# 第5章

# 高维数据实体分辨多分类器方法

## 5.1 引言

在高维数据研究中，使用特征选择降低不相关特征、冗余特征和噪声特征对算法性能的影响是常用的方法。通常，特征选择是从特征中选择满足算法性能要求的一个特征子集。特征子集的基数一般较小，即使在高维数据中，特征子集的基数也显著小于数据的维度。这是由于增加特征个数会提高运算开销，同时也会引入不相关特征和噪声特征，导致算法性能下降。然而与维度较低的数据相比，高维数据的特征中蕴含了更多的信息，如果仅使用一个基数较小的特征子集，会导致算法无法有效利用高维特征中的信息。而当前特征选择方法的局限性同样存在于高维数据的实体分辨问题中。

集成学习是一种独立于具体算法的非参数机器学习方法，它的思想是将一些性能较弱的分类器（至少优于随机分类器）组合形成一个分类性能较强的多分类器系统，系统的输出由各个弱分类器的预测值共同决定。近年来，集成学习的研究如火如荼，研究人员已经从理论和实践上证明了集成学习能够有效提升算法性能。

本章采用集成学习思想，设计针对高维数据实体分辨的多分类器方法，提升高维数据实体分辨性能。

## 5.2 分类器度量

### 5.2.1 分类器性能度量

通常，二分类器从功能上可以看成一个将状态类别为 $N$ 的样本空间（源空间）映射到 2 类空间（目标空间）的映射函数，如图 5-1 所示。

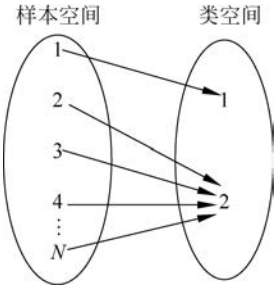


图 5-1　二分类器映射图

本章将 1 类(实体分辨中的匹配类)作为正类,2 类(实体分辨中的不匹配类)作为负类,在一般的分类器中对应为 1 类和－1 类(或 1 类和 0 类)。通常使用分类器的性能指标评估特征选择算法的效果,本节基于曹建军定义的分类器性能度量指标,定义针对高维数据实体分辨的二分类器性能度量指标[1]。

分类正确率(Classification Success Rate)的定义如式(5-1)

$$P=\frac{\text{正确分辨的样本数}}{\text{参加分辨的样本总数}}\times 100\% \tag{5-1}$$

虚警率(False Alarm Rate)的定义如式(5-2)

$$R_{fa}=\frac{\text{不匹配误分为匹配的样本数}}{\text{参加分辨的不匹配样本总数}}\times 100\% \tag{5-2}$$

漏诊率(Fault not Be Recognized Rate)的定义如式(5-3)

$$R_{fn}=\frac{\text{匹配误分为不匹配的样本数}}{\text{参加分辨的匹配样本总数}}\times 100\% \tag{5-3}$$

定义二类分类器输出的分布矩阵如式(5-4)所示

$$p=[p_{ii'}],i,i'=1,2 \tag{5-4}$$

式(5-4)中,$p_{ii'}$的定义如式(5-5)

$$p_{ii'}=\frac{\text{第 } i \text{ 类被分为第 } i' \text{ 类的样本数}}{\text{参加分辨的第 } i \text{ 类样本数}}\times 100\% \tag{5-5}$$

$p_{ii}(i=1,2)$为第 $i$ 类样本的分类正确率,可由式(5-6)计算得出

$$p_{ii}=1-p_{ii'} \tag{5-6}$$

式(5-6)中,$i\neq i'$。

分类正确率 $P$ 可通过式(5-7)计算得出

$$P=P_i p_{ii}+P_{i'}p_{i'i'} \tag{5-7}$$

式(5-7)中 $i\neq i'$,$P_i$ 为第 $i$ 类样本的先验概率,给定测试样本集,$P_i$ 可由式(5-8)计算

$$P_i=\frac{N_i}{N_i+N_{i'}} \tag{5-8}$$

式(5-8)中,$N_i$ 为第 $i$ 类的样本数,$N_{i'}$为第 $i'$类的样本数,$P_{i'}$的定义与 $P_i$ 相同。

虚警率 $R_{fa}$ 可通过式(5-9)计算

$$R_{fa}=p_{21} \tag{5-9}$$

漏诊率 $R_{fn}$ 可通过式(5-10)计算

$$R_{fn}=p_{12} \tag{5-10}$$

$P$,$R_{fa}$ 和 $R_{fn}$ 之间存在式(5-11)的关系

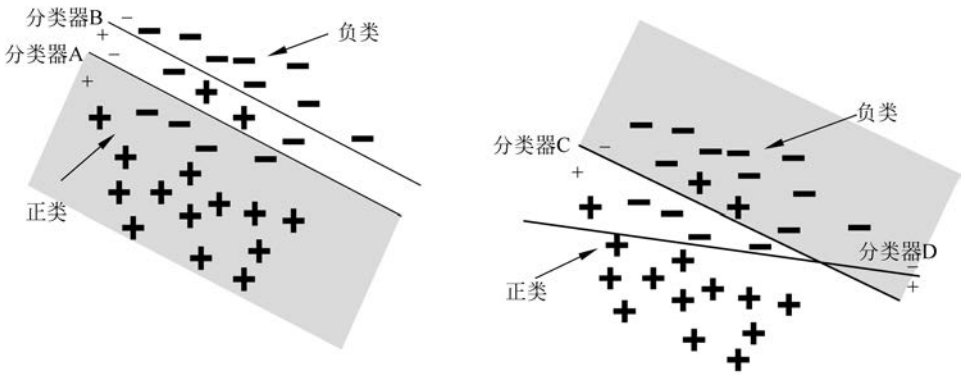
$$1-P=(1-P_1)R_{fa}+P_1R_{fn} \tag{5-11}$$

根据式(5-11)以及虚警率和漏诊率的定义可以看出,虚警率与漏诊率是一对互相矛盾的指标,虚警率高时漏诊率较低,反之亦然。而分类正确率能够综合反映虚警率与漏诊率,因此采用分类正确率更能有效判定分类器的分类能力。

## 5.2.2　分类器相似性度量

存在多个分类器时(多分类系统),输出结果相似的分类器对难分数据可能具有相同的

预测值，因此它们的组合并不能有效提高分类性能。反之，分类器输出相似度具有一定的差异度时（多样性），它们的组合能够在一定程度上提高分类性能。下面用示例说明该情况，如图 5-2 所示。




(a) 输出相似分类器示意图　　(b) 输出不相似分类器示意图

图 5-2　不同相似度分类器输出示意图

在图 5-2(a)中，分类器 A 与分类器 B 具有相似的分类超平面，即具有相似的预测性能，因此它们的组合无法对阴影中的 4 个负类进行正确的判别；图 5-2(b)中分类器 C 与分类器 D 的分类超平面与(a)相比具有更大的差异性，即增加了输出的差异性，它们的组合仅无法正确判别阴影中的 2 个正类样本，从而提升了组合分类器的整体分类性能，这种具有输出不相似性的分类器在分类效果上是互补的，即具有较强的多样性。

给定训练样本集，以及功能类型和参数设置相同的二类分类器，对于确定的特征子集 St（即给定特征向量），通过特征向量构建新的训练样本并训练分类器，然后用测试样本评估分类器的分类性能，可以将特征子集 St 映射为一个确定的分类器 $\Lambda_{\mathrm{St}}$ 和一个分类器输出分布矩阵 $\boldsymbol{p}$，如式(5-12)所示

$$\Lambda(\mathrm{St})=(\Lambda_{\mathrm{St}},\boldsymbol{p}) \tag{5-12}$$

因此，在分类输出结果上互补的分类器是通过具有互补性的特征子集训练得到的，即互补特征子集。而分类器 $\Lambda_{\mathrm{St}}$ 的相似性可以由特征子集 St 的相似度和输出分布矩阵 $\boldsymbol{p}$ 的相似度来衡量，我们将其分别称为输入相似性和输出相似性。

下面定义二类分类器间的相似性度量指标（类似于集成学习分类器的多样性）。

**定义 5-1**　（**输入相似性**）定义分类器的输入相似性为输入特征子集的相似度，对两个分类器的非空特征子集 $\mathrm{St}_A$ 与 $\mathrm{St}_B$，采用谷元距离度量它们的相似度[2]，如式(5-13)

$$S_d(\mathrm{St}_A,\mathrm{St}_B)=1-\frac{|\mathrm{St}_A|+|\mathrm{St}_B|-2|\mathrm{St}_A\cap\mathrm{St}_B|}{|\mathrm{St}_A|+|\mathrm{St}_B|-|\mathrm{St}_A\cap\mathrm{St}_B|} \tag{5-13}$$

由式(5-13)可知 $S_d\in[0,1]$，当 $S_d=0$ 时，表示两个特征子集之间没有相同的组成元素；当 $S_d=1$ 时，表示两个特征子集完全相同，即具有完全一致的组成元素，使用它们构成的样本进行训练得出的分类器也完全相同。因此，$S_d$ 越大两个特征子集的相似度越高，即分类器的输入相似性越高。

**定义 5-2**　（**输出相似性**）定义分类器的输出相似性为分类器输出分布矩阵的相似度，即对 $p'=[p'_{ii'}]$，$p''=[p''_{ii'}]$，其中 $i=1,2$，$i'=1,2$，用标准化的皮尔逊相关系数度量它们之

间的相似度，如式(5-14)：

$$S_c(p',p'')=\frac{1}{2}\left(1+\frac{\sum_{i=1}^{2}\sum_{i'=1}^{2}(p'_{ii'}-\overline{p'})(p''_{ii'}-\overline{p''})}{\sqrt{\sum_{i=1}^{2}\sum_{i'=1}^{2}(p'_{ii'}-\overline{p'})^2\sum_{i=1}^{2}\sum_{i'=1}^{2}(p''_{ii'}-\overline{p''})^2}}\right) \tag{5-14}$$

式(5-14)中，$\overline{p'},\overline{p''}$ 分别为输出分布矩阵 $p'p''$ 中元素的平均值，即 $\overline{p'}=\frac{1}{4}\sum_{i=1}^{2}\sum_{i'=1}^{2}p'_{ii'}$。由式(5-14)可知 $S_c\in[0,1]$，当 $S_c=0$ 时，表示 $p',p''$ 完全负相关，即两个分类器的输出分布矩阵完全相反，分类结果完全不同，两个分类器的相似性最弱；当 $S_c=1$ 时，表示 $p',p''$ 完全正相关，即两个分类器的输出分布矩阵完全相同，分类结果完全一致。

下面给出分类器的输出相似性与输入相似性关系的定理，并进行证明[1]。

**定理 5-1**　假设 $\Lambda(\mathrm{St}_A)=(\Lambda_{\mathrm{St}_A},p_A)$ 且 $\Lambda(\mathrm{St}_B)=(\Lambda_{\mathrm{St}_B},p_B)$，如果 $S_c(p_A,p_B)<1$，则有 $S_d(\mathrm{St}_A,\mathrm{St}_B)<1$。

**证明**：假设 $\mathrm{St}_A=\mathrm{St}_B$，由公式(5-12)和定理的前提条件可得 $S_c(p_A,p_B)=1$，即若 $S_d=(\mathrm{St}_A,\mathrm{St}_B)=1$，那么 $S_c(p_A,p_B)=1$，该命题成立，同时该命题的逆否命题成立，定理得证。

由定理 5-1 可以得出结论，分类器的输出相似性要强于输入相似性，因此可以通过分类器的输出相似性来衡量分类器间的相似程度。

通过本节叙述可知，使用两个特征子集训练的分类器之间的相似性越低，分类器的多样性越好，它们的互补性越强，由这些分类器组成的多分类器系统能够更好地利用高维特征蕴含的信息，从而进一步提升分类性能。

## 5.3　基于特征选择的多分类器方法

根据“没有免费的午餐”定律，不存在单个分类算法能够适用于所有的问题，这是由于算法中的分类器是运行在特定环境中，其对类别的判定能力与具体问题相关。而集成学习是一种与算法无关的提升分类性能的方法，通过集成学习将性能较弱的分类器(至少优于随机分类器)进行组合，能够形成具有更好分类性能的多分类器方法。本节基于集成学习的思想，针对高维数据实体分辨，设计一种具有较强分类性能的多分类器方法，并基于单目标蚁群算法和多目标蚁群算法实现该方法。

### 5.3.1　系统模型设计

针对高维数据实体分辨问题，提出基于特征选择的的多分类器方法(基分类器为二分类器)(Multiple Classifier System Based on Feature Selection，MCSBFS)，其设计结构模型描述如下：

对于含有 $M$ 个($M$ 为奇数)分类器的多分类器方法，记 $P_m$ 为第 $m$($m\geqslant 2$)个分类器的分类正确率，$q_m$ 为第 $m$ 个分类器输入特征子集的基数，采用以下目标函数构造第 $m$ 个分类

器：

$$\max P_m \tag{5-15}$$

$$\max\left[1-\max_{j=1}^{m-1} S_c(p_j,p_m)\right] \tag{5-16}$$

$$\min q_m \tag{5-17}$$

式(5-15)表明希望所设计的第 $m$ 个分类器，具有最优的分类正确率；式(5-16)比较当前分类器与前 $m-1$ 个分类器的相似性，选择使得第 $m$ 个分类器与其他 $m-1$ 个分类器之间具有最大不相似性的特征子集，即互补特征子集，从而最大化分类器的多样性；式(5-17)说明希望选择基数最小的特征子集。从目标函数的定义可以看出，3 个目标之间互相冲突，因此该模型是一个多目标优化问题。

由于多分类器方法是由多个分类器构成，因此需要对分类器的输出结果进行集成，形成最终的分类预测值。MCSBFS 使用多数投票表决实现分类结果的集成，定义 $f_n^m$ 为第 $n$ 个样本在第 $m$ 个二类分类器上的函数决策值，它的预测值表示如式(5-18)所示

$$f_n^m=\begin{cases}1, & \text{匹配}\\ 2, & \text{不匹配}\end{cases} \tag{5-18}$$

则对于基分类器个数为 $M$ 个($M$ 为奇数)的多分类器系统，第 $n$ 个样本的集成输出结果如式(5-19)所示

$$\text{class}_n=f_n^1\oplus f_n^2\oplus\cdots\oplus f_n^M \tag{5-19}$$

式(5-19)中，$\oplus$表示异或运算，即将多数基分类器的输出作为多分类器系统的分类预测值。

## 5.3.2 方法实现

以上给出了 MCSBFS 方法的模型设计，本节给出基于蚁群算法的方法模型实现。由于 MCSBFS 方法模型是一个多目标优化问题，因此可以采用两种方式求解实现：一种是将多目标优化问题转换为单目标优化问题；另一种是直接作为多目标优化问题求解。针对这两种求解方式，设计两种基于蚁群算法的实现方法。蚁群算法的相关介绍见第 3 章，这里不再赘述。

**1. 单目标蚁群算法实现**

首先给出 MCSBFS 方法的单目标蚁群算法实现，在给定分类器时，对 MCSBFS 方法模型做如下分析：

(1) 在特征子集基数的确定前提下，将式(5-15)与式(5-16)进行加权聚合，转化为单目标优化函数，如式(5-20)所示

$$\max\{r_1P_m+r_2[\max_{j=1}^{m-1} S_c(p_j,p_m)]\} \tag{5-20}$$

其中，$r_1$ 与 $r_2$ 是聚合参数，且有 $r_1>0$，$r_2>0$，$r_1+r_2=1$。在设计第 $m$ 个分类器的过程中，通过聚合参数控制当前分类器的分类性能与分类器多样性之间的平衡，因此求解式(5-20)的关键在于确定聚合参数的值。由于求解第一个分类器模型时，目标函数式(5-16)并不存在，因此可以直接使用式(5-15)评估特征子集。

(2) 式(5-17)的求解关键在于确定分类器的特征子集基数。由于蚁群算法选择特征时需要事先给定特征的个数，另一方面，当特征个数在 1～15 时，分类器能够在分类性能与运算效率之间达到较好平衡。不失一般性，将特征子集基数 $q_m$ 的取值范围限定在[1,20]，确保不丢失边缘解且不会造成高昂的时间开销。

(3) 算法在当前的特征子集基数条件下迭代完成后，需要为当前分类器选择较好的特征子集。设定转换后的优化目标式(5-20)的优先级高于式(5-17)，当特征子集基数不同时，优先选择使得目标式(5-20)的值较大的特征子集；当两个特征子集的评估值相等时，选择特征基数较小的特征子集。需要特别说明的是，求解第一个分类器模型时，由于目标函数式(5-16)并不存在，因此当特征子集基数不同时，优先选择使得目标式(5-15)的值较大的特征子集，当两个特征子集的评估值相等时，选择基数较小的特征子集。

综上，基于单目标蚁群算法的 MCSBFS 方法模型实现伪代码描述如表 5-1 所示。

**表 5-1　分类器的单目标蚁群算法求解**

**输入**：聚合参数值 $r_1$ 与 $r_2$，信息素重要程度值 $\alpha$，启发式信息重要程度值 $\beta$，蚂蚁个数 $N$，蚁群算法最大迭代次数 iter
**输出**：特征子集 St

```
1.    for 1≤q_m≤20 do
2.        初始化蚁群算法信息素矩阵、启发式信息;
3.            while ite(当前迭代次数)<iter do
4.            for 1≤ant≤N do
5.                蚂蚁搜索特征子集;
6.                end for
7.                按照分析(1)选择较好的特征子集并更新信息素矩阵;
8.            end while
9.        按照分析(3)更新较好特征子集 St;
10.   end for
```

表 5-1 中第 1 行表示当前分类器的特征子集基数，第 2～9 行是在当前特征子集基数确定的条件下，通过单目标蚁群算法搜索满足优化目标的特征子集，第 3～8 行是蚁群算法搜索过程，第 4～6 行是在一次循环中，蚂蚁搜索特征子集的过程，第 7 行是根据分析(1)选择较好特征子集并更新信息素矩阵，第 9 行是根据分析(3)更新当前分类器的特征子集。现对多分类器方法的单目标蚁群算法求解作复杂度分析：设分类器个数为 M，数据特征的规模为 $C$，每只蚂蚁搜索特征子集的时间为 $O(C^2)$，因此蚁群算法的运行时间为 $O(\text{iter}\times N\times C^2)$，多分类器方法的求解运算时间为 $O(M\times|q_m|\times\text{iter}\times N\times C^2)$。

### 2. 多目标蚁群算法实现

本节给出 MCSBFS 方法的多目标蚁群算法实现，在给定分类器时，对 MCSBFS 方法的模型分析如下。

(1) 在特征子集基数确定的前提下，将式(5-15)与式(5-16)作为多目标优化问题的两个优化目标进行考虑，如式(5-21)所示

$$\max F=(f_1,f_2)^{\mathrm{T}} \tag{5-21}$$

其中，$f_1$ 为目标式(5-15)，$f_2$ 为目标式(5-16)。多目标优化问题的解为帕累托解，帕累托解的个数通常不唯一，因此算法开始前需要设置帕累托档案并指定档案规模，用帕累托档案记

录帕累托解，在算法迭代过程中需要基于帕累托支配关系使用新生成的解更新帕累托档案。与5.3.2.1节分析(1)相似，当求解第一个分类器模型时，目标函数式(5-16)并不存在，因此直接通过式(5-15)评估特征子集并进行优劣比较，此时帕累托档案不存在，仅有一个最优解。

(2) 对式(5-17)的分析与5.3.2.1节分析(2)一致，这里不再赘述。

(3) 当完成指定基数条件下特征子集的搜索后，为当前分类器选择特征子集的过程如下：比较帕累托档案中所有解在目标函数 $f_1$ 上的评估值，选择评估值最大的帕累托解(特征子集)；若存在多个使得目标函数 $f_1$ 取值最大的帕累托解，比较它们在目标函数 $f_2$ 上的评估值，选择评估值最大的帕累托解，若当前分类器为第一个分类器，则跳过此步；否则，比较它们的特征基数，选择特征基数最小的帕累托解。

这里采用第2章提出的基于等效路径更新的两档案多目标蚁群优化作为多目标蚁群算法的具体实现，在给定分类器的前提下，基于多目标蚁群算法实现的MCSBFS方法模型求解伪代码如表5-2所示。

**表5-2 分类器的多目标蚁群算法求解**

**输入**：帕累托档案规模 $Np$，信息素重要程度值 $\alpha$，启发式信息重要程度值 $\beta$，蚂蚁个数 $N$，蚁群算法最大迭代次数 iter；
**输出**：特征子集 St

```
1.    for 1≤q_m≤20 do
2.        初始化多蚁群算法信息素矩阵、启发式信息以及帕累托档案;
3.        while ite<iter do
4.            for 1≤ant≤N do
5.                蚂蚁搜索特征子集;
6.            end for
7.                按照分析(1)更新帕累托档案,同时更新信息素矩阵;
8.        end while
9.        按照分析(3)更新较好特征子集 St;
10.   end for
```

与单目标蚁群算法求解实现类似，表5-2中第1行表示当前分类器的特征子集基数，第2～9行是在当前特征子集基数确定的条件下，搜索满足优化目标的特征子集。现对多分类器方法的多目标蚁群算法求解作复杂度分析：设分类器个数为 $M$，数据特征的规模为 $C$，每只蚂蚁搜索特征子集的时间为 $O(C^2)$，更新帕累托档案的时间为 $O(Np^2)$，因此蚁群算法的运行时间为 $O(\text{iter}\times N\times Np^2\times C^2)$，多分类器方法的求解运算时间为 $O(M\times|q_m|\times\text{iter}\times N\times Np^2\times C^2)$。

## 5.4 实验与分析

本节验证评估MCSBFS方法，首先给出实验数据与对比方法，然后进行测试验证。

### 5.4.1 实验设置与对比方法

由于实体分辨问题与分类问题在数学模型上是一致的，且论文将实体分辨作为二分类

问题，因此仅需使用二分类数据验证实体分辨方法即可。实验采用 10 个标准测试数据集，其中 4 个数据集是第 3 章使用的标准测试数据集，其他 6 个数据集来源于 UCI 网站，10 个数据集的相关信息如表 5-3 所示。

**表 5-3　实验数据集信息**

| 数据集 | 实例规模 | 特征个数 | 特征选择范围 |
|---|---|---|---|
| BASEHOCK | 1993 | 4862 | [1,20] |
| PCMAC | 1943 | 3289 | [1,20] |
| RELATHE | 1427 | 4322 | [1,20] |
| MADELON | 2600 | 500 | [1,20] |
| IONOSPHERE | 351 | 35 | [1,20] |
| PRDS | 182 | 13 | [1,11] |
| SONAR | 208 | 61 | [1,20] |
| STATLOGHEART | 270 | 14 | [1,12] |
| CLIMATE | 540 | 18 | [1,16] |
| Z-ALIZADEH | 303 | 56 | [1,20] |

表 5-3 中特征选择范围表示特征子集基数的限定范围，为了防止丢失边缘解，特征个数的下限设置为 1，若原始特征维度大于 20，则上限设定为 20，若小于 20，则上限设定为特征维度减 2(特征中包含类标)。

由于 MCSBFS 方法中包含特征选择和集成学习的思想和技术，为了全面验证方法的有效性和优越性，从特征选择和集成学习两个方面选择 5 个具有代表性的对比算法，即基于蚁群优化的特征选择(Feature Selection Based on Ant Colony Optimization，FSBACO)[3]、信息增益(IG)和 ReliefF、以及集成学习方法 AdaBoost 和随机森林(Random Forest，RF)。FSBACO 方法以查准率和查全率作为特征选择的优化目标，通过加权聚合将其转换为单目标优化问题并采用蚁群算法求解，具有优异的性能表现；IG 和 ReliefF 是常用的特征选择方法，尽管它们独立于分类器选择特征，但具有较强的搜索性能，在实际应用中得到了广泛使用；AdaBoost 和 RF 是具有代表性的两种集成学习算法，AdaBoost 是一种迭代式的集成学习方法，通过调整样本权重逐个训练基分类器，RF 则将特征选择与 Bootstrap 结合，具有较强的鲁棒性和分类性能。

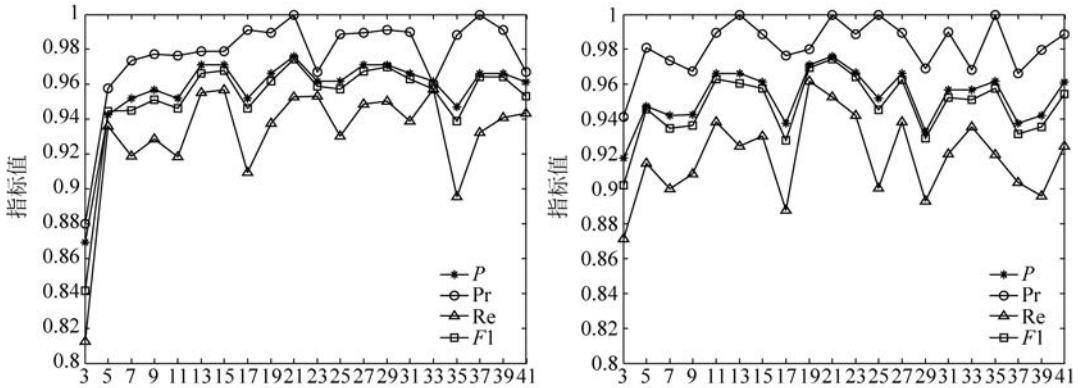
使用分类正确率 $P$、查准率(Precision)Pr、查全率(Recall)Re 和 $F1$ 作为实验的测量指标。

## 5.4.2　实验验证与结果分析

由于 MCSBFS 使用多目标蚁群算法和单目标蚁群算法两种方式实现，因此需要对这两种实现方法的性能作比较分析，将采用多目标蚁群算法实现的方法作为 Method-1，采用单目标蚁群算法实现的方法作为 Method-2。

首先分析 Method-1 和 Method-2 方法中基分类器个数的参数敏感性。Method-1 的参数设置如下：$\alpha=1$，$\beta=2$，$Q=0.2$，信息素初始浓度 $\tau(0)=100$，迭代次数为 80，帕累托档案规模为 40。除了帕累托档案规模，Method-2 的参数设置与 Method-1 相同，为了使得基分

类器具有更好的分类性能，设置 Method-2 中的聚合参数 $r_1=0.6$，$r_2=0.4$。以 SONAR 为测试数据集，采用支持向量机作为分类器，选择径向基核函数，宽度设置为 0.4，平衡参数为 100，将 20 轮 5 重交叉检验的均值作为输出，在基分类器个数从 3 增加到 41 的情况下，两种方法指标值的变化趋势如图 5-3 所示。




(a) Method-1基分类器敏感性分析　　(b) Method-2基分类器敏感性分析

图 5-3　基分类器参数敏感性分析

首先观察 Method-1 的参数敏感性分析图。可以看出，4 个指标值总体呈现先上升后下降的趋势。在基分类器个数为 3 的时候，4 个指标的数值最低，随着基分类器个数的增加，虽然指标值有一定幅度的波动，但其趋势是在不断上升，当基分类器个数达到 21 时，除查全率 Re，其他 3 个指标值达到最好。尽管在基分类器个数达到 33 时，查全率 Re 才达到最好，但与分类器为 21 时的 Re 值相比，并未有显著提升。因此当基分类器个数为 21 时，Method-1 方法的综合性能最好。

观察 Method-2 的参数敏感性分析图。与 Method-1 方法相比，Method-2 相关指标的参数敏感性变化趋势并不显著，但可以看出，当基分类器个数同样为 21 时，除查全率 Re 指标外，其他 3 个指标值达到最好。造成两种方法的参数取值一致的原因可能是由于 Method-1 与 Method-2 方法的求解模型一致，实现方式上的差异对多分类器方法的影响要弱于基分类器个数对多分类器方法的影响。

下面对 MCSBFS 方法的性能作比较分析，Method-1 和 Method-2 的参数设置不变，基分类器个数设定为为 21；FSBACO 的参数设置与原文相同；IG 和 ReliefF 方法使用特征选择范围内的最好结果作为输出；AdaBoost 与 RF 方法使用决策树作为分类器，基分类器个数设定为 100。使用 20 轮 5 重交叉检验的均值作为输出，分类器仍然使用支持向量机（参数不变），6 种方法在 4 个指标上的测试结果如表 5-4～表 5-13 所示。

**表 5-4　BASEHOCK 数据集实验结果**

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| Method-1 | 91.97 | 86.56 | 99.39 | 0.9252 |
| Method-2 | 82.09 | 73.73 | **99.70** | 0.8475 |
| FSBACO | 91.12 | 85.32 | 99.31 | 0.9178 |

续表

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| IG | 90.82 | 84.98 | 99.10 | 0.9149 |
| ReliefF | 89.92 | 83.77 | 99.10 | 0.9077 |
| AdaBoost | **92.73** | **88.72** | 97.90 | **0.9306** |
| RF | 88.91 | 86.76 | 93.73 | 0.8933 |

**表 5-5　PCMAC 数据集实验结果**

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| Method-1 | **87.85** | 81.43 | 98.58 | **0.8915** |
| Method-2 | 78.07 | 70.17 | **98.89** | 0.8205 |
| FSBACO | 87.03 | 80.24 | 98.66 | 0.8846 |
| IG | 83.17 | 75.65 | 98.28 | 0.8548 |
| ReliefF | 82.66 | 74.94 | 98.68 | 0.8518 |
| AdaBoost | 87.65 | **85.21** | 92.16 | 0.8828 |
| RF | 83.17 | 80.24 | 90.75 | 0.8453 |

**表 5-6　RELATHE 数据集实验结果**

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| Method-1 | 76.31 | 69.83 | 99.62 | 0.8209 |
| Method-2 | 65.24 | 61.07 | **100** | 0.7581 |
| FSBACO | 76.10 | 69.68 | 99.23 | 0.8180 |
| IG | 62.93 | 59.89 | 97.57 | 0.7417 |
| ReliefF | 62.93 | 59.88 | 97.70 | 0.7420 |
| AdaBoost | **83.11** | **89.42** | 78.28 | **0.8347** |
| RF | 59.85 | 57.80 | 99.87 | 0.7300 |

**表 5-7　MADELON 数据集实验结果**

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| Method-1 | **91.81** | **92.23** | **91.33** | **0.9178** |
| Method-2 | 91.19 | 91.56 | 90.75 | 0.9112 |
| FSBACO | 90.62 | 92.01 | 89.00 | 0.9046 |
| IG | 87.96 | 88.26 | 87.62 | 0.8792 |
| ReliefF | 88.54 | 88.97 | 87.90 | 0.8841 |
| AdaBoost | 59.04 | 59.03 | 59.43 | 0.5916 |
| RF | 55.58 | 59.61 | 59.79 | 0.5340 |

**表 5-8　IONOSPHERE 数据集实验结果**

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| Method-1 | **97.72** | **96.59** | **100** | **0.9826** |
| Method-2 | 94.29 | 93.64 | 98.25 | 0.9586 |
| FSBACO | 96.30 | 95.11 | 99.10 | 0.9703 |
| IG | 90.87 | 89.71 | 96.66 | 0.9302 |
| ReliefF | 84.90 | 82.26 | 97.82 | 0.8926 |

续表

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| AdaBoost | 90.89 | 89.69 | 96.89 | 0.9310 |
| RF | 92.88 | 91.34 | 98.21 | 0.9462 |

**表 5-9 PRDS 数据集实验结果**

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| Method-1 | **80.77** | **78.67** | **100** | **0.8788** |
| Method-2 | 75.84 | 74.91 | 99.20 | 0.8530 |
| FSBACO | 79.61 | 78.58 | 98.53 | 0.8718 |
| IG | 70.39 | 73.77 | 91.62 | 0.8143 |
| ReliefF | 63.21 | 70.69 | 84.03 | 0.7642 |
| AdaBoost | 59.86 | 69.19 | 80.25 | 0.7394 |
| RF | 72.54 | 72.20 | **100** | 0.8375 |

**表 5-10 SONAR 数据集实验结果**

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| Method-1 | **98.57** | **100** | **97.40** | **0.9865** |
| Method-2 | 97.14 | 98.82 | 95.02 | 0.9685 |
| FSBACO | 95.68 | **100** | 90.48 | 0.9495 |
| IG | 81.78 | 85.40 | 74.41 | 0.7880 |
| ReliefF | 78.87 | 80.23 | 74.27 | 0.7688 |
| AdaBoost | 77.46 | 79.17 | 71.41 | 0.7416 |
| RF | 78.79 | 89.01 | 64.80 | 0.7381 |

**表 5-11 STATLOGHEART 数据集实验结果**

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| Method-1 | **89.63** | **90.33** | 91.20 | **0.9066** |
| Method-2 | 86.30 | 85.94 | 90.44 | 0.8796 |
| FSBACO | 87.78 | 86.91 | **92.09** | 0.8938 |
| IG | 81.48 | 80.76 | 87.20 | 0.8376 |
| ReliefF | 63.33 | 66.52 | 68.09 | 0.6706 |
| AdaBoost | 78.89 | 80.05 | 83.00 | 0.8138 |
| RF | 82.96 | 82.68 | 88.67 | 0.8520 |

**表 5-12 CLIMATE 数据集实验结果**

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| Method-1 | **95.37** | **95.18** | **100** | **0.9752** |
| Method-2 | 93.52 | 93.36 | **100** | 0.9655 |
| FSBACO | 95.19 | 95.14 | 99.79 | 0.9740 |
| IG | 92.78 | 93.37 | 99.20 | 0.9615 |
| ReliefF | 91.11 | 91.76 | 99.18 | 0.9531 |

续表

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| AdaBoost | 91.30 | 93.76 | 96.97 | 0.9531 |
| RF | 91.85 | 91.99 | 99.80 | 0.9572 |

**表 5-13　Z-ALIZADEH 数据集实验结果**

| 实验方法 | $P$/% | Pr/% | Re/% | $F1$ |
|---|---|---|---|---|
| Method-1 | **93.40** | **92.26** | **99.08** | **0.9555** |
| Method-2 | 90.10 | 89.13 | 98.11 | 0.9338 |
| FSBACO | 91.74 | 94.71 | 93.44 | 0.9403 |
| IG | 84.14 | 88.77 | 89.31 | 0.8899 |
| ReliefF | 72.91 | 77.45 | 87.45 | 0.8212 |
| AdaBoost | 88.13 | 90.55 | 92.97 | 0.9168 |
| RF | 85.50 | 85.86 | 95.27 | 0.9023 |

从两个方面分析表 5-4～表 5-13 中的结果，即两种不同实现方式的性能比较以及 MCSBFS 方法与其他方法的比较。

首先分析两种实现方式的性能差异。通过统计算法提供的最优值可以看出，与单目标蚁群算法实现的 Method-2 相比，多目标蚁群算法实现的 Method-1 在测试数据集上取得了更好的效果。在分类正确率 $P$、查准率 Pr 和 $F1$ 指标方面，Method-1 在所有测试数据集上均优于 Method-2，而在查全率 Re 指标上，Method-1 在 5 个数据集上优于 Method-2。造成该结果的原因是，Method-2 将模型通过聚合转换为单目标优化问题求解，由于将多目标优化问题转换为单目标优化问题时，算法仅能够搜索帕累托前沿的某一部分，造成其他部分帕累托解的丢失，从而导致无法寻找到较好的帕累托解。

其次，对多目标蚁群算法实现的方法 Method-1 与其他方法的性能作比较分析。可以看出，除了在 PCMAC 数据集的 Re 指标和 SONAR 数据集的 Pr 指标上，Method-1 在所有测试数据上都优于 FSBACO，特别是在 $F1$ 指标上，Method-1 都取得了更高的评估值，说明 Method-1 具有更好的分类性能，可以得出结论，使用互补特征子集能够进一步提高特征信息的利用率，从而提高算法的分类性能。观察 IG 和 ReliefF 方法，一方面，在多数测试数据集上，IG 的分类性能要好于 ReliefF，另一方面，IG 和 ReliefF 的指标值都要低于 Method-1，这说明仅使用特征选择方法无法获得更好的分类效果；同时 IG 和 ReliefF 的分类性能指标值也要低于 FSBACO，该结论表明，独立于分类器的特征选择无法取得更好的分类性能。最后分析集成学习方法 AdaBoost 与 RF，AdaBoost 在 BASEHOCK 和 RELATHE 数据集上的分类结果好于 Method-1，在 PCMAC 数据集上，查准率 Pr 也要好于 Method-1，但是在其他测试数据集上，Method-1 的分类性能好于 AdaBoost，这说明集成学习在能够一定程度上有效提升分类器的分类性能，然而由于 AdaBoost 没有使用特征选择，因此其效果要弱于 Method-1。此外，虽然 RF 结合了抽样和特征选择，但是 Method-1 的分类性能也要优于 RF，表明 Method-1 采用的互补特征子集能够更加有效提升集成分类器的分类性能。

## 本章小结

为了有效解决当前特征选择方法在高维数据实体分辨问题中特征利用率较低的不足，基于集成学习思想，提出针对高维数据实体分辨的多分类器方法。首先定义分类器性能度量和分类器相似性；然后设计集成分类器，使用分类正确率、特征个数和分类器间的相似性作为优化目标，从而选择互补特征子集并训练分类器，使得每个特征子集在具有较小规模的同时，多个互补特征子集能够有效利用高维数据蕴含的丰富信息。采用多目标蚁群算法和单目标蚁群算法两种方式实现模型，在标准测试数据集上与对比方法的实验结果显示，使用多目标蚁群算法实现的多分类器方法具有优越性。

## 本章参考文献

[1] 曹建军. 基于提升小波包和改进蚁群算法的自行火炮在线诊断研究[D]. 石家庄：军械工程学院，2008.

[2] Richard O D，Perer E H. Pattern Classification and Scene Analysis[M]. New York：John Willey and Sons，2001：131-132.

[3] 曹建军，刁兴春，杜鹢，等. 基于蚁群特征选择的相似重复记录分类检测[J]. 兵工学报，2010，31(9)：1222-1227.