网格环境中相似资源的模糊聚类研究①

Research on Similar Resources Based on Fuzzy Clustering in Grid Environment

张险全 王 亮 陈未如 （沈阳化工学院 计算机科学与技术学院 辽宁 沈阳 110142）

摘 要: 采用模糊聚类的方法对网格系统中的最大相似性资源进行聚类，并且采用多维向量的方式表示网格资源，根据用户对资源向量各维的关心程度的不同进行加权，提出了加权欧氏距离法计算资源之间的相似度，通过构造 F－分布统计量的方法确定最佳分类，并且给出了理论证明。最后，通过仿真实验将网格资源进行分类，实验结果表明文中的聚类方法以及确定最佳分类的方法能够有效的将具有最大相似性的网格资源聚在同一簇内。

关键词: 网格计算 网格资源 模糊聚类 加权欧氏距离 F－分布统计量

1 引言

网格计算[1,2]是下一代并行分布式计算方法，它聚集了大量的、分散的和异构的资源，求解科学、工程与商业应用中的大规模高性能问题[3]。资源管理问题是构造网格系统的基础问题之一，资源组织是否合理直接影响网格系统的效率。由于网格资源的复杂性和多样性，所以对网格资源的组织和管理都十分困难。在科学技术和经济管理中经常采用分类的方法，把大量的资源有条理的组织起来。本文也借助这种分类的思想将网格中的资源进行聚类分析。由于网格资源非常丰富、异构性强，很难精确计算资源的相似度，所以本文介绍了一种模糊聚类最大相似性资源的方法，并分析了分类的合理性。

本文将从以下几个方面来介绍网格环境中相似资源的模糊聚类。首先介绍了网格系统中的资源表示方式，接着介绍了网格资源聚类模型、最大相似资源模糊聚类方法步骤，然后介绍了如何构造 F－统计量来确定最佳分类，最后通过仿真实验对网格环境中的最大相似资源模糊聚类方法以及确定最佳分类方法进行验证。

2 资源描述

网格资源之间虽然没有严格的属性区分，但是在某些层面上具有一定的关联，因此更适合于模糊聚类。最大相似性资源的聚类分析是以资源的属性为基础的，所以资源描述的方式直接影响聚类的结果。在本文中采用多维向量来表示网格资源。综合分析网格资源属性，本文定义以下 6 个特征属性来描述网格系统中的资源[4]：(1) 资源类别（a_1）：主要描述资源的服务类型，例如计算资源、存储资源以及网络资源等；(2) 服务能力（a_2）：表示资源在单位时间内向用户提供的服务量；(3) 服务质量（a_3）：表示资源给用户所提供服务的满意程度或者是服务的使用者同服务的提供者之间关于服务所能提供的质量的一种约定[5]；(4) 使用价格（a_4）：表示用户使用该资源花费的代价；(5) 网络带宽（a_5）：表示资源实体之间的网络情况以及数据传输速度限制。(6) 资源负载（a_6）：表示资源在一段时间内的负载平衡情况。根据以上特征属性，网格资源用一个 6 维数据向量表示为：

$$R_0 = (a_1, a_2, a_3, a_4, a_5, a_6) \quad (1)$$

3 网格资源的模糊聚类分析

假设网格系统中有 n 个资源节点，每个资源节点用 m 维向量表示，则资源集合 R：$R = \{R_1, R_2, R_3, \ldots, R_n\}$，其中第 i 个资源实体：$R_i = (R_{i1}, R_{i2}, \ldots, R_{im})$，$i \in [1, n]$。

① 基金项目：辽宁省教育厅科学技术研究项目（20060675）

3.1 网格资源聚类模型

为了更好的将网格资源进行分类,首先需要建立一个聚类模型。网格资源聚类模型可以采用如下方式描述:对于资源节点集合:$R=\{R_1,R_2,R_3,\ldots,R_n\}$,$i\in[1,n]$,找出其子集构成的集合$\{G_j\}$,$j\in[1,r]$,满足$\bigcup_{i=1}^{r}G_i=R$且$\forall i,j\in[1,r],i\neq j\Rightarrow G_i\cap G_j=\varnothing$,其中 r 表示子集合个数,把每个子集合称为一个虚拟组织(VO),也称为簇。下面给出一些文中所用到的相关概念的定义:

定义 1. 簇质心:我们把每个簇的向量中心称为簇质心。假设第 j 簇的容量为 n_j,则第 j 簇记为:$R_1^{(j)}$,$R_2^{(j)}\ldots\ldots R_{nj}^{(j)}$,第 j 簇的簇质心记为:$\bar{R}^{(j)}=(\bar{R}_1^{(j)},\bar{R}_2^{(j)}\ldots\ldots\bar{R}_m^{(j)})$,其中 $\bar{R}_k^{(j)}$ 为第 j 簇的资源向量第 k 维的均值,$\bar{R}_k^{(j)}$ 的定义如下:

$$\bar{R}_{\cdot k}^{(j)}=\frac{1}{n_j}\sum_{i=1}^{n_j}R_{ik}^{(j)}\quad(k=1,2,\ldots,m)$$

定义 2. 总体质心:把资源集合 R 的中心向量称为总体质心,记为 $\bar{R}=(\bar{R}_1,\bar{R}_2\ldots\ldots\bar{R}_m)$,其中 $\bar{R}_k$ 是资源向量第 k 维的均值,即

$$\bar{R}_{\cdot k}=\frac{1}{n}\sum_{i=1}^{n}R_{ik}$$

定义 3. 簇内距:同一个簇中的所有样本到簇质心的距离平方和的算术平方根,则第 j 簇的簇内距计算公式定义为:$d^j=\sqrt{\sum_{i=1}^{n_j}\|R_i^{(j)}-\bar{R}^{(j)}\|^2},j\in[1,r]$

其中 $R_i^{(j)}$ 是第 j 簇的第 i 个资源节点,$\bar{R}^{(j)}$ 是第 j 簇的簇质心。

定义 4. 簇间距:簇容量与簇质心到总体质心之间的欧氏距离的乘积,则第 j 簇的簇间距的计算公式定义为:

$D^j=\sqrt{n_j\|\bar{R}^{(j)}-\bar{R}\|^2}$　j∈[1,r]　,其中 $\bar{R}^j$,$\bar{R}$,n_j 分别是簇质心、总体质心和第 j 簇的簇容量。

3.2 数据标准化

网格系统中的资源是用多个特征属性来描述的,各个指标的量纲通常是不一样的,为了使有不同的量纲的量也能进行比较,一般需要对数据作适当的变换即平移.标准差变换和平移.极差变换[6]。

平移.标准差变换公式:

$$R_{ik}=\frac{R_{ik}-\bar{R}_{\cdot k}}{S_k}\quad(i=1,2,\ldots,n;k=1,2,\ldots,m)\qquad(2)$$

其中 $\bar{R}_k$ 是网格资源的第 k 维的均值,其公式为:$\bar{R}_{\cdot k}=\frac{1}{n}\sum_{i=1}^{n}R_{ik}$,$S_k$ 是资源样本的标准差,其公式为:

$$S_k=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R_{ik}-\bar{R}_{\cdot k}\right)^2}$$

由统计学相关理论可知,经过变换后每个变量的均值为 0,标准差为 1,并且消除了量纲的影响,但是 R_{ik} 并不一定在区间[0,1]上。由于网格资源之间的相似系数是在区间[0,1]上的,所以通过变换后的数据如果不在区间[0,1]上的需要把数据压缩到区间[0,1]上。通常采用平移.极差变换公式:

$$R_{ik}''=\frac{R_{ik}'-\min\limits_{1\le i\le n}\{R_{ik}'\}}{\max\limits_{1\le i\le n}\{R_{ik}'\}-\min\limits_{1\le i\le n}\{R_{ik}'\}}\quad(j=1,2,\ldots,m)\qquad(3)$$

3.3 建立模糊相似矩阵

根据传统的聚类方法确定相似系数,建立模糊相似矩阵,资源 R_i 与资源 R_j 之间的相似程度 $r_{ij}=R(R_i,R_j)$ 且 r_{ij}[0,1]。根据网格资源的特点确定 $r_{ij}=R(R_i,R_j)$ 的方法可以采用距离法,计算公式如下:

$$R_{ik}''=\frac{R_{ik}'-\min\limits_{1\le i\le n}\{R_{ik}'\}}{\max\limits_{1\le i\le n}\{R_{ik}'\}-\min\limits_{1\le i\le n}\{R_{ik}'\}}\quad(j=1,2,\ldots,m)\qquad(4)$$

其中 $\sqrt{\sum_{k=1}^{m}\left(R_{ik}-R_{jk}\right)^2}$ 表示资源 R_i 与资源 R_j 之间的欧氏距离,C 是欧氏距离的比例系数,是一个常数。该公式是采用网格资源在 m 维的欧氏空间中的距离来度量两个资源之间的相似程度。

网格资源中的性能指标比较多,采用欧氏距离公式计算相似度是将所有的性能指标都是平等的看待所得到的综合相似度,而综合相似度是对网格资源综合性能的反映,但是有时候人们只关心网格资源的部分指标,只要那些人们关心的指标具有很高的相似性即可聚为一簇,如果采用欧氏距离公式将会导致本来应该聚为一簇的网格资源却由于某些并不关心的指标所

影响而没有聚在同一簇中。为了解决这个问题,针对某些指标关心程度的不同,对每个指标进行加权,用户关心的指标权值大,反之则小,权值的大小表示用户关心的程度。因此,本文提出了通过加权欧氏距离来计算资源间的相似度。加权欧氏距离是将网格资源向量各维之间的距离乘以各自的权值系数,权值系数的取值大小是由用户关心程度决定的。资源 R_i 与资源 R_j 之间的加权欧氏距离 d_{ij} 计算公式为:

$$d_{ij}=\sqrt{\sum_{k=1}^{m}C_k\left(R_{ik}-R_{jk}\right)^2}$$

,其中 $d_{ij}\leqslant 1$,C_k 是常数,表示用户对网格资源的第 k 个指标的关心程度,因此,利用加权欧氏距离计算资源之间的相似度公式变为:

$r_{ij}=1-d_{ij}$ (5),$\forall K\in[1,m]$,$C_k\equiv C$ 如果恒成立,C 为常数,那么加权欧氏距离公式就变成了欧氏距离公式。

3.4 聚类最大相似资源

在经典数学中,我们对有限集合的划分是以等价关系为依据的也即是一种等价关系决定一个分类,同样的在模糊数学中,有限论域上的模糊等价关系可以决定一个分类。因此,我们需要将模糊相似矩阵转化成模糊等价矩阵。基于模糊等价矩阵聚类的方法主要有两种:传递闭包法和 Boole 矩阵法。大量的实验证明这两种方法在分类效果上是一致的,当网格资源数量较大时,传递闭包法聚类效率较高,因此,本文采用传递闭包法聚类。下面给出求模糊等价矩阵的定理。

定理 1. 假设 $R\in\mu_{n\times n}$ 是模糊相似矩阵,则存在一个最小自然数 k($k\leqslant n$),使得传递闭包 $t(R)=R^k$,对于一切大于 k 的自然数 z,恒有 $R^k=R^z$,此时,$t(R)=R^k$ 是模糊等价矩阵。

定理 1 的证明在此省略,详细证明可以参考文献[7]。通常使用平方法求传递闭包 t(R)。从模糊相似矩阵 R 出发,依次求平方 $R\to R^2\to R^4\to\cdots\to R^{2^r}\to\cdots$,当第一次出现 R^k。$R^k=R^k$ 时(表明 R_k 具有传递性),R_k 就是所求的传递闭包 t(R)。

得到了模糊等价矩阵之后,再让 λ 由大到小依次取 λ-截矩阵进行聚类,根据 λ-截矩阵就可以得到相对应的一系列分类从而形成动态的聚类图。尽管能够得到一系列的分类,但是很难确定在网格环境下的最佳分类。下面介绍最佳分类的确定方法。

4 确定最佳分类

网格资源模糊聚类的目的就是尽量让具有最大相似性资源聚在同一个簇内,使得所得的分类具有高内聚性。从模糊聚类分析过程中我们可以看出对于各个不同的 $\lambda\in[0,1]$,可得到不同的分类,从而形成一种动态聚类图,这样尽管可以全面的了解资源样本的分类情况,但是不能确定哪一种分类是最合理的,因此,如何确定最佳值是聚类最大相似资源的一个关键问题。

一个最佳的分类应该是每个簇与簇之间的资源差别相对比较大,而同一个簇内的资源之间的差别相对比较小,所以根据簇间距与簇内距之间的关系来构造统计量,然后根据方差分析原理来判断簇与簇之间的差异是否显著来确定最佳分类。本文通过构造 F-统计量的方法将簇内距和簇间距联系起来。下面是构造 F-统计量所使用的资源原始数据如表 3.1 所示:

表 1　资源向量原始数据表

样本	指标					
	1	2	…	k	…	m
R_1	R_{11}	R_{12}	…	R_{1k}	…	R_{1m}
R_2	R_{21}	R_{22}	…	R_{2k}	…	R_{2m}
…	…	…	…	…	…	…
R_i	R_{i1}	R_{i2}	…	R_{ik}	…	R_{im}
…	…	…	…	…	…	…
R_n	R_{n1}	R_{n2}	…	R_{nk}	…	R_{nm}
$\bar{R}$	$\bar{R}_1$	$\bar{R}_2$	…	$\bar{R}_k$	…	$\bar{R}_m$

式(*)即是所构造的 F-统计量:

$$F=\frac{\sum_{j=1}^{r}\left(D^j\right)^2}{\sum_{j=1}^{r}\left(d^j\right)^2}=\frac{\sum_{j=1}^{r}n_j\,\|\bar{R}^j-\bar{R}\|^2/(r-1)}{\sum_{j=1}^{r}\sum_{i=1}^{n_j}\|R_i^j-\bar{R}^j\|^2/(n-r)}\quad(*)$$

分子 $\sum_{i=1}^{r}(D^i)^2$ 反映了簇与簇之间的距离,而分母 $\sum_{i=1}^{r}(d^i)^2$ 反映了簇内的资源之间距离。因此,F 值越大,说明簇与簇之间的距离大,表明簇与簇之间的差异大,分类效果也就越好。

定理 2 如果 n 个网格资源通过模糊聚类得到了一种分类,并且分类数 $1<r<n$,d^i 和 D^i 分别是第 i 类的簇内距和簇间距,那么统计量:

$$F=\frac{\sum_{j=1}^{r}\left(D^{j}\right)^{2}}{\sum_{j=1}^{r}\left(d^{j}\right)^{2}}\sim F\left(r-1,n-r\right)$$

证明 在网格系统中,网格资源动态的加入网格系统是随机的,并且资源之间是独立的,随着时间的增加在一定时间内资源的动态加入数量是一个泊松过程。把所有资源组成的空间叫做资源样本空间 Ω,那么 Ω 中的资源样本应该服从同一个分布。由中心极限定理可知,网格资源总体是近似服从正态分布。不妨设总体 $R\sim N(\mu,\sigma^{2})$。

费歇引理:

$\sum_{i=1}^{n}\frac{\left(X_i-\bar{X}\right)^{2}}{\sigma^{2}}\sim\chi^{2}(n-1)$,其中 $X_i\sim N(\mu,\sigma^{2})$ 由 $R_i^{j}\in\Omega$ 即 $R_i^{j}\sim N(\mu,\sigma^{2})$,根据费歇引理可知:

$$\sum_{j=1}^{r}\sum_{i=1}^{n_j}\frac{\|R_i^{j}-\bar{R}^{j}\|^{2}}{\sigma^{2}}\sim\sum_{j=1}^{r}\chi^{2}\left(n_j-1\right)=\chi^{2}\left(n-r\right)\quad(a)$$

其中 $\sum_{i=1}^{r}n_i=n$,而 $\bar{R}^{i}\sim N(\mu,\frac{\sigma^{2}}{n_i})$,由统计学理论可得:

$$\sum_{j=1}^{r}n_j\|\bar{R}^{j}-\bar{R}\|\Big/\sigma^{2}=\sum_{j=1}^{r}\frac{\|\bar{R}^{j}-\bar{R}\|^{2}}{\sigma^{2}/n_j}\sim\chi^{2}\left(r-1\right)\quad(b)$$

将(b)式除以(a)式,根据 F－分布的定义可得:

$$\frac{\sum_{j=1}^{r}\frac{\|\bar{R}^{j}-\bar{R}\|^{2}}{\sigma^{2}/n_j}}{\sum_{j=1}^{r}\sum_{i=1}^{n_j}\frac{\|R_i^{j}-\bar{R}^{j}\|^{2}}{\sigma^{2}}}=\frac{\sum_{j=1}^{r}\left(D^{j}\right)^{2}}{\sum_{j=1}^{r}\left(d^{j}\right)^{2}}=F\sim F\left(r-1,n-r\right)$$ 证毕。

由定理 2 可知统计量 F 是服从自由度为 r－1,n－r 的 F－分布的。根据统计学相关理论对 F 进行分析:如果 $F>F_{\alpha}(r-1,n-r)$ $(\alpha=0.05)$,则根据数理统计方差分析理论可知簇与簇之间的差异是显著的,也即是说明了分类比较合理。如果满足不等式 $F>F_{\alpha}(r-1,n-r)$ 的 F 值不止一个,那么就取使得 $v=Max(F-F\alpha)$ 成立的分类即可。如果所有分类都不能使得 $F>F\alpha(r-1,n-r)$ $(\alpha=0.05)$ 成立,那么说明最佳分类数 $r\notin(1,n)$,而 $r\in[1,n]$ 且 r 为整数,所以 r =1 或者 r = n。通过聚类后得到的分类数 r 有 3 种情况:(1) r = n;(2) 1 < r < n;(3) r = 1。第一种情况说明网格资源之间没有相似性的;第二种情况说明有一些资源存在着一定的相似性的;第三种情况说明 n 个网格资源之间的相似性大,可以聚为一类。通常情况下网格资源聚类属于第二种情况,因此首先根据定理 2 判断是否能够得到最佳分类,如果不能得到最佳分类,那么就根据资源间的相似度直接判断分类数 r =1 还是 r = n。

5 仿真实验及结果分析

本实验采用模糊聚类法和传统的层次聚类法对网格中的资源进行聚类,最后对实验结果进行对比分析。实验的主要目的是将资源类别相同的、服务质量、价格、带宽以及负载都比较接近的网格资源聚类到同一簇中。

5.1 网格资源的物理实体到逻辑实体的映射

当网格资源节点加入到网格系统后,需要将资源节点映射成逻辑实体,也即是将资源节点用量化的多维向量表示出来。下面介绍一些相关的规则。

(1) 资源类别映射:资源类别通常是用字符串表示的,采用映射函数 F:S→V 量化,集合 S 是资源类别字符串组成的,集合 V 是通过映射得到的值域,通常资源类别比较相近的资源通过映射后所得的值应该在集合 V 中。例如,若是计算资源 V∈[1,10],若是存储资源 V∈[10,20],若既是计算资源也是存储资源,则 V 值就是 10 左右。

(2) 服务能力映射:根据资源节点加入网格系统时提供的参数进行确定。例如,如果加入的资源节点是计算资源,那么就根据处理器单位时间内执行的指令数来表示服务能力。

(3) 服务质量映射:采用 QoS 评价函数将服务质量进行量化。网格 QoS 模型可以参考文献[8]。QoS 评价函数也可以根据资源日志产生,资源日志描述了资源在一段时间内提供服务的情况。

(4) 价格映射:价格也是根据资源节点加入网格系统时提供的参数进行确定。

(5) 带宽映射:带宽可以通过计算得到,例如通过发送请求到收到确认消息来计算带宽。

(6) 负载映射:根据资源节点加入网格系统时提供的资源使用率来确定。例如,如果资源节点是计算资源,那么可以通过获取资源节点的 CPU 利用率。

5.2 采用模糊聚类法进行聚类

由于网格中的资源节点是随机加入或者离开的，因此随机产生的网格逻辑资源是有一定的代表意义的。假设网格逻辑资源数据如表 2。在本实验中，资源描述采用 6 维数据向量表示。

表 2　原始数据表

资源	原始数据					
	资源类别	服务能力(MIPS/S)	服务质量(%)	价格	带宽(Mb/S)	负载(%)
R1	71	46	20	52	100	5
R2	85	16	60	90	150	10
R3	20	15	60	18	80	80
R4	60	6	50	55	90	20
R5	7	18	30	20	60	80
R6	68	8	40	68	100	15
R7	10	55	80	39	40	30
R8	90	37	90	83	120	25
R9	8	25	70	32	80	5
R10	95	18	50	76	138	18

首先，根据资源的每一维特征属性值的均值和方差利用公式(2)和公式(3)将原始数据进行标准化处理，然后，采用加权欧式距离法计算相似度。在加权欧氏距离公式中的比例系数取值为：$C_1=0.1$，$C_2=0.03$，$C_3=0.1$，$C_4=0.1$，$C_5=0.1$，$C_6=0.3$。利用公式(5)构造资源的模糊相似关系矩阵。最后，采用平方法求传递闭包得到模糊等价矩阵，再让 λ 由大到小依次取 λ－截矩阵进行聚类，分类结果如下：

(1) 取 $\lambda=1.00$，分 10 簇，即 $\{R_1\}$，$\{R_2\}$，$\{R_3\}$ $\{R_4\}$，$\{R_5\}$，$\{R_6\}$，$\{R_7\}$，$\{R_8\}$，$\{R_9\}$，$\{R_{10}\}$；(2) 取 $\lambda=0.91$，分 9 簇，即 $\{R_1\}$，$\{R_2\}$，$\{R_3\}$，$\{R_4,R_6\}$，$\{R_5\}$，$\{R_7\}$，$\{R_8\}$，$\{R_9\}$，$\{R_{10}\}$；(3) 取 $\lambda=0.89$，分 8 簇，即 $\{R_1\}$，$\{R_2,R_{10}\}$，$\{R_3\}$，$\{R_5\}$，$\{R_4,R_6\}$，$\{R_7\}$，$\{R_8\}$，$\{R_9\}$；(4) 取 $\lambda=0.84$，分 7 簇，即 $\{R_1\}$，$\{R_2,R_4,R_6,R_{10}\}$，$\{R_3\}$，$\{R_5\}$，$\{R_7\}$，$\{R_8\}$，$\{R_9\}$；(5) 取 $\lambda=0.81$，分 5 簇，即 $\{R_1,R_2,R_4,R_6,R_{10}\}$，$\{R_3,R_5\}$，$\{R_7\}$，$\{R_8\}$，$\{R_9\}$；(6) 取 $\lambda=0.79$，分 4 簇，即 $\{R_3,R_5\}$，$\{R_1,R_2,R_4,R_6,R_8,R_{10}\}$，$\{R_7\}$，$\{R_9\}$；(7) 取 $\lambda=0.75$，分 3 簇，即 $\{R_1,R_2,R_4,R_6,R_8,R_{10}\}$，$\{R_3,R_5\}$，$\{R_7,R_9\}$；(8) 取 $\lambda=0.73$，分 2 簇，即 $\{R_1,R_2,R_4,R_6,R_7,R_8,R_9,R_{10}\}$，$\{R_3,R_5\}$；(9) 取 $\lambda=0.75$，分 1 簇，即 $\{R_1,R_2,R_3,R_4,R_5,R_6,R_7,R_8,R_9,R_{10}\}$；

将得到的分类利用定理 2 中的统计量 F 计算出各种 λ 值相对应的 F 值、$F_{0.05}(r-1,n-r)$ 和 $F-F_{0.05}$，其结果如表 4 所示。

表 4　分类结果表

	1	0.91	0.89	0.84	0.81	0.79	0.75	0.73	0.57
r 类	10	9	8	7	5	4	3	2	1
F	0	22.5	22	3.28	5.2	7.35	10.2	3.34	0
F0.05		239	19.4	8.94	6.19	4.76	4.74	5.32	
F－F0.05		−216.5	2.6	−5.66	−0.99	2.59	5.46	−1.98	

通过表 4 可知，有 3 个 F 满足不等式 $F>F_{0.05}$ 成立，再考查(F－F0.05)，差异最大的是 $\lambda=0.75$，$F-F_{0.05}=5.46$ 取得最大值。因此，第(7)类为最佳分类，此时，10 个网格资源可以分为 3 簇，第 1 簇：$\{R_1,R_2,R_4,R_6,R_8,R_{10}\}$ 第 2 簇：$\{R_3,R_5\}$ 第 3 簇：$\{R_7,R_9\}$。

5.3 采用层次聚类法进行聚类

实验数据如表 2 原始数据表，依次迭代直到聚为一类，则所有分类情况如下：

(1) 分 10 类，即 $\{R_1\}$，$\{R_2\}$，$\{R_3\}$，$\{R_4\}$，$\{R_5\}$，$\{R_6\}$，$\{R_7\}$，$\{R_8\}$，$\{R_9\}$，$\{R_{10}\}$；(2) 分 9 类，即 $\{R_1\}$，$\{R_2\}$，$\{R_3\}$，$\{R_4,R_6\}$，$\{R_5\}$，$\{R_7\}$，$\{R_8\}$，$\{R_9\}$，$\{R_{10}\}$；(3) 分 8 类，即 $\{R_1\}$，$\{R_2,R_9\}$，$\{R_3\}$，$\{R_4,R_6\}$，$\{R_5\}$，$\{R_7\}$，$\{R_8\}$，$\{R_{10}\}$；(4) 分 7 类，即 $\{R_1\}$，$\{R_2,R_9\}$，$\{R_3,R_5\}$，$\{R_4,R_6\}$，$\{R_7\}$，$\{R_8\}$，$\{R_{10}\}$；(5) 分 6 类，即 $\{R_1\}$，$\{R_2,R_4,R_6,R_9\}$，$\{R_3,R_5\}$，$\{R_7\}$，$\{R_8\}$，$\{R_{10}\}$；(6) 分 5 类，即 $\{R_2,R_4,R_6,R_9,R_{10}\}$，$\{R_1\}$，$\{R_3,R_5\}$，$\{R_7\}$，$\{R_8\}$；(7) 分 4 类，即 $\{R_1\}$，$\{R_3,R_5\}$，$\{R_7\}$，$\{R_2,R_4,R_6,R_8,R_9,R_{10}\}$，(8) 分 3 类，即 $\{R_3,R_5\}$，$\{R_7\}$，$\{R_1,R_2,R_4,R_6,R_8,R_9,R_{10}\}$，(9) 分 2 类，即 $\{R_3,R_5,R_7\}$，$\{R_1,R_2,R_4,R_6,R_8,R_9,R_{10}\}$，(10) 分 1 类，即 $\{R_1,R_2,R_3,R_4,R_5,R_6,R_7,R_8,R_9,R_{10}\}$。

对比分析两种聚类方法的结果：从两种聚类法的聚类过程到最终的结果可知，主要的差别是资源 R_9 被划分到不同的簇中，在模糊聚类中 $\{R_7,R_9\}$ 在同一簇；而在

层次聚类中 $\{R_1, R_2, R_4, R_6, R_8, R_9, R_{10}\}$ 在同一簇。从原始数据表中资源类别的角度来看，R_9 与 R_7 的资源类别指标参数比较接近，而与 $R_1, R_2, R_4, R_6, R_8, R_{10}$ 的资源类别指标参数相差很大，因此，R_9 与可以 R_7 可以认为是同类资源，而与 $R_1, R_2, R_4, R_6, R_8, R_{10}$ 不是同类资源节点；从服务质量、价格、网络带宽和负载的角度来看，R_9 与 R_7 之间的差异比较小，而与簇 $\{R_1, R_2, R_4, R_6, R_8, R_{10}\}$ 中的资源 R_2, R_8, R_{10} 都相差比较大；因此，R_9 与 R_7 划分到同一簇中是比较合理的。

导致 R_9 不能较好的被划分到自己所属簇中的主要原因是层次聚类法不能较准确的计算资源的相似度，从而使得分类的粒度比较大，最终将会导致聚类的结果偏差较大，不能较好的把满足用户需求的资源进行合理聚类。通过以上的对比分析可知，模糊聚类法能够较好的将网格资源进行合理的聚类。

6 结束语

本文提出了加权欧氏距离计算资源相似度的方法，采用模糊等价矩阵将网格资源进行聚类，并且构造了 F－分布统计量来确定最佳分类。最后，通过仿真实验进行对比分析，实验结果表明本文中的模糊聚类网格资源的方法能够根据用户的需求，把用户所关心的满足资源性能指标要求的网格资源进行聚类。

参考文献

1 Ian Foster, Carl Kesselman. The Grid: Blueprint for a New Computing Infrastructure. Elsevier Inc, Singapore, Second Edition, 2004.

2 Huaglory Tianfield. Towards Agent Based Grid Resource Management. IEEE International Symposium on Cluster Computing and the Grid 2005 (CCGrid 2005), 2005. 590 － 597.

3 Foster I, Kesselman C. 网格计算. 第 2 版. 北京: 机械工业出版社, 2005.

4 刘晓锋, 吴亚娟, 李明东. 一种基于模糊聚类的资源发现策略. 西华师范大学, 2007, (9).

5 Rashid AI － Ali, Gergor von Laszewski. QoS Support for High － Performance Scientific Grid Applications. IEEE International Symposium on Cluster Computing and the Grid. 2004: 134 － 143.

6 高新波. 模糊聚类分析及其应用. 西安: 西安电子科技大学出版社, 2004.

7 杨纶标. 模糊数学原理及应用. 广州: 华南理工大学出版社, 2005.

8 刘丽, 杨扬, 陈冬娥. 基于 QoS 的网格计算经济模型. 计算机应用研究, 2006, (10).