第 5 章　文件切分方案评价模型研究与构建

5.1　文件切分方案评价指标体系的建立

对于一个较大的基因序列文件，应将其切分为若干个小的文件，然后对切分后的文件进行两两比对[1]。假设将一个较大的基因序列文件切分为 m 个大小相同子文件，m 值的不同即切分文件个数不同会对分布式集群的存储性能、计算性能、节点的平均计算量产生影响。因此，建立文件切分方案评价指标体系，通过分析评价指标对文件切分方案进行优劣评价，即寻找最优的切分方案。文件最优的切分方案就是确定文件切分个数 m，使得文件分发到分布式集群中的各个节点后，进行基因序列比对的总体效率最优。要使得切分方案最优就要保证文件分发后，在满足数据本地化的前提下，达到分布式集群中各个节点的计算负载均衡、存储量最小和节点平均计算量最小。因此，影响分布式集群总体效率的指标包括以下 3 个方面。

（1）计算负载均衡——集群中各个节点上分发到的任务量均衡，避免出现任务量上的节点先执行完成等待其他节点的情况。此处假定任务量与文件大小之间关系表达式为 $rw=f(\text{size})$。

（2）节点存储量最小——集群中各个节点上分发到的存储量越小越好且不超过计算机存储上限，使得整个集群占用的存储空间最小。

（3）节点平均计算量最小——集群中各个节点上的计算量越小越好，使得整个集群的总体计算量最小。

文件分发后进行序列比对的集群总体效率最优设计问题是一个多目标

规划问题，要实现影响分布式集群总体效率的 3 个指标整体效果的最大化。因此首先需要对 3 个指标实现归一化处理。

当文件的切分份数 m 值确定之后，当前文件切分方案下的节点平均任务计算量、当前节点的任务计算量、节点平均存储量、当前节点的存储量、计算量下限（当前文件切分方案下各节点中分配的最小计算量）和节点平均计算量可以通过第 2 章构建的文件分发模型式(2-16)求得。

在寻找到相关影响指标的同时要对相关数据进行归一化处理[2]。

(1) 负载均衡的归一化处理。假设在文件切分数量为 $k(k=1,2,\cdots,m)$ 时的各节点平均任务计算量为 r_{nk}，r_{ik} 表示当前节点的任务计算量，则不同切分数量下的当前任务负载均衡的归一化处理公式如式(5-1)所示。

$$\frac{\sum_{i=1}^{n} \mid r_{ik}-r_{nk} \mid}{\max\limits_{k}\sum_{i=1}^{n} \mid r_{ik}-r_{nk} \mid} \quad (k=1,2,\cdots,m) \tag{5-1}$$

(2) 节点存储量的归一化处理。假设在文件切分数量为 $k(k=1,2,\cdots,m)$ 时的各节点最小存储量为 $c_{\min}^{nk}$，则不同切分数量下的存储最小化归一化处理公式如式(5-2)所示。

$$\frac{c_{\min}^{nk}}{\max\limits_{k} c_{\min}^{nk}} \quad (k=1,2,\cdots,m) \tag{5-2}$$

(3) 节点平均计算量的归一化处理。在不同的当前文件切分方案下节点的平均计算量是不同的，假设 $j_{\min}$ 表示在不同切分数量下的计算量下限，j_{nk} 为当前分发方案下的节点平均计算量，则节点的平均计算量的归一化处理公式如式(5-3)所示。

$$\frac{j_{nk}-j_{\min}}{j_{\min}} \tag{5-3}$$

5.2 利用层次分析法建立文件切分模型

文件切分方案以确定切分后文件的个数 m 为目的来测定负载均衡、存储量和节点的平均计算量这 3 个指标，整个过程均需要评价者的参与。层次分析法是一种需要评价者做出相应反应或指示的综合评价方法，在层次分析法中，首先需要将与决策有关的元素划分到目标层、准则层、方案层[3-5]，并在此基础之上进行定性和定量分析。层次分析法包括以下 5 个基本步骤。

步骤 1：建立层次结构模型。

步骤 2：构造成对比较阵。

步骤 3：计算权向量并做一致性检验。

步骤 4：计算组合权向量并做组合一致性检验。

步骤 5：进行决策。

根据层次分析法的基本过程，首先对文件切分方案建立层次结构模型，在建立层次结构模型之前，需要根据 3.1 节中的分析与描述构建层次分析法所需的因素集。所谓因素集就是将所有影响研究对象的因素集合起来。一般以 U 表示集合，假设影响因素有 n 个，因素集就表示为 $U=\{u_1,u_2,\cdots,u_n\}$，$u_i(i=1,2,\cdots,n)$表示的因素具有程度不同的模糊性。文件切分方案评价的影响因素有负载均衡、存储量、平均计算量，因此因素集 $U=\{u_1,u_2,u_3\}$，即 $U=\{$负载均衡，存储量，平均计算量$\}$。

模糊综合判定可分为一级模糊综合判定[6]和多级模糊综合判定[7]。本书主要使用一级模糊综合判定。一级模糊综合判定的基本方法和步骤如下。

1. 建立因素集

因素集是以影响判定对象的各种因素为元素所组成的一个普通集合，通

常用 U 表示，即 $U=\{u_1,u_2,\cdots,u_m\}$，$u_i(i=1,2,\cdots,m)$ 代表影响因素，这些因素可以是模糊的，也可以是非模糊的，但它们对因素集 U 的关系，要么 $u_i\in U$，要么 $u_i\notin U$，二者必居且仅居其一，因此因素集本身是普通集合。

2. 建立权重集

权重是一个相对的概念，是针对某一个指标而言的。某一个指标的权重是指该指标在整体评价中的相对重要程度。一般来说，各个因素的重要程度是不一样的。因此，应根据各因素的重要程度赋予相应的权重 $a_i(i=1,2,\cdots,m)$，并要求 $\sum_{i=1}^{m}a_i=1,a_i>0$。由各权重所组成的集合 $A=(a_1,a_2,\cdots,a_n)$ 称为因素权重集，简称权重集。

3. 建立备择集（评价集、评语集）

备择集是评判者对评判对象可能作出的各种总的评判结果所组成的集合，通常用 V 表示，即 $V=\{v_1,v_2,\cdots,v_n\}$，v_i 表示各种可能的评判结果，如评判一个教师的素质，评判结果可分为很好、好、一般和差，则备择集 $V=\{$很好，好，一般，差$\}$。

4. 单因素模糊评价

单独从一个因素出发进行评判，以确定评判对象对备择元素的隶属程度，称为单因素模糊评判。

设评判对象按因素集中第 i 个因素 u_i 进行评判，对备择集中第 j 个元素 v_i 的隶属程度为 r_{ij}，则按第 i 个因素 u_i 评判的结果为

$$R_i=(r_{i1},r_{i2},\cdots,r_{in})\quad(i=1,2,\cdots,m)$$

以各单因素评判结果为行组成的矩阵，称为单因素评判矩阵，即

$$\boldsymbol{R}=(r_{ij})=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix}$$

单因素评判矩阵是一个难点，主要是确定隶属度往往带有主观性，常用确定隶属度的方法有专家打分法、线性函数法（三角形法和梯形法）等。

5. 模糊综合评判

模糊综合评判就是综合考虑所有因素，得出正确的评判结果。

从单因素评判矩阵 $\boldsymbol{R}$ 可以看出：$\boldsymbol{R}$ 的第 i 行反映了第 i 个因素隶属于备择集各元素的程度，而 $\boldsymbol{R}$ 的第 j 列反映了所有因素隶属于备择元 v_j 的程度，若用第 j 列元素之和 $R_j=\sum_{i=1}^{m} r_{ij}$，$j=1,2,\cdots,n$ 来反映所有因素的综合影响，则并不能够同时考虑各因素的重要程度。若考虑到权重集，则便能合理地反映所有因素的综合影响，因此，$\boldsymbol{F}$ 综合评判可表示为

$$\boldsymbol{B}=\boldsymbol{A}*\boldsymbol{R}=(a_1,a_2,\cdots,a_m)*\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix}=(b_1,b_2,\cdots,b_n)$$

$b_j=\{b_1,b_2,\cdots,b_n\}$ 称为 $\boldsymbol{F}$ 综合评判指标，简称评判指标。b_j 的含义是：综合考虑所有因素的影响时，评判对象对备择元 v_j 的隶属度。

在 $\boldsymbol{B}=\boldsymbol{A}*\boldsymbol{R}$ 中，运算“$*$”的取法不同，结果也不完全相同，但相差不大。“$*$”的取法主要有如下几种：

模型Ⅰ：$M(\wedge,\vee)b_j=\bigvee_{i=1}^{m}(a_i\wedge r_{ij})(j=1,2,\cdots,n)$

这里“$\wedge$”“$\vee$”表示取小、取大运算。

模型Ⅱ：$M(\cdot,\vee)b_j=\bigvee_{i=1}^{m}(a_i\cdot r_{ij})(j=1,2,\cdots,n)$

这里“·”是普通乘法。

模型Ⅲ：$M(\wedge,\oplus)b_j=\min\left\{1,\sum_{i=1}^{m}(a_i\wedge r_{ij})\quad(j=1,2,\cdots,n)\right\}$

模型Ⅳ：$M(\cdot,\oplus)b_j=\min\left\{1,\sum_{i=1}^{m}(a_i\cdot r_{ij})\quad(j=1,2,\cdots,n)\right\}$

模型Ⅴ：指数模型 $b_j=\bigwedge_{i=1}^{m}r_{ij}^{a_i}$

模型Ⅰ是一种制约性主因素突出型模型，不宜应用于因素太多或太少的情况。

模型Ⅱ和模型Ⅲ与模型Ⅰ比较，能较好地反映单因素评价结果和因素的重要程度。

模型Ⅳ不仅考虑了所有因素的影响，而且保留了单因素评判的全部信息，该模型称为加权平均型模型，在实践中常常用该模型。

模型Ⅴ的最大特点是评判对象的综合评判指标等于所有 $r_{ij}^{a_i}$ 的最小值，因此为了有效提高评判对象的综合评判指标，务必全面改善所有单因素指标才能达到目的，该模型是一种制约性全面促进型模型。在实际应用时，根据问题的要求，选择恰当的模型进行运算。

6. 评判指标的处理

对评判指标 $b_j=\{b_1,b_2,\cdots,b_n\}$，可根据以下几种方法确定评判对象的具体结果。

（1）最大隶属度法[8]。若 $b_l=\max\{b_1,b_2,\cdots,b_n\}$，则评判结果隶属于 v_l。

（2）加权平均法[9]。取以 b_j 为权重，对各个备择元素 v_j 进行加权平均的值为评判结果，即

$$V=\frac{\sum_{j=1}^{n}b_jv_j}{\sum_{j=1}^{n}b_j}$$

此法要求将 v_j 中非数量性备择元素数量化。

(3) F 分布法[10]。对 b_j 进行归一化，令 $b=\sum_{j=1}^{n} b_j$ ，则第 i 个指标归一化处理之后的和 B_i' 为

$$B_i'=\frac{b_1}{b}+\frac{b_2}{b}+\cdots+\frac{b_n}{b}=b_1'+b'_2+\cdots+b'_n$$

这样，b_j' 反映了评判对象在所评判的特性方面的分布状态，即所占的百分比。

通过分析得知，对文件切分方案评价的目标为对文件切分方案优劣进行综合评价，将其作为递阶层次结构最高层的元素。采用一种文件切分方案，在分布式集群环境下按照第 2 章式(2-12)和式(2-16)所示的文件分发模型实现文件分发之后，完成所有文件的两两比对任务，分布式集群整体性能是评价该切分方案优劣的唯一标准。影响分布式集群整体性能的主要因素包括各个节点的计算负载是否均衡、各个节点被分配文件存储量是否最小和各个节点的平均计算量是否最小 3 方面，即文件切分方案综合评价所需要考虑的准则。负载均衡 U_1、存储量 U_2、节点平均计算量 U_3 这 3 个指标共同构成了准则层元素的集合，上述准则对目标实现的重要程度可能不同，但是彼此之间独立，不存在支配关系，因此它们是有优先级顺序的同级关系。确定文件切分的个数 m 作为实现对文件切分方案优劣进行综合评价这个目标的措施方案，即将文件切分为大小尽可能均匀的 $k(k=1,2,\cdots,m)$ 份，放置在层次结构模型的最底层。

明确各个层次的因素及其位置后，将它们之间的关系用连线连接起来，构成递阶层次结构，如图 5.1 所示。

假设 3 个目标 U_1、U_2、U_3 的重要程度分别为 a_1、a_2、a_3，这 3 个系数可以由层次分析方法等确定得出，则当前文件切分方案的评价模型如式(5-4)所示。

$$\min_k\left(a_1\frac{\sum_{i=1}^{n}|r_{ik}-r_{nk}|}{\max_k\sum_{i=1}^{n}|r_{ik}-r_{nk}|}+a_2\frac{c_{\min}^{nk}}{\max_k c_{\min}^{nk}}+a_3\frac{j_{nk}-j_{\min}}{j_{\min}}\right) \tag{5-4}$$

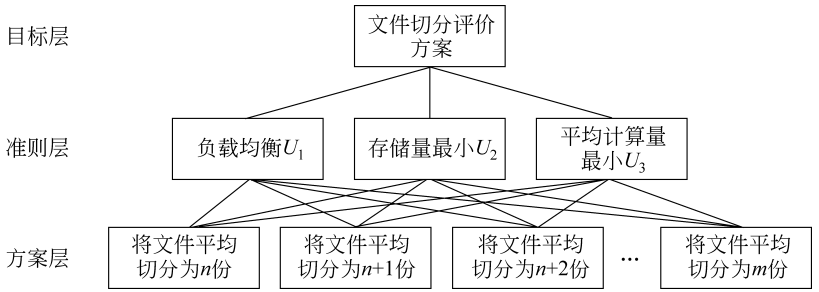

图 5.1　递阶层次结构图

同样对各个子指标也通过专家打分的方式获取到对应的权重。特别地，如果假定 3 个目标的重要程度是相等的，则目标函数变成如式(5-5)所示。

$$\min_k\left(\frac{1}{3}\frac{\sum_{i=1}^{n}\mid r_{ik}-r_{nk}\mid}{\max_k\sum_{i=1}^{n}\mid r_{ik}-r_{nk}\mid}+\frac{1}{3}\frac{c_{\min}^{nk}}{\max_k c_{\min}^{nk}}+\frac{1}{3}\frac{j_{nk}-j_{\min}}{j_{\min}}\right)\tag{5-5}$$

由于 3 个目标同等重要，因此无须进行构造成对比较矩阵，也无须进行步骤三、步骤四。

5.3　文件切分最优个数的确定

在确定文件切分评价模型后，根据现有节点个数 n 及不同切分数量可以筛选出最优的文件切分数量 m。其具体确定公式如式(5-6)所示。

$$m=\operatorname{opt}\left(\min_k\left(\frac{1}{3}\frac{\sum_{i=1}^{n}\mid r_{ik}-r_{nk}\mid}{\max_k\sum_{i=1}^{n}\mid r_{ik}-r_{nk}\mid}+\frac{1}{3}\frac{c_{\min}^{nk}}{\max_k c_{\min}^{nk}}+\frac{1}{3}\frac{j_{nk}-j_{\min}}{j_{\min}}\right)\right)\tag{5-6}$$

文件切分个数 m 值确定算法具体计算步骤如下。

第一步：利用式(2-12)所示文件分发模型确定最优的均衡化任务分配方

案，并计算出分布式集群中计算量最大节点的总计算量 c_{max}。

第二步：利用式(2-16)所示文件分发模型计算获得分布式集群中在每个节点的最大计算量不变的情况下，使得各个节点的总存储量最小化的任务重新分发方案。

第三步：在当前分布式集群中节点数量 n 不变的情况下，进一步计算不同切割数量下的相关优化结果的取值。

第四步：利用 3.1 节中的方法处理各项优化数据，并利用相关评价方法计算出最优 m 值。

文件切分个数 m 值确定算法相关伪代码详细描述见表 5.1。

表 5.1　m 值确定算法相关伪代码详细描述

输入：假定的 m 值
输出：集群存储占用总和、计算量总和、误差总和

```
%定义并初始化变量参数
n←分布式集群数量
s←文件大小
m←切分文件数量
for i=m1 to m2 do
    m←i+n;
    %将文件切分为 m 等份
    s←1/m*ones(m,1);
    %在当前切分数量下计算出最优任务指派方案
    [result,deci,result1,scqk,scdx]←taskassign_n(m,n,s);
    x←result1(:,length(result1(1,:)));
    y←abs(x-sum(x)/length(x));
    %计算在当前切分数量下计算量 r1,计算量误差 r2 和存储量总和 r3
    r1(i),r2(i),r3(i)←sum(x,y,scdx);
end for
q←三个目标的权重;
%标准化处理
rs3,rs2,rs1←normalized(r3,r1,r2);
%总体评价结果
reuslt←q(1)*rs3+q(2)*rs2+q(3)*rs1;
```

在文件切分个数 m 值确定的实验中基于计算的方便，假定将一个充分大的基因序列文件等份的切分成 m 份，并通过相关理论寻找最优的 m 取值。假定分布式集群中节点数量分别为 3、4、5 的不同切分大小情况下，计算获得了存储大小总和、计算量总和、误差总和大小等相关数据的取值情况及其对应评价结果。具体结果见表 5.2、表 5.3 和表 5.4。

表 5.2　节点数为 3 时对应评价结果

m 取值($n=3$)	4	5	6	7	8	9	10
存储大小总和	0.75	0.666 667	0.722 222	0.714 286	1	1	1
计算量总和	0	0.333 333	0.666 667	1	1.333 333	1.666 667	2
误差总和大小	0	0.8	0	0	1	0	0
评价分值	0.250	0.544	0.333	0.405	0.889	0.611	0.667

表 5.3　节点数为 4 时对应评价结果

m 取值($n=4$)	5	6	7	8	9	10	11
存储大小总和	0.8	0.833 333	0.809 524	0.833 333	0.888 889	1	1
计算量总和	0	0.25	0.5	0.75	1	1.25	1.5
误差总和大小	1	0.416 667	1.071 429	0	0	0.5	0.227 273
评价分值	0.600	0.472	0.738	0.444	0.519	0.778	0.742

表 5.4　节点数为 5 时对应评价结果

m 取值($n=5$)	6	7	8	9	10	11	12
存储大小总和	1	0.857 143	0.825	0.833 333	0.84	0.845 455	0.9
计算量总和	0	0.2	0.4	0.6	0.8	1	1.2
误差总和大小	0	0.964 286	0.5625	1	0	0	0.5625
评价分值	0.333	0.663	0.574	0.778	0.502	0.560	0.821