数比较少时(绝大多数实际模式识别系统都是如此)，AdaBoost 算法不能使基分类器之间具有较大差异；②神经网络以经验风险最小化为原则，导致在小样本上训练的基分类器产生过学习，使其泛化能力的提高受到了限制，所以扰动特征空间的方式比较适合实际模式识别系统.

6　结　论

随着粗糙集理论的发展，越来越多的研究者将其应用于模式识别领域，基于粗糙集理论的集成特征选择即是典型代表. 针对粗糙集理论应用于集成特征选择所面临的问题，本文在给出马氏距离分布熵概念及其计算方法的基础上，设计了基于广义可变精度邻域粗集模型和主成分分析法的选择性集成算法，实例结果验证了所设计选择性集成算法是有效和可行的，实验结果同时也表明基分类器差异性和平均分类精度与集成分类器性能的确切关系是值得研究的课题. 此外，如何能快速有效地设置邻域阈值和可变精度阈值，以寻找使基分类器满足分类精度要求的属性约简集合，是下一步研究的方向.

参考文献：

[1] DIETTERICH T G. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization [J]. Machine Learning, 2000,40(2): 139-158.

[2] 胡清华，于达仁，谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报，2008,19(3): 640-649.
HU Qinghua, YU Daren, XIE Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation [J]. Journal of Software, 2008,19(3): 640-649.

[3] 肖迪，胡寿松. 实域粗糙集理论及属性约简[J]. 自动化学报，2007，33(3):253-258.
XIAO Di, HU Shousong. Real rough set theory and attribute reduction [J]. Acta Automatica Sinica, 2007, 33(3):253-258.

[4] 张东波，王耀南. 基于粗糙集约简的神经网络集成及其遥感图像分类应用[J]. 中国图象图形学报，2008，13(3):481-487.
ZHANG Dongbo, WANG Yaonan. Neural network ensemble based on rough sets reduction and its application to remote sensing image classification [J]. Journal of Image and Graphics, 2008, 13(3):481-487.

[5] 冯林，王国胤，李天瑞. 连续值属性决策表中的知识获取方法[J]. 电子学报，2009,37(11):2432-2438.
FENG Lin, WANG Guoyin, LI Tianrui. Knowledge acquisition from decision tables containing continuous-valued attributes [J]. Acta Electronica Sinica, 2009, 37(11):2432-2438.

[6] ZHOU Zhihua, WU Jianxin, TANG Wei. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence, 2002, 137(1/2): 239-263.

[7] 唐耀华，高静怀，包乾宗. 一种新的选择性支持向量机集成算法[J]. 西安交通大学学报，2008，42(10):35-39.
TANG Yaohua, GAO Jinghuai, BAO Qianzong. Novel selective support vector machine ensemble learning algorithm [J]. Journal of Xi′an Jiaotong University, 2008, 42(10):35-39.

[8] KAMINSKA B, ARI K, BELL I, et al. Analog and mixed-signal benchmark circuits: first release [C]// Proceedings of the 1997 IEEE International Conference on Test. Piscataway, NJ, USA: IEEE, 1997: 183-190.

[9] 梁霖，徐光华. 基于克隆选择的粗糙集属性约简算法[J]. 西安交通大学学报，2005,39(11):1231-1234.
LIANG Lin, XU Guanghua. Reduction of rough set attribute based on immune clone selection [J]. Journal of Xi′an Jiaotong University, 2005,39(11):1231-1234.

[10] 刘红，陈光禹，宋国明，等. 基于 AdaBoost 集成网络的模拟电路单软故障诊断[J]. 仪器仪表学报，2010，31(4):851-856.
LIU Hong, CHEN Guangyu, SONG Guoming. Adaboost based ensemble of neural networks in analog circuit fault diagnosis [J]. Chinese Journal of Scientific Instrument, 2010,31(4):851-856.

(编辑　刘杨)