第十章　聚类分析与判别分析

聚类分析(cluster analysis)起源于分类学,在古老的分类学中,人们主要依靠经验和专业知识来实现分类,很少利用数学工具进行定量的分类,从而使分类的结果不可避免地具有主观性和随机性,并且不能揭示被分类对象的内在联系和区别。特别是当被分类的对象受到多个因素或指标影响时,由此做出的分类的可靠性就会更低。随着人类科学技术的发展,对分类的要求越来越高,于是人们逐渐将数学工具引入分类学中,形成了数值分类学。之后又将多元分析的技术引入数值分类学,形成了聚类分析。

聚类分析是一种根据研究对象特征对研究问题进行分类的多元分析方法,它主要是依据样本间相似性的度量标准将数据集自动分成几个群组,而且使同一个群组内的样本之间相似度尽量高,而不同群组的样本之间相似度尽量低的一种方法。

目前,聚类分析已经在各个领域得到广泛应用。在经济领域,对商业区或住宅区进行聚类,确定自动取款机(ATM)的设置地点;通过对消费者行为的研究,对市场进行细分,确定目标市场;在医学、生物学领域,对各种病症进行分类分析,通过挖掘出的一些骨骼的形状和大小对生物进行分类,对基因进行分类,以获得对种群的认识;在数据挖掘领域,作为其他数学算法的预处理步骤,获得数据分布状况,从而集中对特定的类做进一步的研究等。

判别分析与聚类分析不同,是在已知分组的前提下,根据已经确定分类的对象的某些观测指标和所属类别来判断未知对象所属类别的一种统计学方法。与聚类分析的不同之处在于:判别分析法的第一步是要对所研究对象进行分类,然后进一步选择对观测对象能够进行较全面描述的变量,进而按照一定的判别准则,建立一个或者多个判别函数,用研究对象的大量资料确定判别函数中的待定系数,并计算判别指标。对一个未确定分组的对象只要将其带入判别函数就可以判断其所属分类。在实际中,判别分析在气候分类、农业区划、土地类型划分中有着广泛的应用。再如,日常生活中,通过收集网上众多店铺经营的商品种类、品牌、价格、交易量等数据信息,来分析判别店铺的星级。

第一节　聚类分析方法概述

一、聚类分析的基本思想

我们一般认为,所研究的样本或指标之间存在不同程度的相似性。于是根据一批样本的多个观测指标,具体找出一些能够度量样本或指标之间相似程度的统计量。以这些统计量为划分类型的依据,把一些相似程度较大的样本(或指标)聚为一类,关系疏远的聚合到另一个大的分类单位,重复这个过程,直到把所有样本(或指标)都聚成一类,这样就可以形成一个由分散到统一的系统。

二、聚类分析方法

聚类分析方法可分为两大类：样品聚类分析（case cluster analysis，又称 Q 型聚类分析）和指标聚类分析（variable cluster analysis，又称 R 型聚类分析）。

Q 型聚类分析：对样品进行分类，没有唯一“正确”的分类方法。由实际工作者决定所需的分类数和分类情况。

R 型聚类分析：对变量进行分类，在每一类中找出有代表性的变量作为重要变量，利用少数几个重要变量进一步进行回归分析或 Q 型聚类分析。

具体的聚类方法有以下五种：

（一）系统聚类法

基本思想：① 计算 n 个样本两两之间的距离，构造 n 个类，每个只包含一个样本；② 合并最近的两类为一个新的类；③ 计算当前 $n-1$ 个类中，两两之间的距离；④ 如果此时类的个数为 1，聚类过程停止，否则继续重复步骤②、③、④。最后，可根据所研究问题的实际需要决定分类的个数和类。

（二）快速聚类法

基本思想：给定类数 k，确定 k 个点为“聚类种子”；然后将所有样本点按与这 k 个点的距离远近分为 k 类；再以这 k 类的重心为新的“聚类种子”，将所有样本点重新分类。如此下去，直到收敛得到最终的 k 类。

（三）两步聚类法

基本思想：首先，将记录预聚类为许多小子类；然后将小子类再聚类，如采用系统聚类法。

该聚类法主要处理非常大的数据集，能够处理连续变量和分类变量的混合数据，并且可自动确定类的数目。

（四）有序样本聚类法（最优分割法）

基本思想：开始时所有样本分为一类，然后分成两类、三类等，直到分为 n 类。要求分类后产生的离差平方和最小。

在使用该聚类法对样本进行聚类时，不能打乱样本的次序。当样本量 n 不大时，有可能讨论所有可能的分类结果，并在某损失函数意义下，从中得到最优解。

（五）模糊聚类

基本思想：首先，对原始数据进行变换；然后，计算模糊相似矩阵（夹角余弦或相关系数），建立模糊等价矩阵；最后，进行聚类。

第二节　聚类分析的基本概念

聚类分析是将性质相同或相近的个体聚为一类，但是如何衡量性质是否相同或相近，需要选取一些指标对相似性进行度量。不同类型的变量，其相似性的测度方法也有所不同。下面分别介绍数值变量和非数值变量的相似性测度方法。

一、数值变量的相似性测度

对样品进行聚类时，相似性一般用距离来衡量。点与点之间的距离、类与类之间的距离有不同的度量方法。

为方便说明，设 x,y 是两个要度量相似性的聚类变量，它们均含有 m 个分量。

（一）样本间的相似性度量

对样本进行聚类时，相似性一般用数学意义上的“距离”来衡量，距离是空间中的点与点的直观度量。常用的距离有如下几种定义方法：

1. 绝对值距离

$$\text{distance}(x,y) = \sum_{k=1}^{m} |x_k - y_k|$$

2. 欧氏距离

$$\text{distance}(x,y) = \sqrt{\sum_{k=1}^{m} (x_k - y_k)^2}$$

3. 平方欧氏距离

$$\text{distance}(x,y) = \sum_{k=1}^{m} (x_k - y_k)^2$$

4. 切比雪夫距离

$$\text{distance}(x,y) = \max_{1 \leqslant k \leqslant m} |x_k - y_k|$$

5. 明可夫斯基距离

$$\text{distance}(x,y) = \left[\sum_{k=1}^{m} |x_k - y_k|^p \right]^{\frac{1}{p}}$$

在五种距离的定义中，欧氏距离和平方欧氏距离是应用最为广泛的。而明可夫斯基距离是五种距离中最综合的，其他距离只是参数 p 取某些特殊值时的特例。

（二）对指标的相似性度量

对指标聚类时，通常根据相关系数或某种关联性来度量。

1. 夹角余弦（相似系数）

$$r = \frac{\sum_{k=1}^{m} x_k y_k}{\sqrt{\sum_{k=1}^{m} x_k^2} \sqrt{\sum_{k=1}^{m} y_k^2}}$$

2. 皮尔逊相关系数

$$r_{xy} = \frac{\sum_{k=1}^{m} (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m} (x_k - \bar{x})^2} \sqrt{\sum_{k=1}^{m} (y_k - \bar{y})^2}}$$

注意，两个指标的相关系数（或者相似系数）的绝对值越大，说明两个变量的性质越接近。

一般地，对变量聚类时，通常采用相关系数或者相似系数，原因如下：

（1）变量在概率统计意义上通常理解为随机变量，而度量随机变量之间的“距离”通常不能采用欧式距离等空间度量方式；

（2）变量通常有量纲，采用相关系数或者相似系数能够消去量纲。

需要注意的是，前面讲的大部分度量方法受变量的测量单位影响较大，数量级较大的数据变异性也较大，相当于对这个变量赋予了更大的权重，经常导致聚类结果产生很大的偏差。为了克服测量数据数量级等影响，在计算样本或者变量间的距离前，通常对数据进行标准化处理，将原始变量变成均值为0、方差为1的标准化数据。

（三）类与类之间的距离

以上我们介绍了如何测量点与点之间的距离。但在聚类分析的过程中，还需要测定包含若干点的类与类之间的距离。以下介绍几种比较常见的测度方法。

1. 最短距离法

最短距离法将类与类之间的距离定义为一个类中所有个体与另一个类中所有个体间距离的最小者。设 x_i 为类 A 中的任一个体，y_i 为类 B 中的任一个体，d_{ij} 表示个体 x_i 与 y_j 间的距离，D_{AB} 表示类 A 与类 B 间的距离，则最短距离法将类间距离 D_{AB} 定义为：$D_{AB} = \min_{x_i \in G_p, y_j \in G_q} d_{ij}$，如图10-1所示。

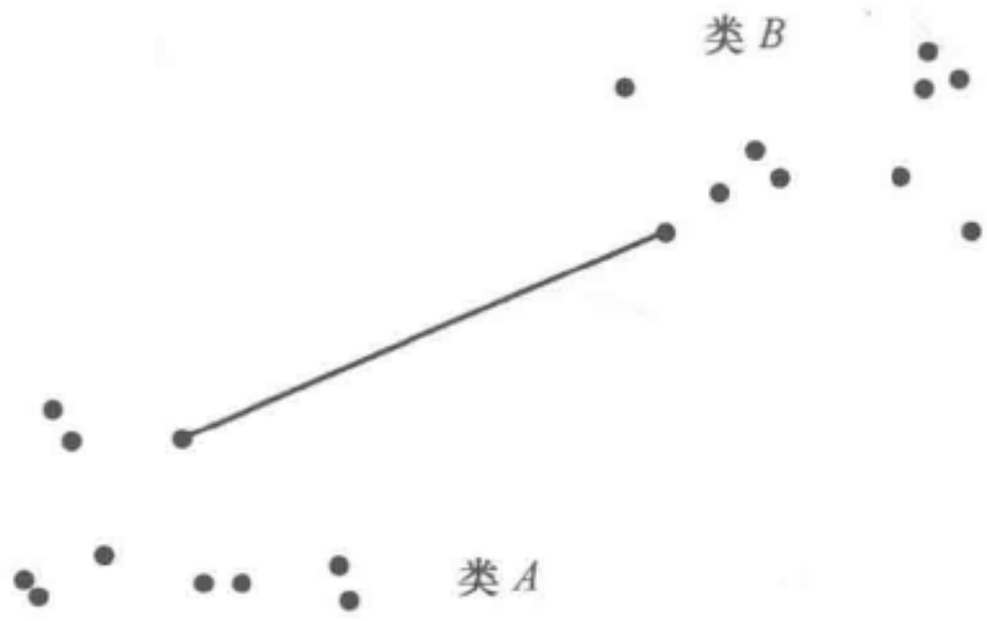

图10-1　最短距离法示意图

虽然最短距离法简单易用，能直观地说明聚类的含义，但是它有连接聚合的趋势，易将大部分个体聚在一类，所以其聚类效果并不好，实际中一般并不采用。

2. 最长距离法

最长距离法将两类间的距离定义为一类中所有个体与另一类中所有个体间距离的最大者，即 $D_{AB} = \max_{x_i \in G_p, y_j \in G_q} d_{ij}$，如图10-2所示。

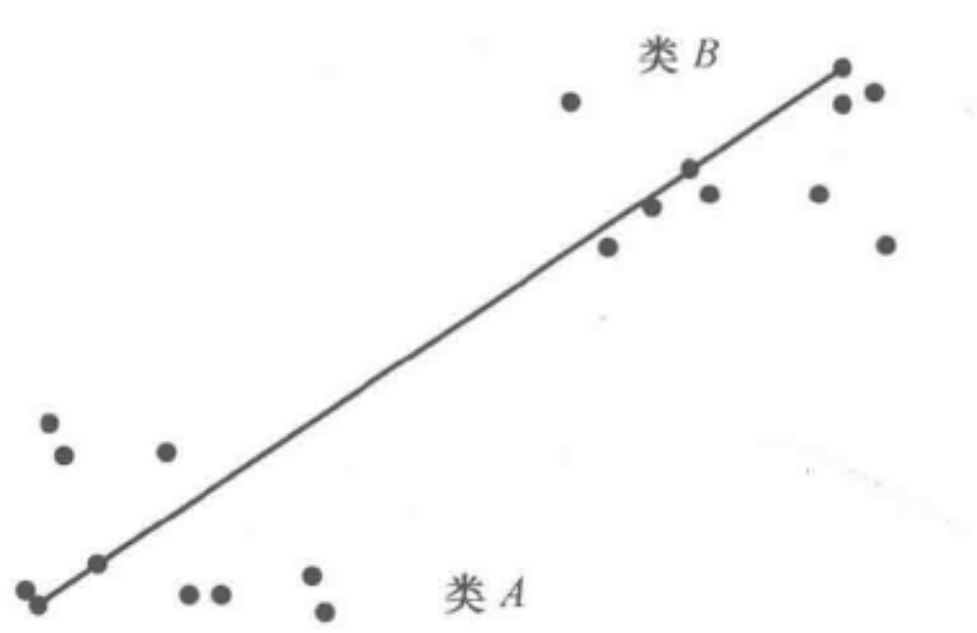

图10-2　最长距离法示意图

最长距离法克服了最短距离法连接聚合的缺陷，但是当数据离散程度较大时，会影响聚类的结果，导致产生较多的类。与最短距离法一样，最长距离法受异常值的影响也较大。

3. 未加权的类间平均法

未加权的类间平均法将变量间的距离定义为一个类中所有个体与另一类中所有个体间距离的平均值，即 $D_{AB}=\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}d_{ij}}{pq}$，如图 10-3 所示。

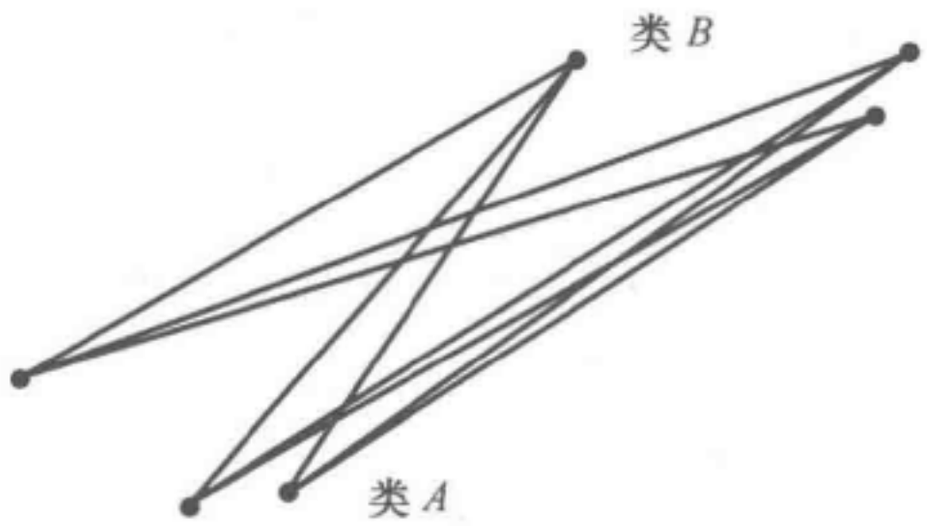

图 10-3　类间平均法示意图

类间平均法充分利用了已知信息，考虑到了所有个体，克服了最短距离法与最长距离法受异常值影响较大的缺陷，是一种聚类效果较好、应用较为广泛的聚类方法。

4. 加权的类间平均法

加权的类间平均法将各自类中的规模作为权数，其余与未加权的类间平均法相同。当群间的资料变异性较大时，加权的类间平均法比未加权的类间平均法更优。

5. 未加权的类间重心法

从物理学的角度看，一个类用它的重心（该类个体的均值）来代表是比较合理的。未加权的类间重心法就是将类间的距离定义为两个类重心间的距离。设类 A、类 B 的重心分别为 $\bar{x}_p$ 和 $\bar{y}_q$，则两个类间的距离为：$D_{AB}=\text{distance}(\bar{x}_p,\bar{y}_q)$，如图 10-4 所示。

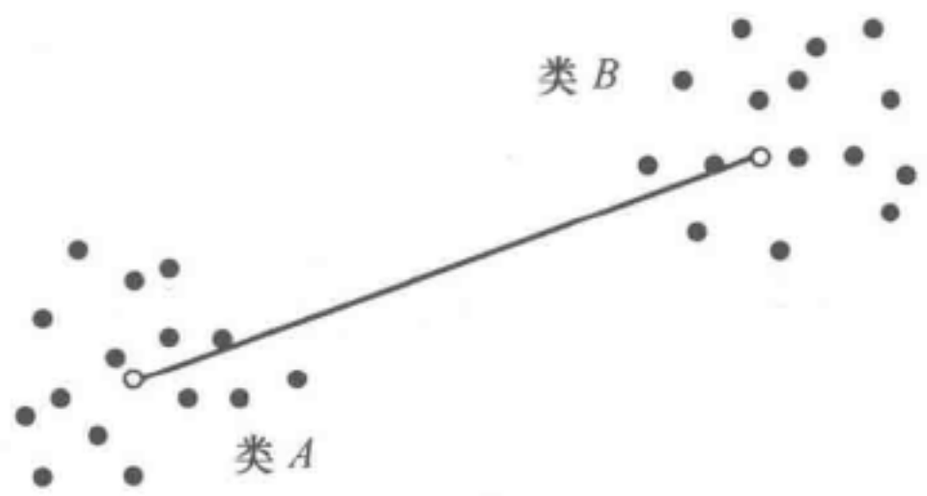

图 10-4　类间重心法示意图

重心法要求用欧氏距离测度点与点之间的距离，每聚一次类，都要重新计算重心。它较少受到异常值的影响，但因为群间距离没有单调递增趋势，在树状聚类图上可能出现图形逆转，从而限制了它的使用。

6. 加权的类间重心法

加权的类间重心法将各自类中的规模作为权数，其余与未加权的类间重心法相同。

当类间的资料变异性较大、类的规模有显著差异时，加权的类间重心法比未加权的类间重心法更优。

7. 离差平方和法

离差平方和法与前几种方法明显不同，它运用了变异数分析的思想。合理的聚类方法是使类内的差异尽量小，而类间的差异尽量大，即类内的离差平方和尽量小，类间的离差平方和尽量大。当类数固定时，使整个类内离差平方和达到极小的分类即为最优。

离差平方和法要求采用平方欧氏距离测度点与点间的距离。由于计算较为烦琐，在过去并未得到重视。但是随着计算机技术的发展，在很大程度上解决了复杂计算的问题，该方法被认为是一种理论上和实际上都非常有效的聚类方法，得到了较为广泛的应用。

二、非数值变量的相似性测度

关联测度常用于测度名义变量的相似性，一般基于列联表来计算。

设 x,y 均是取值为 0,1 的变量，两变量间的列联表如表 10-1 所示。其中，a 表示 x,y 均取 0 时的配对个数；b 表示 x 取 0，y 取 1 时的配对个数；x 共有 $a+c$ 个值取 0，y 共有 $a+b$ 个值取 0；每个变量共有 $a+b+c+d$ 个值。

表 10-1　列联表

y \ x	0	1	求和
0	a	b	$a+b$
1	c	d	$c+d$
求和	$a+c$	$b+d$	$a+b+c+d$

常用的关联测度方法是不匹配系数，即 x,y 取值不相同的个数与取值总数之比：

$$r=\frac{b+c}{a+b+c+d}$$

需要说明的是，适用于非数值变量的测度也一定适用于数值变量，但适用于数值变量的测度基本不能用于非数值变量。不同距离的选择对于聚类的结果有重要影响，因此在选择相似性测度时，一定要结合变量性质。

第三节　系统聚类法实例

一、基本原理

系统聚类法是目前国内外运用最多的一种聚类分析方法，它包含以下步骤：

第一步，计算 n 个样本（或指标）两两之间的距离；

第二步，每个样本各自成为一个类，这样共有 n 个类；

第三步，合并距离最近的两个类为一个新类；

第四步，计算新类与当前其他各类的距离，若类的个数等于 1，转到第五步，否则回到第三步；

第五步，画出聚类图；

第六步，根据聚类图等聚类结果，决定这 n 个样本（或指标）应当分为几类。

二、类间距离的定义

细心的读者已经发现，单个样本与样本的距离是很好计算的，只要运用欧几里德等度量方式即可求出两者之间的距离。但如果一个类包含 n 个样本，另一个类包含 m 个不同的样本，那么这两个类之间的距离如何确定呢？这种类与类之间的距离的定义导致了多种系统聚类方式，下面一一介绍。

1. 最短距离法

最短距离法将两个变量间的距离定义为一个类中所有个体与另一个类中所有个体间距离的最小者。

设 x_i 为群 G_p 中的任一个体，y_i 为群 G_q 中的任一个体，d_{ij}表示个体 x_i 与 y_j 间的距离，D_{pq}表示群 G_p 与群 G_q 间的距离，则最短距离法把群间距离 D_{pq}定义为：

$$D_{pq} = \min_{x_i \in G_p, y_j \in G_q} d_{ij}$$

最短距离法简单易用，能直观地说明聚类的含义，但是它易将大部分个体聚成一个类，易形成延伸的链状结构，且受异常值影响较大，所以最短距离法的聚类效果并不好，实际中较少应用。

2. 最长距离法

最长距离法将两变量间的距离定义为一个类中所有个体与另一个类中所有个体间的距离的最大者，即：

$$D_{pq} = \max_{x_i \in G_p, y_j \in G_q} d_{ij}$$

最长距离法克服了最短距离法连接聚合时的缺陷，但是当数据有较大的离散程度时，易产生较多类。与最短距离法一样，最长距离法受异常值影响较大。

3. 未加权的类间平均法

未加权的类间平均法将变量间的距离定义为一个类 m 中所有个体与另一个类 k 中所有个体间距离的平均值，即：

$$D_{pq} = \frac{\sum_{x_i \in G_p} \sum_{y_j \in G_q} d_{ij}}{n_p n_q}$$

式中，n_p，n_q 分别为类 G_p 和类 G_q 的样本个数。

类间平均法充分利用已知信息，考虑了所有的个体，克服了最短（长）距离法受异常值影响较大的缺陷，是一种聚类效果较好、应用较广的聚类方法。

4. 加权的类间平均法

加权的类间平均法将各自群中的规模作为权数，其余与未加权的类间平均法相同。当群间的信息变异性较大时，加权的类间平均法比未加权的平均法更优。

5. 未加权的类间重心法

从物理学的角度看，一个类用它的重心（该类中个体的均值）来代表是比较合理的。未加权的类间重心法就是将变量间的距离定义为两类重心间的距离。设类 G_p、类 G_q 的重心分别为 $\bar{x}_p$ 和 $\bar{x}_q$，则两群间的距离为 $D_{pq} = \text{distance}(\bar{x}_p, \bar{x}_q)$。

重心法要求使用欧氏距离，每聚一次类，都要重新计算重心。它较少受到异常值的影响，但因为类间距离没有单调递增趋势，在树状聚类图（聚类的重要结果之一，后文会详细讲述）上可能出现图形逆转，限制了它的使用。

6. 离差平方和法

“好”的聚类方法是使类内的差异尽量小，而不同类之间的差异尽量大，也就是说，类内的离差平方和尽量小，类间的离差平方和尽量大。当类数固定时，使整个类内离差平方和达到极小的分类即为最优。它要求采用平方欧氏距离。以前由于计算烦琐限制了它的应用，现在随着计算机技术的发展，计算已不再困难，离差平方和法被认为是一种理论上和实际上都非常有效的聚类方法，应用较为广泛。该方法包含以下步骤：

第一步，每个样品自成一类。

第二步，计算离差平方和增量，类与类合并后增加的离差平方和为：

$$d_{ij}^2 = \frac{n_i n_j}{n_i + n_j}(x_i - x_j)'(x_i - x_j)$$

选择 d_{ij}最小的先合并，其中 n_i 为类 i 中所含的样本个数。

第三步，设类 i 与类 j 合并为新类 k，则原类 m 与新类 k 合并后增加的离差平方和为：

$$d_{mk}^2 = \frac{n_m n_k}{n_m + n_k}(\bar{x}_m - \bar{x}_k)'(\bar{x}_m - \bar{x}_k)$$

第四步，直至全部合并为一个类为止。

三、实例分析

系统聚类法是最常用的聚类分析方法之一，本节中将运用软件 SPSS18.0 对下面的实际问题进行分析。

表 10-2 是主要叶菜类食物的营养成分表（每 100 克食物所含的成分），现需根据表中所列 8 项指标对这 10 种叶菜类食物进行分类。

表 10-2　主要叶菜类食物营养成分表

食物名称	蛋白质（克）	脂肪（克）	碳水化合物（克）	热量（千卡）	无机盐类（克）	钙（毫克）	磷（毫克）	铁（毫克）
黄花（金针菜）	14.1	0.4	60	300	7.0	463	173	16.5
菠菜	2.0	0.2	2.0	18	2.0	70	34	2.5
韭菜	2.4	0.5	4	30	0.9	56	45	1.3
苋菜	2.5	0.4	5	34	2.3	200	46	4.8
油菜（胡菜）	2.0	0.1	4	25	1.4	140	52	3.4
大白菜	1.4	0.3	3	19	0.7	33	42	0.4
小白菜	1.1	0.1	2	13	0.8	86	27	1.2
洋白菜（椰菜）	1.3	0.3	4	24	0.8	100	56	1.9
香菜（芫荽）	2.0	0.3	7	39	1.5	170	49	5.6
芹菜茎	2.2	0.3	2	20	1.0	160	61	8.5

（一）生成 SPSS 数据（.sav 数据）

在 SPSS 中选择 File→Read Text Data 命令，读入.txt 文件，存为.sav 文件，具体操作步骤如图 10-5 至图 10-14 所示。

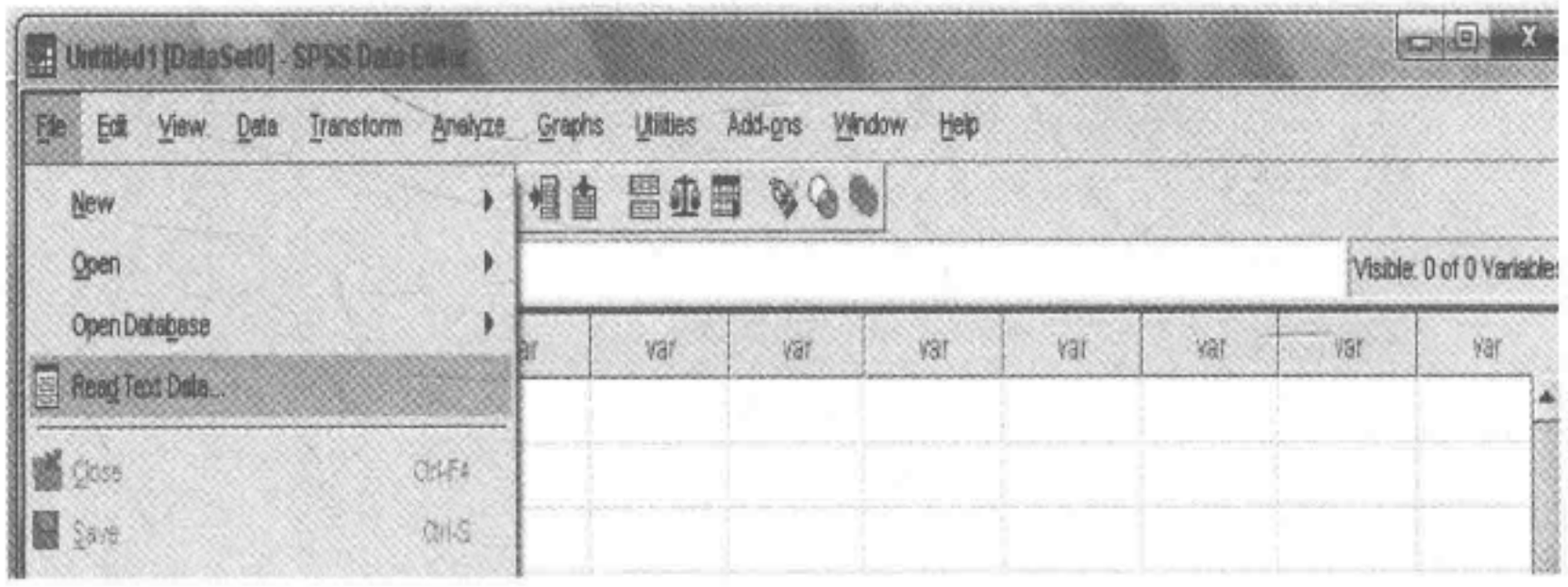

图 10-5　SPSS 数据生成操作步骤 1

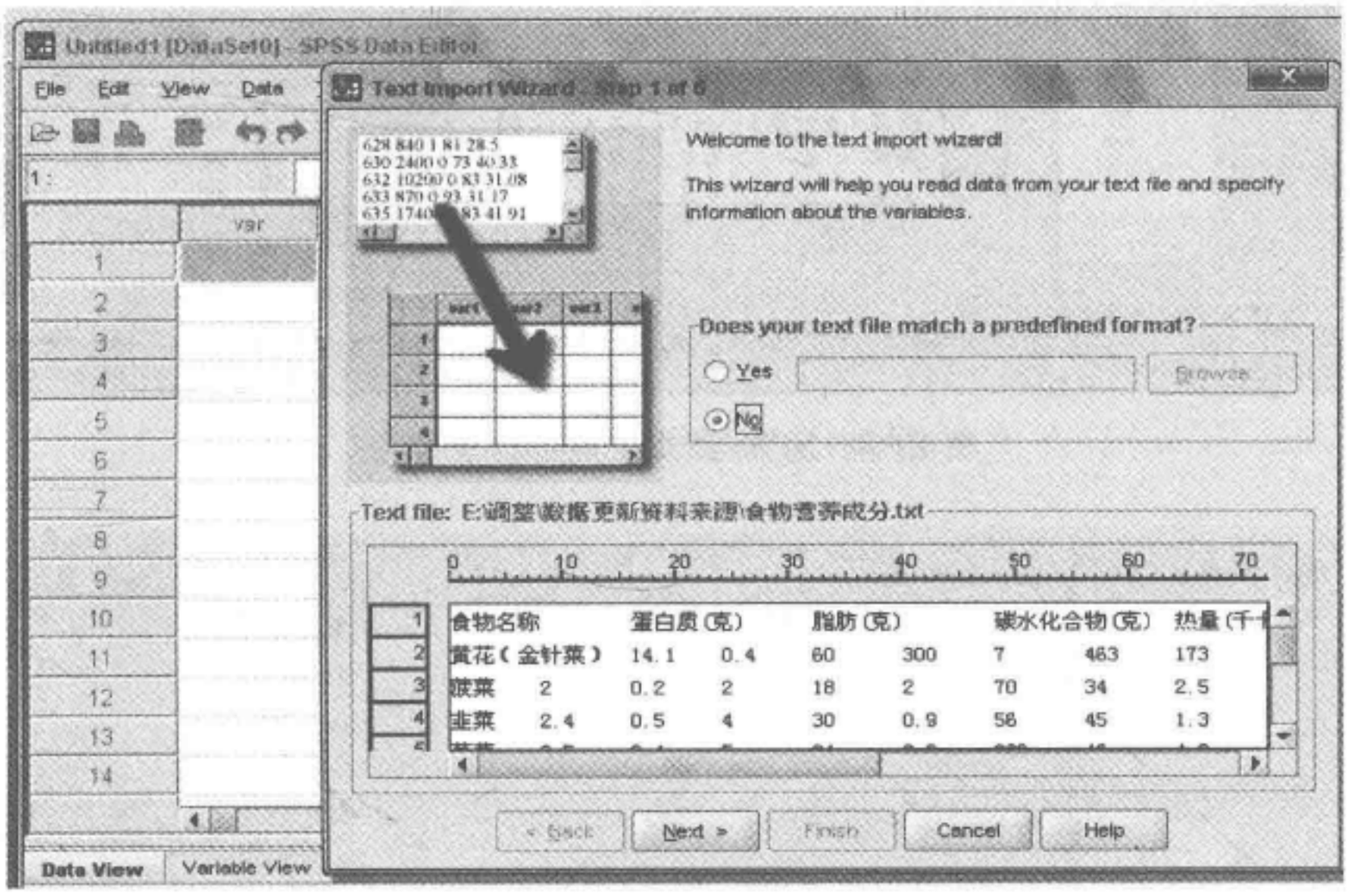

图 10-6　SPSS 数据生成操作步骤 2

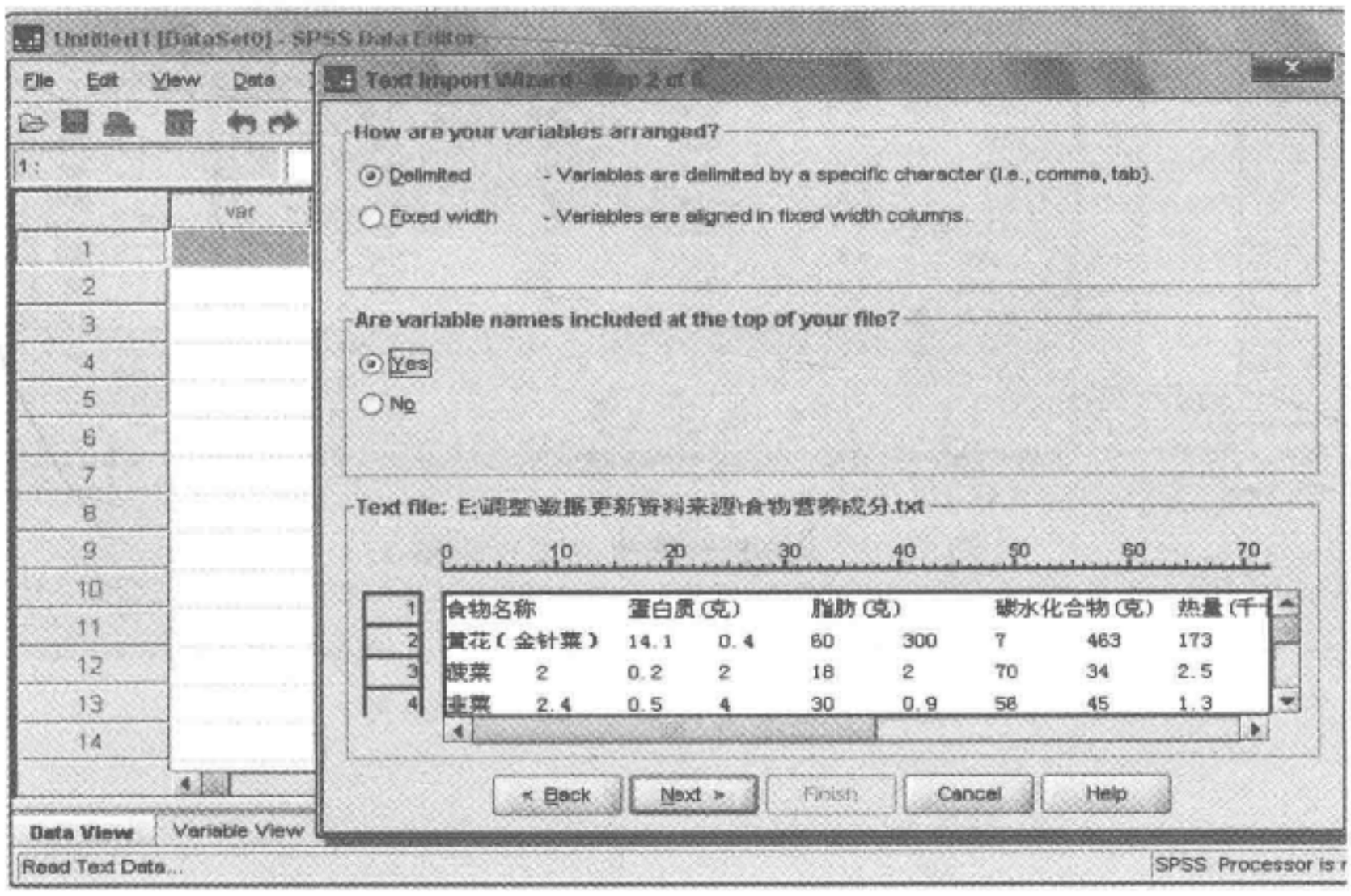

图 10-7　SPSS 数据生成操作步骤 3

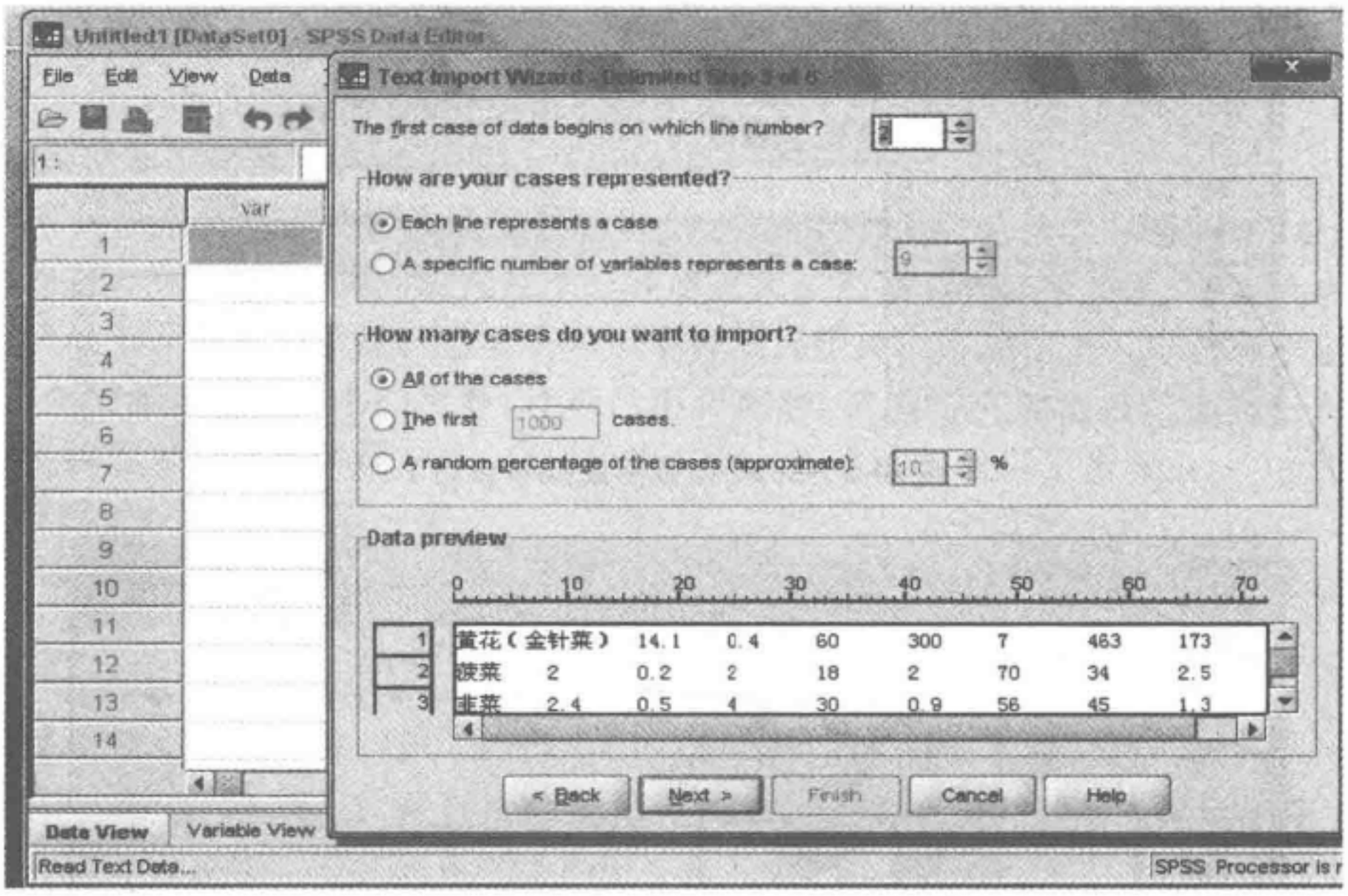

图 10-8　SPSS 数据生成操作步骤 4

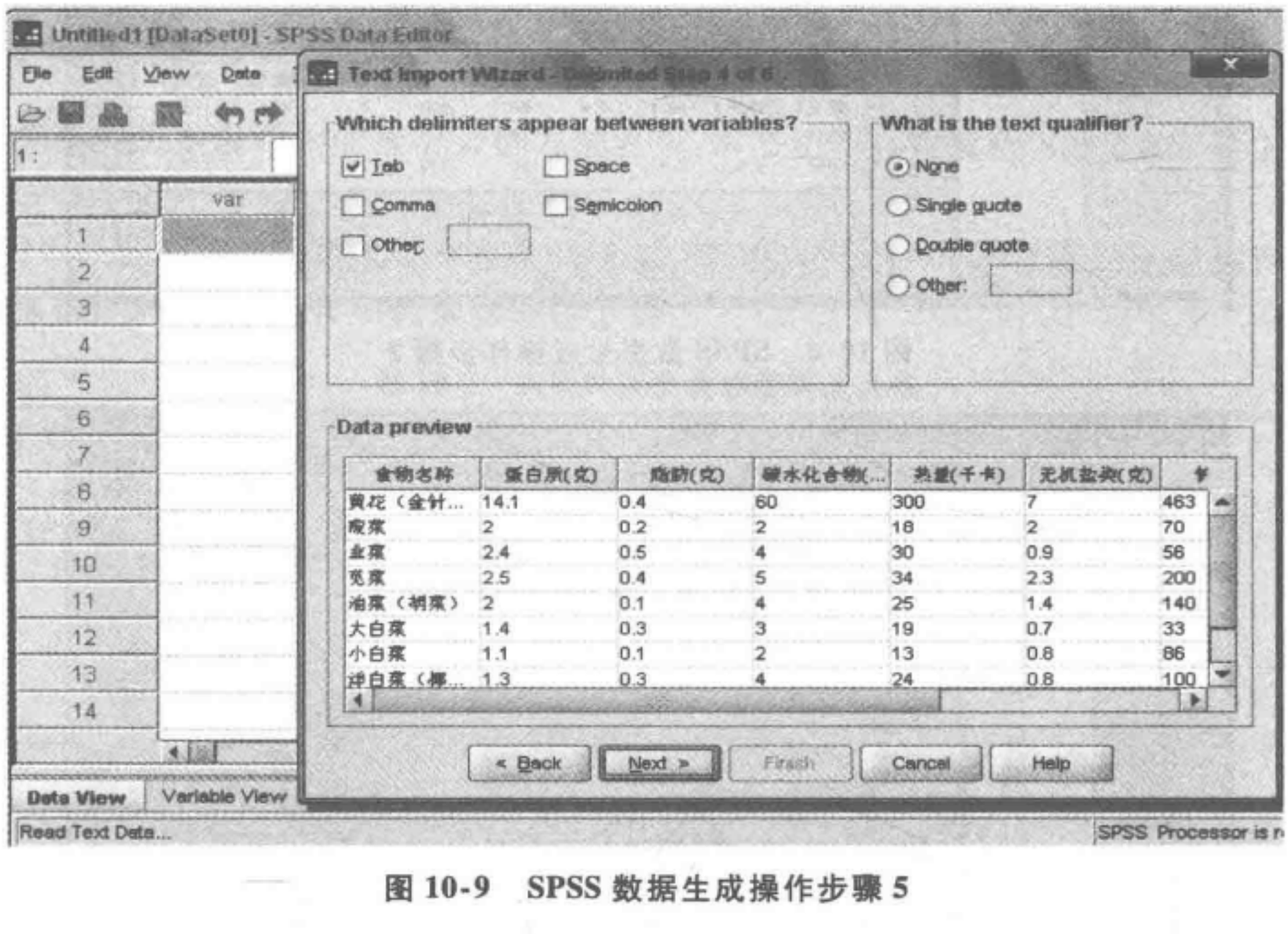

图 10-9　SPSS 数据生成操作步骤 5

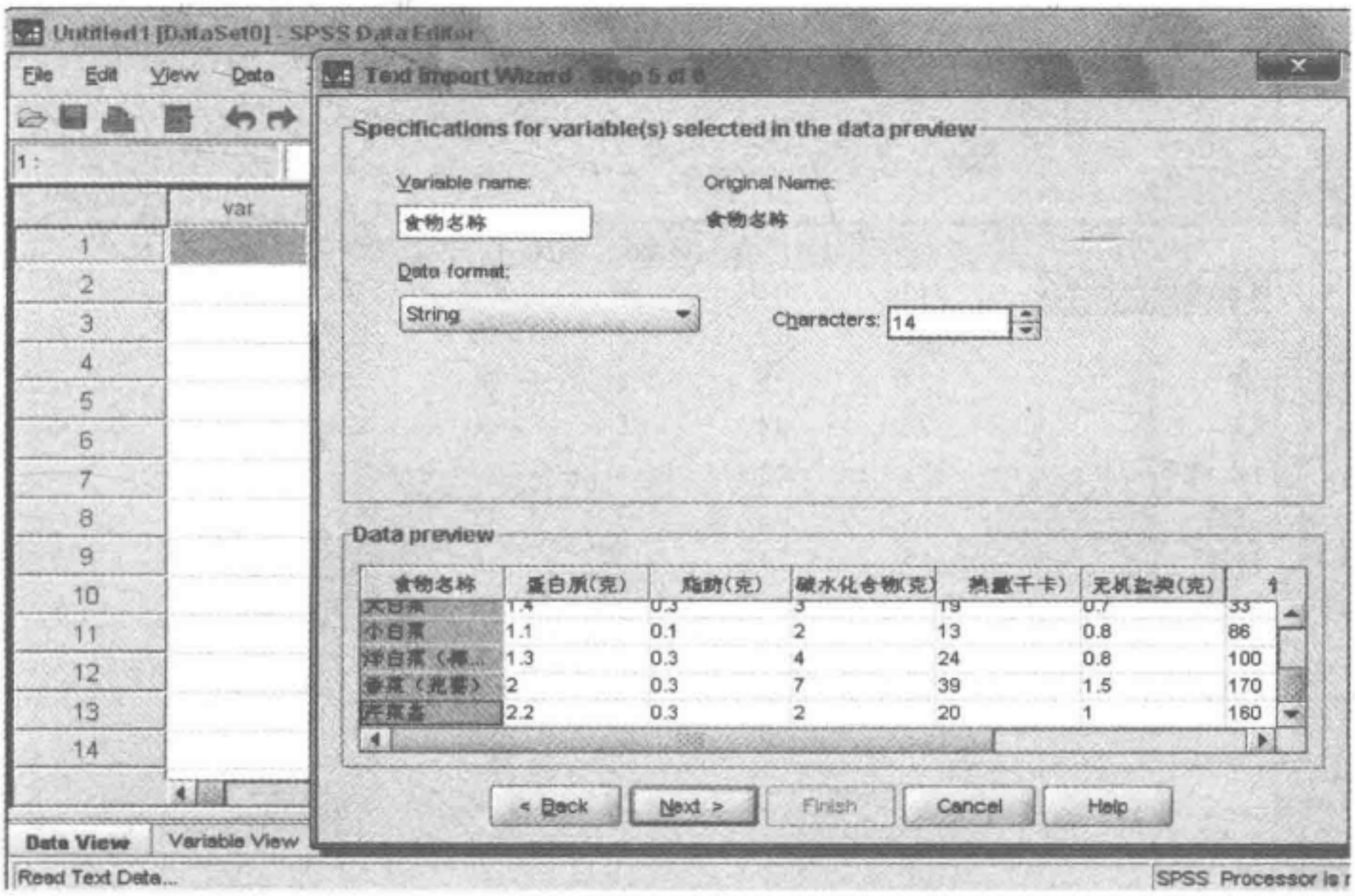

图 10-10　SPSS 数据生成操作步骤 6

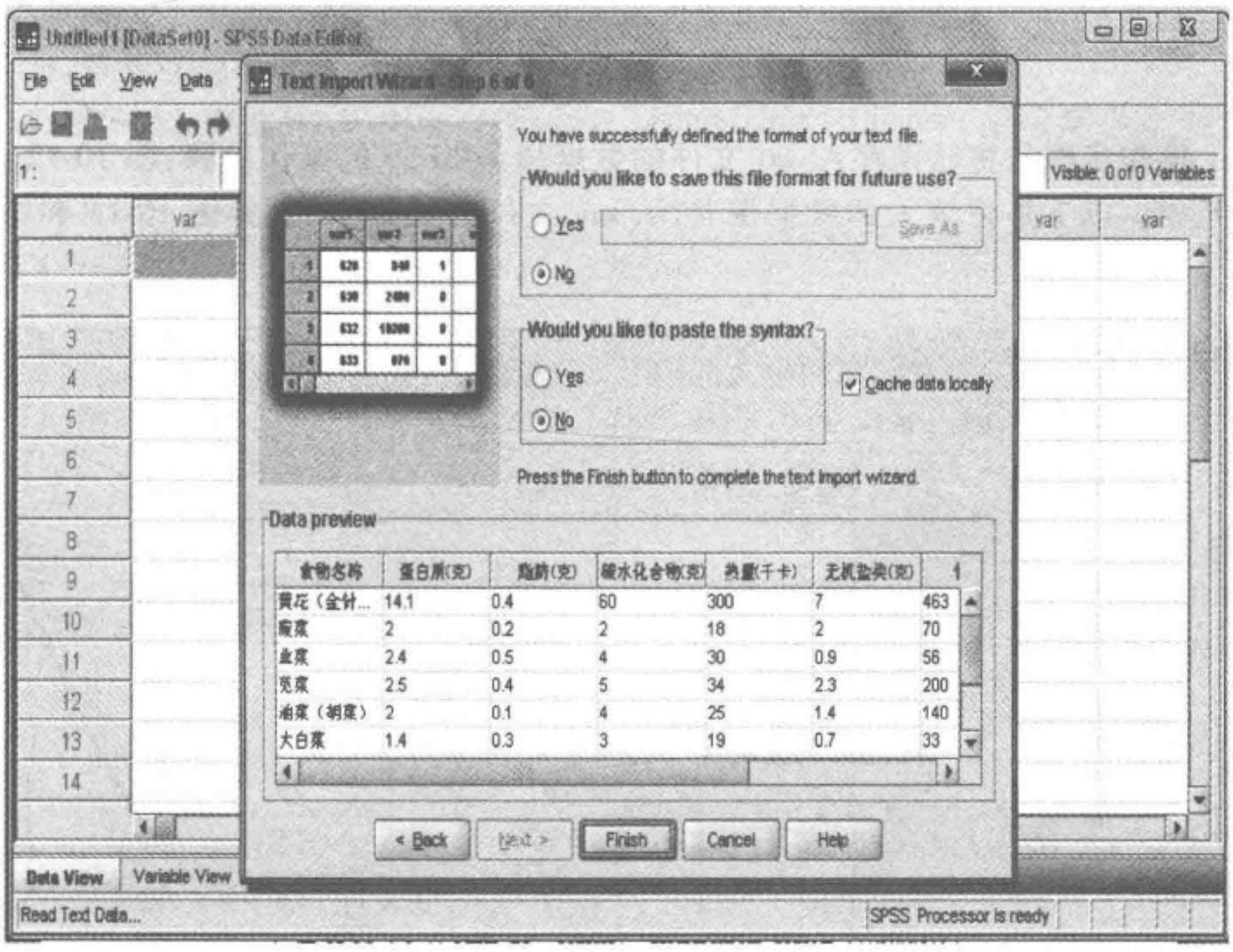

图 10-11　SPSS 数据生成操作步骤 7

*Untitled2 [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : 食物名称 | 黄花（金针菜） | Visible: 9 of 9 Variables

	食物名称	蛋白质(克)	脂肪(克)	碳水化合物(克)	热量(千卡)	无机盐类(克)	钙(毫克)	磷(毫克)
1	黄花（金针菜）	14.1	0.4	60	300	7.0	463	173
2	菠菜	2.0	0.2	2	18	2.0	70	34
3	韭菜	2.4	0.5	4	30	0.9	56	45
4	苋菜	2.5	0.4	5	34	2.3	200	46
5	油菜（胡菜）	2.0	0.1	4	25	1.4	140	52
6	大白菜	1.4	0.3	3	19	0.7	33	42
7	小白菜	1.1	0.1	2	13	0.8	86	27
8	洋白菜（卷菜）	1.3	0.3	4	24	0.8	100	56
9	香菜（芫荽）	2.0	0.3	7	39	1.5	170	49
10	芹菜茎	2.2	0.3	2	20	1.0	160	61
11								
12								
13								
14								

Data View | Variable View

SPSS Processor is ready

图 10-12　SPSS 数据生成操作步骤 8

以上操作步骤是将已经存入.txt 文件的数据读入 SPSS 的操作步骤，图 10-12 是最终读入的数据。以下是将读入的数据保存为.sav 文件的操作步骤，如图 10-13 和图 10-14 所示。

叶菜类营养成分.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

New
Open
Open Database...
Read Text Data...
Close Ctrl-F4
Save Ctrl-S
Save As...
Save All Data
Export to Database...
Mark File Read Only

Visible: 9 of 9 Variable

脂肪(克)	碳水化合物(克)	热量(千克)	无机盐类(克)	钙(毫克)	磷(毫克)
0.4	60	300	7.0	463	173
0.2	2	18	2.0	70	34
0.5	4	30	0.9	56	45
0.4	5	34	2.3	200	46
0.1	4	25	1.4	140	52
0.3	3	19	0.7	33	42
0.1	2	13	0.8	86	27

图 10-13　将读入数据保存为.sav 文件步骤 1

在图 10-14 中单击 Save 即可生成.sav 文件。

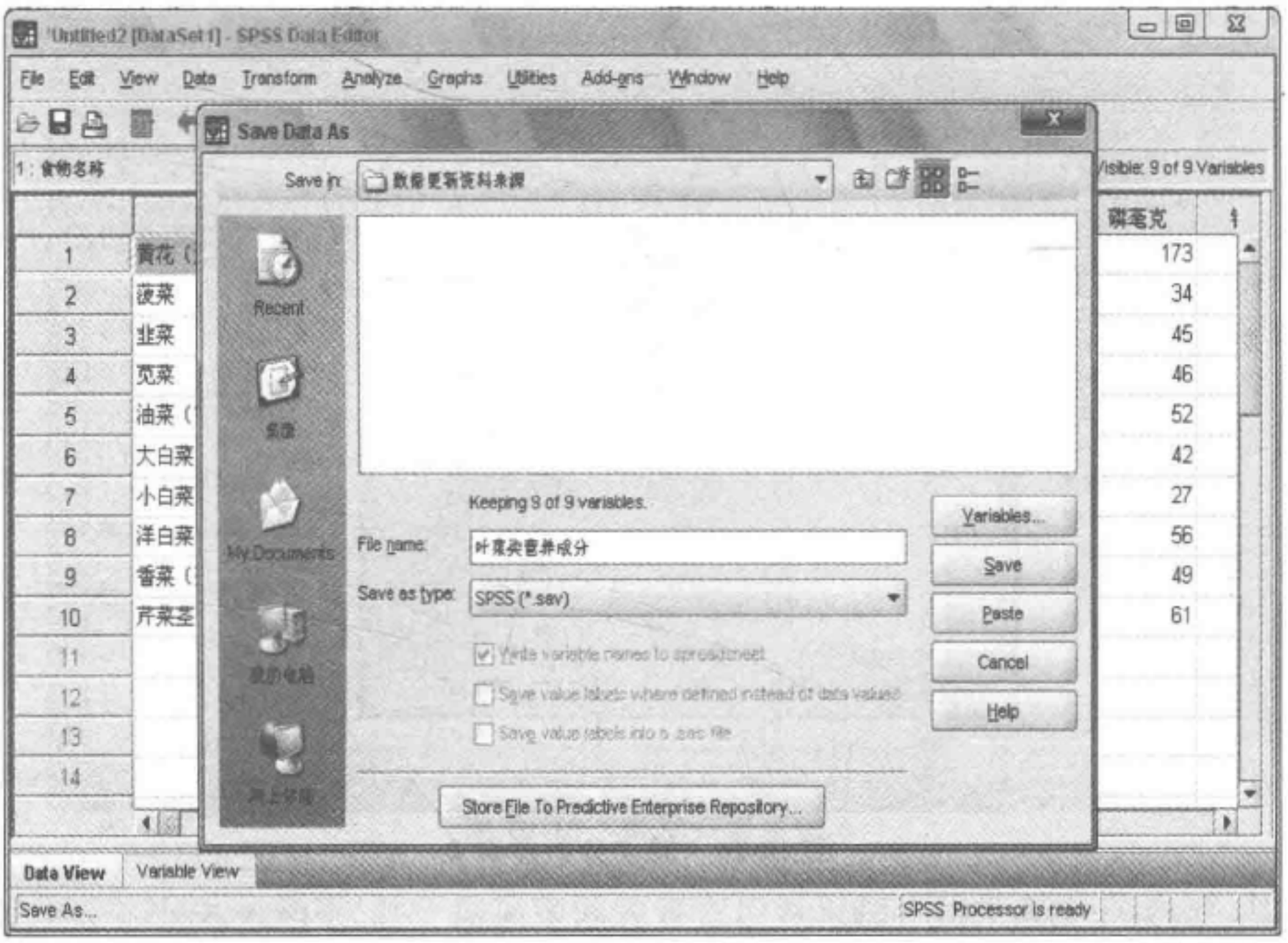

图 10-14　将读入数据保存为.sav 文件步骤 2

（二）对既有数据进行聚类分析

（1）在 SPSS 中选择 Analyze→Classify→Hierarchical Cluster（系统聚类），如图 10-15 所示。

图 10-15　聚类分析步骤 1

将“蛋白质”“脂肪”等 8 个变量选入 Variables，将“食物名称”选入 Label Cases by，如

图 10-16 所示。

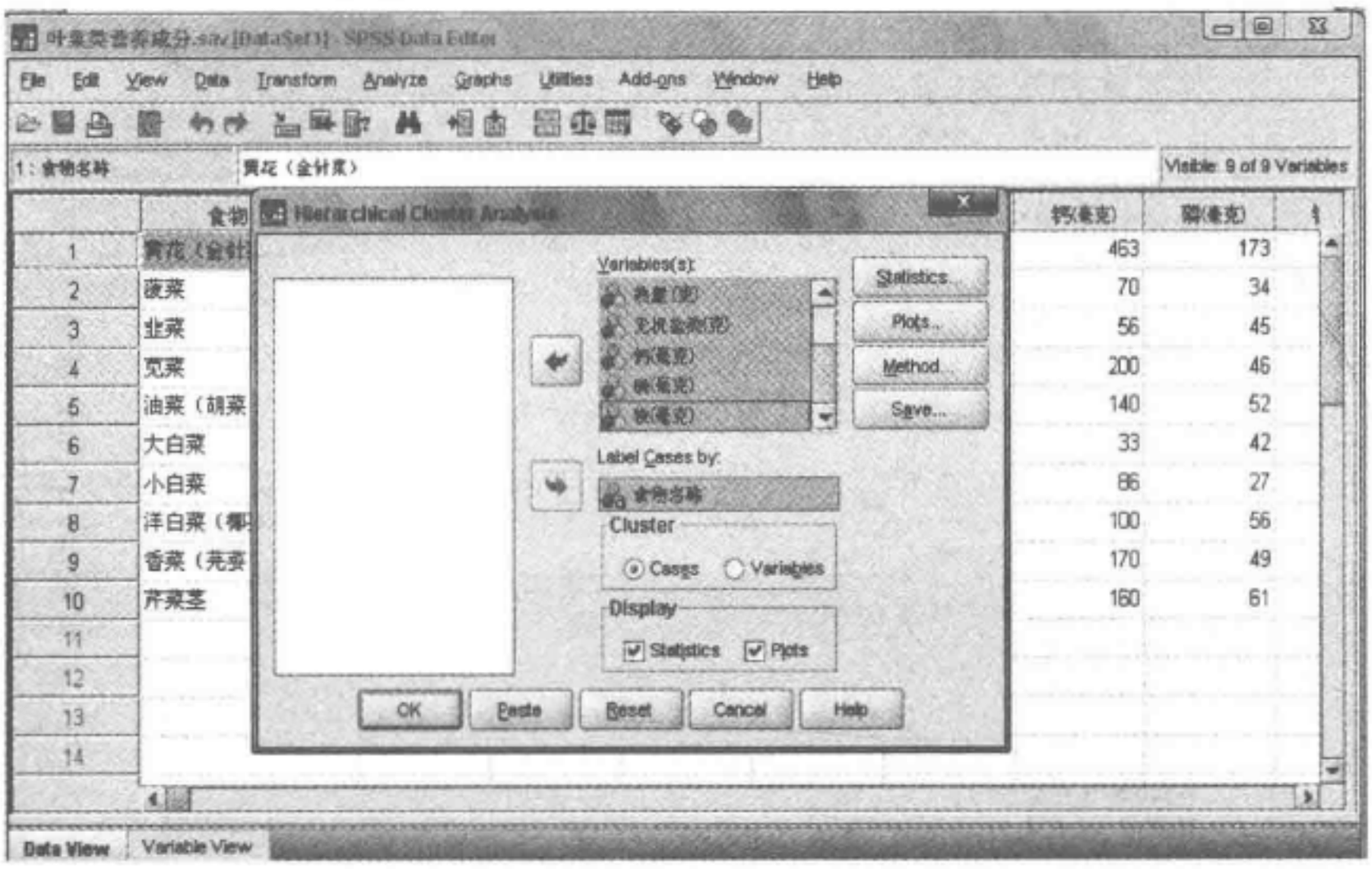

图 10-16　聚类分析步骤 2

在上面的对话框中，我们看到 Cluster 下有两个选项：Cases（样品聚类或 Q 型聚类）和 Variables（变量聚类或 R 型聚类）。在这里，我们选择对样品进行聚类。

（2）单击 Statistics 选项，选 Range of solutions，在本例中我们拟将 10 种叶菜类食物分为两类、三类、四类，则在相应的文本框中分别填入 2 和 4，然后单击 Continue 选项，如图 10-17 所示。

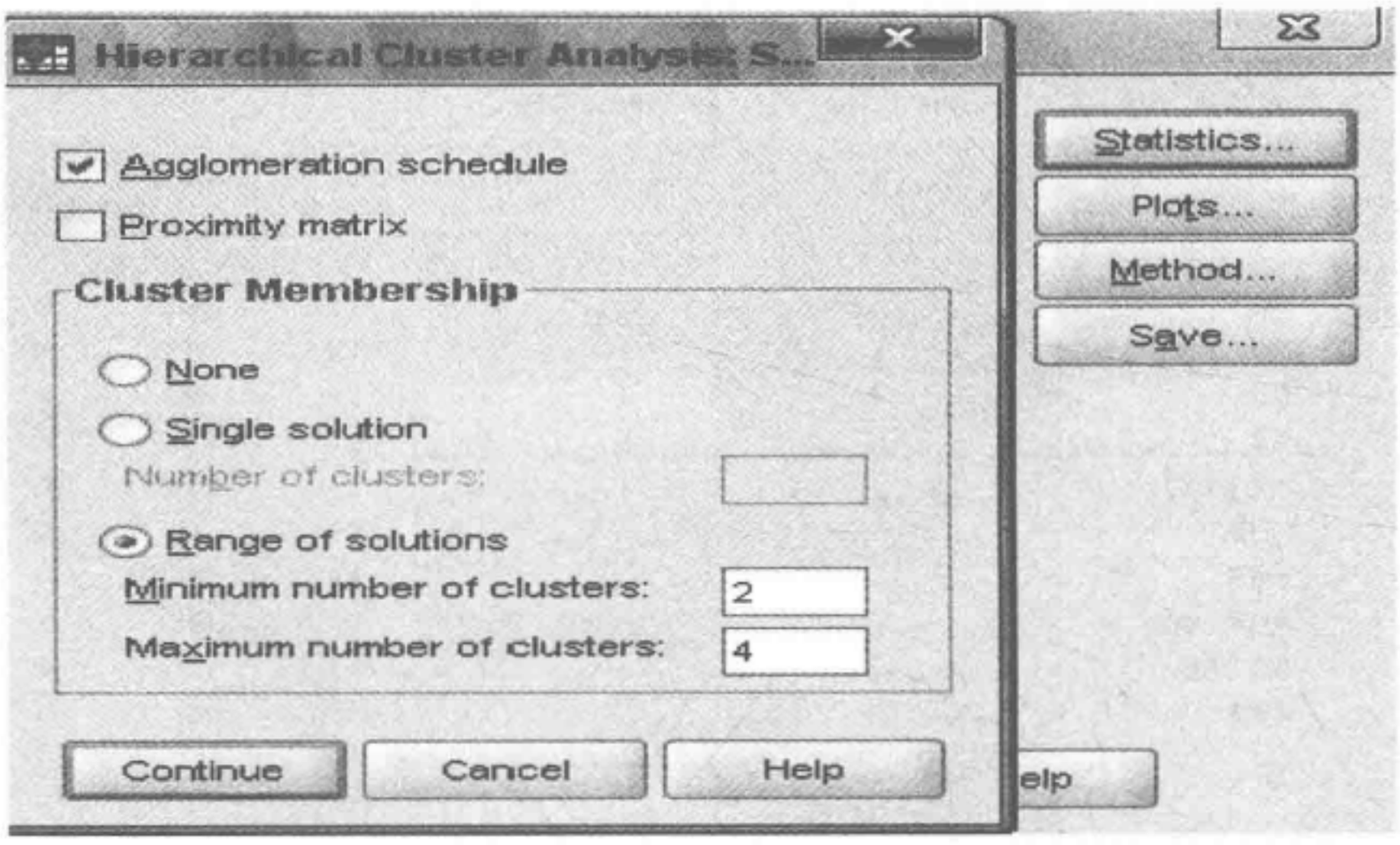

图 10-17　聚类分析步骤 3

（3）单击 Plots 选项，选择 Dendrogram 选项，然后单击 Continue 选项，如图 10-18 所示。

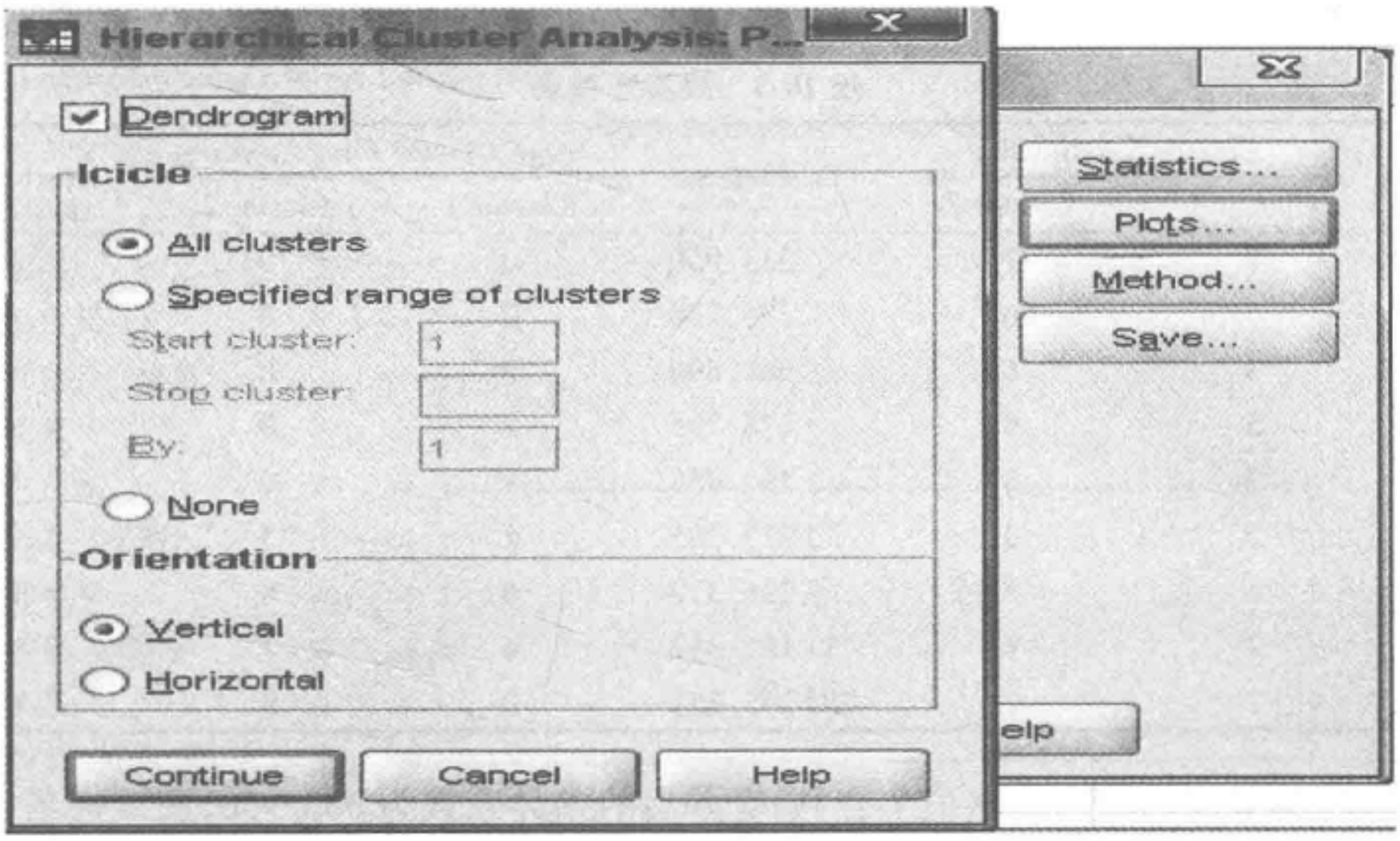

图 10-18　聚类分析步骤 4

（4）单击 Method 选项，对于该项中的内容我们可以使用系统默认的内容，如图 10-19 所示。

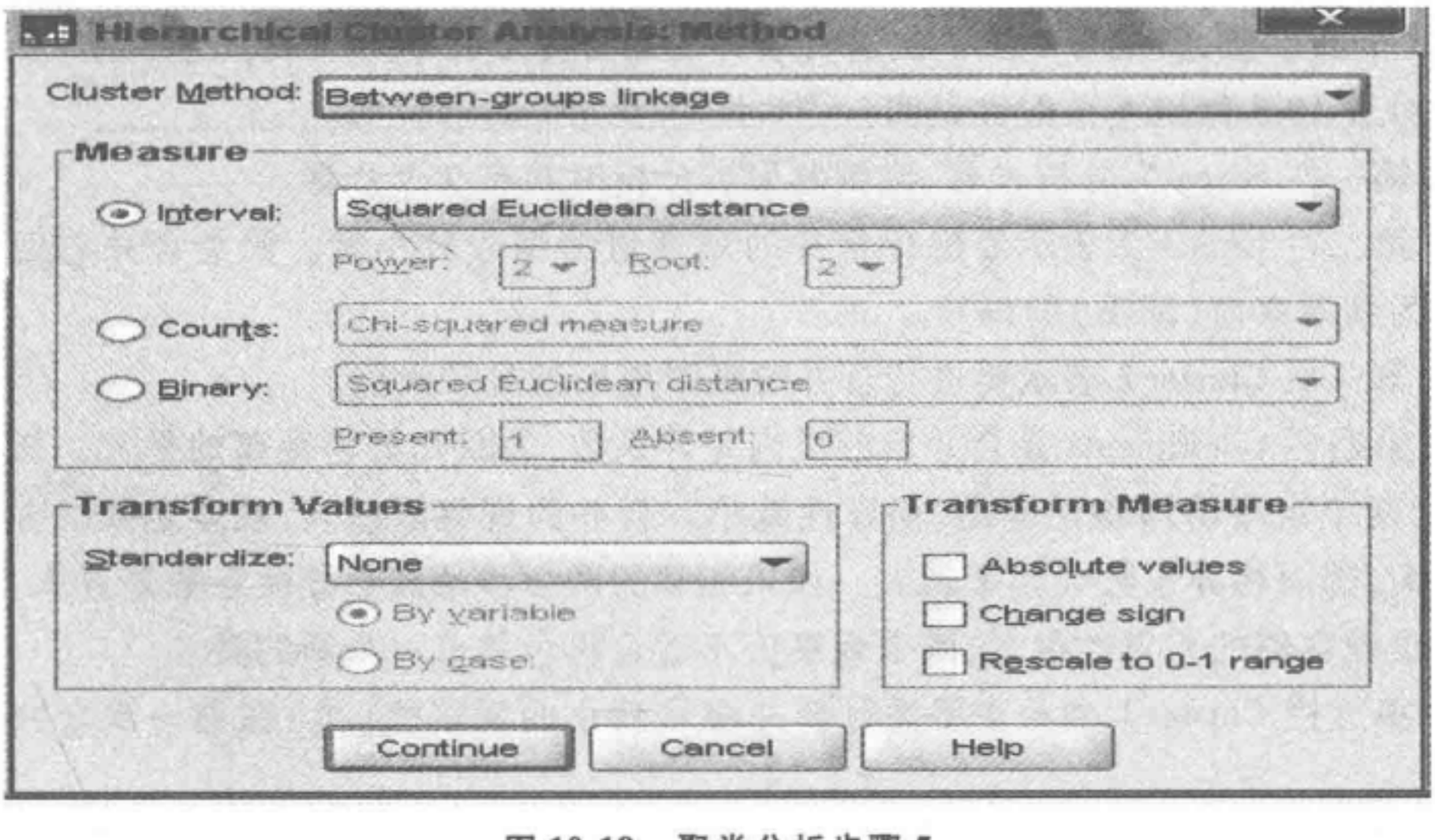

图 10-19　聚类分析步骤 5

（5）在图 10-16 所示的对话框中单击“OK”即可输出分析结果。

（三）聚类结果分析

SPSS18.0 的聚类分析提供了两类展示结果。

1. 聚类过程

这部分是聚类分析的重要结果，它清楚地展示了聚类的过程，直观地给出了各种分类。包含三类结果：

（1）聚类过程表(见表 10-3)

表 10-3　聚类过程表

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	7	333.950	0	0	5
2	5	10	536.250	0	0	4
3	3	6	661.890	0	0	6
4	5	9	878.795	2	0	7
5	2	8	1 294.435	1	0	6
6	2	3	2 217.985	5	3	8
7	4	5	2 235.377	0	4	8
8	2	4	11 161.312	6	7	9
9	1	2	221 255.331	0	8	0

步骤 1 为观察值 2 与观察值 7 合并,合并后的差异系数为 333.950,新观察值(新类)的编号为 2(内有编号成员 2、7)。下次进行合并的地方为步骤 5。

在步骤 5 中,观察值 2 与观察值 8 合并,合并后的差异系数为 1 294.435。其中,观察值 2 前一次合并时出现的地方为步骤 1,观察值 8 前一次合并时出现的地方为步骤 0,2 个观察值合并后的新编号为 2(以编号较小者作为新观察值),在新编号 2 类中,包含编号 2、7、8 共 3 个观察值,新编号 2 类下次进行合并的地方为步骤 6,以下可同理类推。

现对表 10-3 做以下一般性说明:

① 第一栏 Stage 为分析步骤,该例题聚类分析时共有 9 个步骤。

② 第二栏 Cluster 1 表示要进行合并的观察值的编号较小者。聚类合并以后编号较小者作为新观察值(新类)的编号。

③ 第三栏 Cluster 2 表示要进行合并的观察值的编号较大值。

④ 第四栏 Coefficients 是合并后的类内差异系数,为欧几里德距离的平方。该数值越小,表示两个观察值同构性越高,相异性越小。合并的观察值越多,观察之间的差异性会越来越大,因而相异系数会越来越大。如果相邻的两个步骤其相异性系数差异太大,则说明新类中观察值的差异性很大,两个观察值不适合再合并成一个新的类。

⑤ 第五栏 Cluster 1 表示正要进行合并编号较小的观察值(类)在前一次合并时出现的步骤。

⑥ 第六栏 Cluster 2 表示正要进行合并编号较大的观察值(类)在前一次合并时出现的步骤。

（2）树状聚类图

树状聚类图的横轴为距离,纵轴为各个对象(即初始小类),如图 10-20 所示。

菠菜和小白菜两类之间的距离最短,它们首先聚在一起;直至最后,由洋白菜(椰菜)和香菜(芫荽)聚在一起的类与黄花(金针菜)聚成一大类。

至此,系统聚类过程完成。由于树状聚类图能直观明确地展示聚类的过程,所以在实际问题的分析中得到了广泛应用。

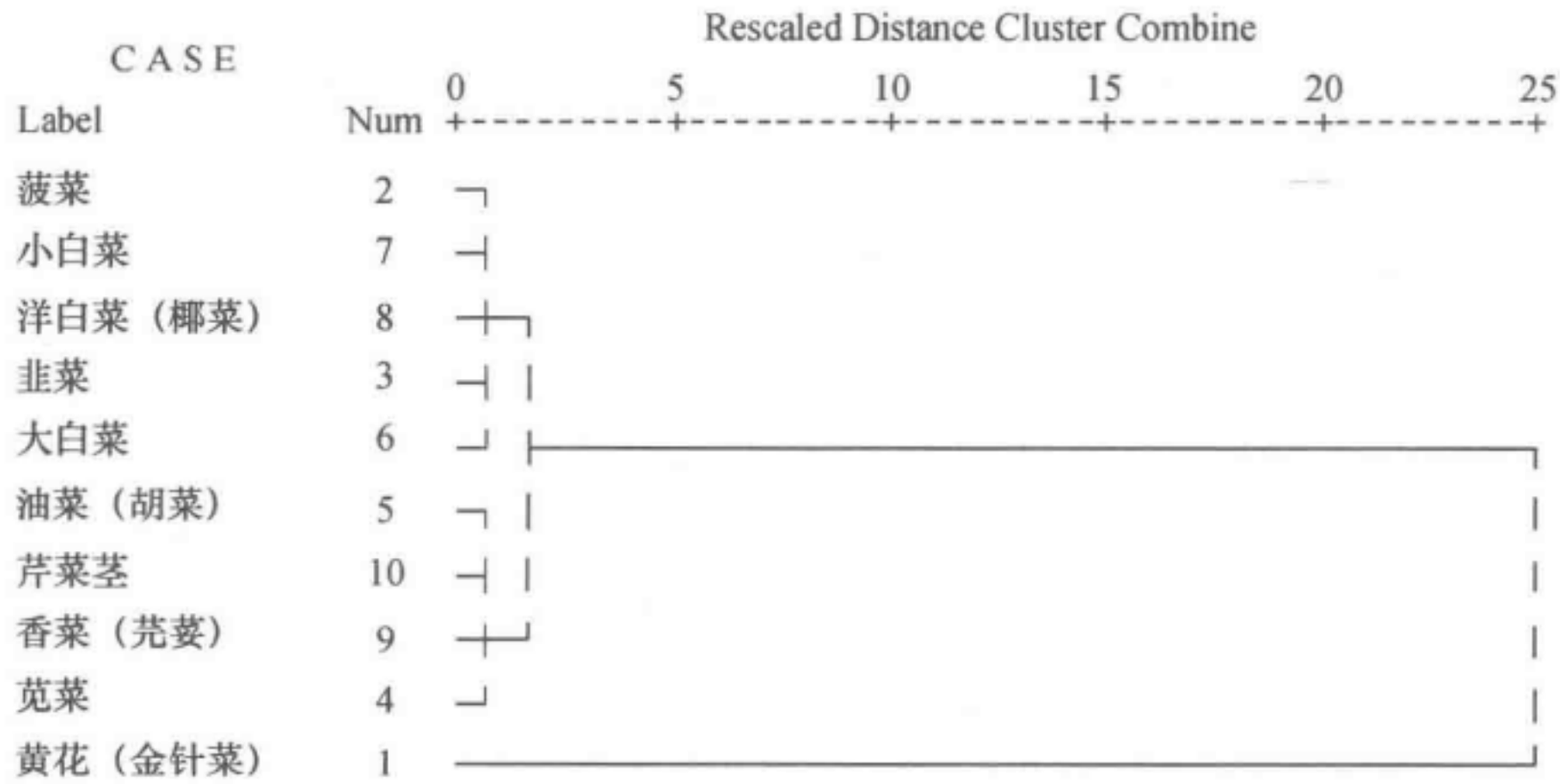

图 10-20　树状图

（3）横向冰柱图

横向冰柱图与树状聚类图类似，也是一种聚类结果的展示形式，两者只是坐标轴不同而已。横向冰柱图的纵轴为各个对象，横轴为距离，结果的说明和树状聚类图相同，如图 10-21 所示。

Horizontal Icicle

Case	Number of clusters								
	1	2	3	4	5	6	7	8	9
9:香菜(芫荽)	×	×	×	×	×	×	×	×	×
	×	×	×	×	×	×			
10:芹菜茎	×	×	×	×	×	×	×	×	×
	×	×	×	×	×	×	×	×	
5:油菜(胡菜)	×	×	×	×	×	×	×	×	×
	×	×	×						
4:苋菜	×	×	×	×	×	×	×	×	×
	×	×							
6:大白菜	×	×	×	×	×	×	×	×	×
	×	×	×	×	×	×	×		
3:韭菜	×	×	×	×	×	×	×	×	×
	×	×	×	×					
8:洋白菜(椰菜)	×	×	×	×	×	×	×	×	×
	×	×	×	×	×				
7:小白菜	×	×	×	×	×	×	×	×	×
	×	×	×	×	×	×	×	×	×
2:菠菜	×	×	×	×	×	×	×	×	×
	×								
1:黄花(金针菜)	×	×	×	×	×	×	×	×	×

图 10-21　横向冰柱图

2. 聚类结果

最终的聚类结果如表 10-4 所示。

表 10-4　聚类结果表

Case	4 Clusters	3 Clusters	2 Clusters
1:黄花(金针菜)	1	1	1
2:菠菜	2	2	2
3:韭菜	2	2	2
4:苋菜	3	3	2
5:油菜(胡菜)	4	3	2
6:大白菜	2	2	2
7:小白菜	2	2	2
8:洋白菜(椰菜)	2	2	2
9:香菜(芫荽)	4	3	2
10:芹菜茎	4	3	2

从表 10-4 可以看到,当我们将 10 种叶菜类食物分为两类时,黄花(金针菜)为一类,其余 9 种菜为一类;分为三类时,分别为黄花(金针菜),菠菜、韭菜、大白菜、小白菜、洋白菜(椰菜),苋菜、油菜(胡菜)、香菜(芫荽)、芹菜茎;分为四类时,也可从表 10-4 直接得出分类结果。

第四节　快速聚类法

系统聚类事先不需要确定分几类,而是通过聚类过程一层层进行,最后得出所有可能的类别结果,再根据具体情况确定最后需要的类别。系统聚类法可通过树状图来直观地观察分类结果,但其缺点是计算量较大,对大批量数据的聚类效率不高。本节介绍的快速聚类法也叫 **K-均值聚类**,计算量较小,效率比系统聚类高。

一、快速聚类的基本过程

快速聚类法不是把所有可能的结果都列出来,而是要求分析者事先把分类数指定好,然后确定各聚类中心,再计算出各样本到聚类中心的距离,最后按距离的远近进行分类。快速聚类即 K-均值聚类中的“K”就是指所要指定的类别个数,而“均值”就是聚类中心。本方法的大致步骤如下:

第一步,将变量进行标准化处理;

第二步,确定 K 值,即要分为几类。这是分析问题者自己确定的,在实际情况中,往往是通过对实际问题反复尝试,得到不同分类结果进行比较,最后得到最优的类别数量。

第三步,确定 K 个类别的初始聚类中心,这一步要求在用于聚类的全部样本中,选择 K 个样本作为 K 个类别的初始聚类中心,与确定类别数一样,原始聚类中心的确定也需要分析者根据实际问题和经验来综合考虑。在 SPSS 软件中,是由系统自动指定初始聚类中心的。

第四步,根据确定的 K 个初始聚类中心,依次计算每个样本到 K 个聚类中心的欧式

距离，并根据最短距离原则将所有的样本分别分到 K 个类别中。

第五步，根据所分成的 K 个类别，计算出各类别中每个变量的均值，并以均值点作为新的 K 个聚类中心，再计算各样本到新中心的距离，并重新分类。

第六步，重复上述第五步的内容，直到达到聚类终止条件为止。聚类终止的条件有：① 迭代次数达到指定的最大迭代次数（SPSS 中默认最大迭代次数为 10）；② 新确定的聚类中心与上一次迭代形成的中心点的最大偏移量小于指定的量（SPSS 中默认是 0.02）。

综上可见，K-均值聚类是根据事先确定的 K 个聚类类别反复迭代直到每个样本都分到特定类别中为止。由于类别的数据是人为主观决定的，不同研究者可能有不同的分类结果，因此具体如何分类需要根据实际情况，以及研究者对问题的理解程度、知识经验决定。

二、快速聚类的应用

下面以 31 个地区的城镇居民消费性支出数据为例，来说明 K-均值聚类法的应用。

居民消费在社会经济的持续发展中有着重要的作用，全国各地区居民消费类别有明显差异。本节根据 31 个地区 8 项消费性支出指标数据，利用 K-均值聚类法进行分类，并对结果进行分析。表 10-5 是各地区城镇居民的消费性支出指标数据。

表 10-5　各地区城镇居民家庭平均每人全年消费性支出　　单位：元

地区	食品	衣着	居住	家庭设备用品及服务	医疗保健	交通和通信	教育文化与娱乐服务	其他商品和服务
北京	6 392.90	2 087.91	1 577.35	1 377.77	1 327.22	3 420.91	2 901.93	848.49
浙江	6 118.46	1 802.29	1 418.00	916.16	1 033.70	3 437.15	2 586.09	546.36
天津	5 940.44	1 567.58	1 615.57	1 119.93	1 275.64	2 454.38	1 899.50	688.73
福建	5 790.72	1 281.25	1 606.27	972.24	617.36	2 196.88	1 786.00	499.30
广东	6 746.62	1 230.72	1 925.21	1 208.03	929.50	3 419.74	2 375.96	653.76
河北	3 335.23	1 225.94	1 344.47	693.56	923.83	1 398.35	1 001.01	395.93
山西	3 052.57	1 205.89	1 245.00	612.59	774.89	1 340.90	1 229.68	331.14
吉林	3 767.85	1 570.68	1 344.41	710.28	1 171.25	1 363.91	1 244.56	506.09
黑龙江	3 784.72	1 608.37	1 128.14	618.76	948.44	1 191.31	1 001.48	402.69
河南	3 575.75	1 444.63	1 080.10	866.72	941.32	1 374.76	1 137.16	418.04
甘肃	3 702.18	1 255.69	910.34	597.72	828.57	1 076.63	1 136.70	387.53
青海	3 784.81	1 185.56	923.52	644.01	718.78	1 116.56	908.07	332.49
宁夏	3 768.09	1 417.47	1 181.71	716.22	890.05	1 574.57	1 286.20	500.12
新疆	3 694.81	1 513.42	898.38	669.87	708.16	1 255.87	1 012.37	444.20
内蒙古	4 211.48	2 203.59	1 384.45	948.87	1 126.03	1 768.65	1 641.17	710.37
辽宁	4 658.00	1 586.81	1 314.79	785.67	1 079.81	1 773.26	1 495.90	585.78
山东	4 205.88	1 745.20	1 408.64	915.00	885.79	2 140.42	1 401.77	415.55
西藏	4 847.58	1 158.60	726.59	376.43	385.63	1 230.94	477.95	481.82
广西	4 372.75	926.42	1 166.85	853.59	625.45	1 973.04	1 243.71	328.27
海南	4 895.96	636.14	1 103.76	616.33	579.89	1 805.11	1 004.62	284.90
四川	4 779.60	1 259.49	1 126.65	876.34	661.03	1 674.14	1 224.73	503.11

（续表）

地区	食品	衣着	居住	家庭设备用品及服务	医疗保健	交通和通信	教育文化与娱乐服务	其他商品和服务
云南	4 593.49	1 158.82	835.45	509.41	637.89	2 039.67	1 014.40	284.95
安徽	4 369.63	1 225.56	1 229.64	678.75	737.05	1 356.57	1 479.75	435.62
江西	4 195.38	1 138.84	1 109.82	854.60	524.22	1 270.28	1 179.89	345.66
湖北	4 429.30	1 415.68	1 187.54	867.33	709.58	1 205.48	1 263.16	372.90
湖南	4 322.09	1 277.47	1 182.33	903.81	776.85	1 541.40	1 418.85	402.52
重庆	5 012.56	1 697.55	1 275.96	1 072.38	1 021.48	1 384.28	1 408.02	462.79
贵州	4 013.67	1 102.41	890.75	673.33	546.84	1 270.49	1 254.56	306.24
陕西	4 381.40	1 428.20	1 126.92	723.73	935.38	1 194.77	1 595.80	435.67
上海	7 776.98	1 794.06	2 166.22	1 800.19	1 005.54	4 076.46	3 363.25	1 217.70
江苏	5 243.14	1 465.54	1 234.05	1 026.32	805.73	1 935.07	2 133.25	514.41

资料来源：《中国统计年鉴 2011》。

（一）实验操作

利用 SPSS18.0 软件中“Analyze-Classify-K-Means Cluster Analysis”命令，对城镇居民消费性支出进行动态聚类分析，具体操作如下：

（1）选择[Analyze]→[Classify-K-Means Cluster Analysis]，进入主对话框。

（2）在主对话框中将用于聚类的所有变量选入[Variables]；在[Number of Clusters]下输入想要分类的数目，本题中分类数目为 10。

（3）点击[Iterate]，在[Maximum Iterations]下输入最大迭代次数（SPSS 中默认是 10 次），点击[Continue]回到主对话框；点击[Save]并选择[Cluster membership]，点击[Continue]回到主对话框；点击[Options]并选择[Initial cluster centers]和[ANOVA tables]，点击[Continue]回到主对话框。点击[OK]。

（二）聚类结果

表 10-6 给出的是初始聚类中心。

表 10-6　初始聚类中心

	聚类									
	1	2	3	4	5	6	7	8	9	10
食品	6 392.90	7 776.98	5 940.44	4 381.40	6 746.62	5 243.14	3 052.57	4 847.58	4 211.48	4 372.75
衣着	2 087.91	1 794.06	1 567.58	1 428.20	1 230.72	1 465.54	1 205.89	1 158.60	2 203.59	926.42
居住	1 577.35	2 166.22	1 615.57	1 126.92	1 925.21	1 234.05	1 245.00	726.59	1 384.45	1 166.85
家庭设备用品及服务	1 377.77	1 800.19	1 119.93	723.73	1 208.03	1 026.32	612.59	376.43	948.87	853.59
医疗保健	1 327.22	1 005.54	1 275.64	935.38	929.50	805.73	774.89	385.63	1 126.03	625.45
交通和通信	3 420.91	4 076.46	2 454.38	1 194.77	3 419.74	1 935.07	1 340.90	1 230.94	1 768.65	1 973.04
教育文化与娱乐服务	2 901.93	3 363.25	1 899.50	1 595.80	2 375.96	2 133.25	1 229.68	477.95	1 641.17	1 243.71
其他商品和服务	848.49	1 217.70	688.73	435.67	653.76	514.41	331.14	481.82	710.37	328.27

该表列出了每一类别的初始聚类中心，本例的这些中心是由 SPSS 自动生成的，实则为数据集中的某一条记录。

表 10-7 给出的是迭代过程表。

表 10-7　迭代历史记录[a]

迭代	聚类中心内的更改									
	1	2	3	4	5	6	7	8	9	10
1	410.130	.000	410.702	368.467	.000	.000	552.202	.000	371.498	379.690
2	.000	.000	.000	215.211	.000	.000	150.418	.000	215.943	.000
3	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

a. 由于聚类中心内没有改动或改动较小而达到收敛。任何中心的最大绝对坐标更改为.000。当前迭代为 3。初始中心间的最小距离为 1 052.416。

从表 10-7 中可以看出每次迭代过程中类别中心的变化，随着迭代次数的增加，类别中心点的变化越来越小，本例只用了 3 次就已经收敛了。

表 10-8 给出的是最终聚类中心。

表 10-8　最终聚类中心

	聚类									
	1	2	3	4	5	6	7	8	9	10
食品	6 255.68	7 776.98	5 865.58	4 389.15	6 746.62	5 243.14	3 607.33	4 847.58	4 358.45	4 660.45
衣着	1 945.10	1 794.06	1 424.42	1 326.53	1 230.72	1 465.54	1 380.85	1 158.60	1 845.20	995.22
居住	1 497.68	2 166.22	1 610.92	1 143.28	1 925.21	1 234.05	1 117.34	726.59	1 369.29	1 058.18
家庭设备用品及服务	1 146.96	1 800.19	1 046.09	824.85	1 208.03	1 026.32	681.08	376.43	883.18	713.92
医疗保健	1 180.46	1 005.54	946.50	750.20	929.50	805.73	878.37	385.63	1 030.54	626.07
交通和通信	3 429.03	4 076.46	2 325.63	1 317.61	3 419.74	1 935.07	1 299.21	1 230.94	1 894.11	1 872.99
教育文化与娱乐服务	2 744.01	3 363.25	1 842.75	1 371.43	2 375.96	2 133.25	1 106.36	477.95	1 512.95	1 121.87
其他商品和服务	697.43	1 217.70	594.02	394.49	653.76	514.41	413.14	481.82	570.57	350.31

表 10-8 中的数据表示各个类别在各个变量上的均值，将全国 31 个地区分为 10 个类别，并迭代出各个类别中食品、衣着、居住、家庭设备用品及服务、交通和通信、医疗保健、教育文化与娱乐服务、其他商品和服务的平均值。

表 10-9 给出的是分类后各个变量在类别之间的方差分析表。

表 10-9　方差分析

	聚类		误差		*F*	Sig.
	均方	df	均方	df		
食品	3 687 517.862	9	68 006.380	21	54.223	.000
衣着	236 836.524	9	47 164.828	21	5.021	.001
居住	264 507.313	9	20 610.833	21	12.833	.000

（续表）

	聚类		误差		F	Sig.
	均方	df	均方	df		
家庭设备用品及服务	208 624.103	9	19 427.288	21	10.739	.000
医疗保健	95 440.104	9	31 574.250	21	3.023	.018
交通和通信	1 992 777.239	9	23 243.581	21	85.735	.000
教育文化与娱乐服务	1 232 493.220	9	19 173.069	21	64.283	.000
其他商品和服务	97 620.178	9	9 048.226	21	10.789	.000

注：F 检验应仅用于描述性目的，因为选中的聚类将被用来最大化不同聚类中的案例间的差别。观测到的显著性水平并未据此进行更正，因此无法将其解释为是对聚类均值相等这一假设的检验。

表 10-9 是对聚类结果的类别间距离进行方差分析，方差分析表明：除了医疗保健的概率值大于 0.001 外，其他类别间距离差异的概率值均小于 0.001，即总体聚类效果良好。

每个地区所属的类别，SPSS 会自动存在数据表中。表 10-10 就是每个地区所属的类别。

表 10-10　聚类成员

案例号	聚类	距离	案例号	聚类	距离
1	1	410.130	17	9	392.240
2	1	410.130	18	8	.000
3	3	410.702	19	10	379.690
4	3	410.702	20	10	469.859
5	5	.000	21	10	435.660
6	7	416.019	22	10	407.949
7	7	627.487	23	4	233.439
8	7	481.582	24	4	408.849
9	7	338.958	25	4	199.338
10	7	227.065	26	4	260.324
11	7	357.848	27	4	830.191
12	7	461.764	28	4	585.631
13	7	385.269	29	4	349.724
14	7	337.555	30	2	.000
15	9	464.148	31	6	.000
16	9	431.887			

根据以上结果，我们可以得出表 10-11 的各地区分类实际情况。

表 10-11　各地区消费结构特征聚类结果

类别	地区	食品	衣着	居住	家庭设备用品及服务	医疗保健	交通和通信	教育文化与娱乐服务	其他商品和服务
1	北京、浙江	6 255.68	1 945.10	1 497.68	1 146.97	1 180.46	3 429.03	2 744.01	697.43
2	天津、福建	5 865.58	1 424.42	1 610.92	1 046.09	946.50	2 325.63	1 842.75	594.02
3	广东	6 746.62	1 230.72	1 925.21	1 208.03	929.50	3 419.74	2 375.96	653.76
4	河北、山西、吉林、黑龙江、河南、甘肃、青海、宁夏、新疆	3 607.33	1 380.85	1 117.34	681.08	878.37	1 299.21	1 106.36	413.14
5	内蒙古、辽宁、山东	4 358.45	1 845.20	1 369.29	883.18	1 030.54	1 894.11	1 512.95	570.57
6	西藏	4 847.58	1 158.60	726.59	376.43	385.63	1 230.94	477.95	481.82
7	广西、海南、四川、云南	4 660.45	995.22	1 058.18	713.92	626.07	1 872.99	1 121.87	350.31
8	安徽、江西、湖北、湖南、重庆、贵州、陕西	4 389.15	1 326.53	1 143.28	824.85	750.20	1 317.61	1 371.43	394.49
9	上海	7 776.98	1 794.06	2 166.22	1 800.19	1 005.54	4 076.46	3 363.25	1 217.70
10	江苏	5 243.14	1 465.54	1 234.05	1 026.32	805.73	1 935.07	2 133.25	514.41

根据聚类分析结果，可以看出各地区消费结构的差异：北京和浙江两地食品支出占消费的比重最大，居住、家庭设备用品及服务、医疗保健的花费比较接近，交通和通信、娱乐服务消费差不多是居住和家庭设备用品及服务支出的 2 倍；天津和福建除了食品支出以外，其他各类消费支出比重都比较接近；广东食品支出高，仅次于上海，其他各项与北京和浙江基本持平；河北等 9 个省区食品支出最小，仅为上海的一半，衣着、居住等其他消费支出也比较小；内蒙古、辽宁、山东三省各项消费支出比较稳定；西藏地区在家庭设备用品及服务、医疗保健、交通和通信、教育文化与娱乐服务的支出均最小，部分支出仅为发达地区的 1/3；广西等 4 个省区在各项消费支出上的比重也比较稳定，但与北京、浙江、广东相比，数值明显较小；安徽等 7 个省区在衣着、交通和通信、教育文化与娱乐服务方面的消费稳定在 1 300 元左右；上海除了在医疗保健上的支出略低于北京和浙江，各项消费支出在各类别中均占最高位；江苏的各项消费支出具有层次，其中衣着、居住、家庭设备用品及服务、医疗保健的支出平均为 1 100 元左右，交通和通信、教育文化与娱乐服务支出为 2 000 元左右，是前者的 2 倍。

第五节　判 别 分 析

一、判别分析概述

判别分析又称“分辨法”，是在分类确定的条件下，根据某一研究对象的各种特征值判别其类型归属问题的一种多元统计分析方法。

其基本原理是按照一定的判别准则，建立一个或多个判别函数，用研究对象的大量资料确定判别函数中的待定系数，并计算判别指标。据此即可确定某一样本属于何类。

在市场调研中，一般根据事先确定的因变量（例如产品的主要用户、普通用户和非用户、自有房屋或租赁、电视观众和非电视观众）找出相应处理的区别特性。在判别分析中，因变量为类别数据，有多少类别就有多少类别处理组；自变量通常为可度量数据。通过判别分析，可以建立能够最大限度地区分因变量类别的函数，考查自变量的组间差异是否显著，判断哪些自变量对组间差异贡献最大，评估分类的程度，根据自变量的值将样本归类。

判别分析从不同的角度可进行不同的分类。本节主要根据判别标准的不同，将判别分析分为距离判别法、Fisher 判别法、贝叶斯判别法，这也是常用的分类法。

二、距离判别法

距离判别法就是以距离为依据实现判别。

设有 k 个判别类别，有分别来自 k 个类别总体的 k 个样本，每个样本是一个 p 维样本，包含 p 个判别变量：$x_1, x_2, \cdots, x_p (p > k)$，且判别变量均为数值型，服从正态分布。

距离判别法的基本思想是根据各样品与各总体之间的距离远近做出判别。其通过建立关于各总体的距离判别函数式，得出各样品与各总体之间的距离值，判别样品属于距离值最小的那个总体。

这里以两个总体为例来说明距离判别的原理。对于多个总体，需要两两判别，方法与两个总体相同。设有两个总体 G_1 和 G_2，x 是一个 p 维样品，定义 x 到 G_1、G_2 的距离分别为 $d(x, G_1)$ 和 $d(x, G_2)$。此处，距离选用马氏距离，定义如下：

$$\begin{aligned} d^2(x, G_1) &= (x - \mu_1)' \sum\nolimits_1^{-1} (x - \mu_1) \\ d^2(x, G_2) &= (x - \mu_2)' \sum\nolimits_2^{-1} (x - \mu_2) \end{aligned} \tag{10-1}$$

式(10-1)中，μ_1、$\sum_1$ 为总体 G_1 的均值向量和协方差阵，μ_2、$\sum_2$ 为总体 G_2 的均值向量和协方差阵。当总体均值未知时，可选用样本均值作为估计值。

显然，马氏距离是点 x 到各类别中心的平方欧式距离，以判别变量的协方差阵调整后的距离。

于是，根据 $d^2(x, G_1)$ 和 $d^2(x, G_2)$ 进行判断：

$$\begin{cases} x \in G_1, & \text{若 } d^2(x, G_1) < d^2(x, G_2) \\ x \in G_2, & \text{若 } d^2(x, G_1) > d^2(x, G_2) \\ \text{待定}, & \text{若 } d^2(x, G_1) = d^2(x, G_2) \end{cases}$$

进一步，设 $W(x) = d^2(x, G_2) - d^2(x, G_1)$ 为判别函数，则有：

$$\begin{cases} x \in G_1, & \text{若 } W(x) > 0 \\ x \in G_2, & \text{若 } W(x) < 0 \\ \text{待定}, & \text{若 } W(x) = 0 \end{cases}$$

以下对于判别函数的计算，分两种情况考虑：

（1）当各总体的协方差阵相等，即 $\sum_1 = \sum_2 = \sum$ 时

$$d^2(x, G_1) - d^2(x, G_2) = (x - \mu_1)' \sum_1^{-1} (x - \mu_1) - (x - \mu_2)' \sum_2^{-1} (x - \mu_2) \tag{10-2}$$

整理式(10-2)得到如下结果：

$$d^2(x, G_1) - d^2(x, G_2) = -2(x - \bar{\mu})' \sum^{-1} (\mu_1 - \mu_2) = -2W(x) \tag{10-3}$$

所以，经过整理得到判别函数 $W(x)$ 为：

$$W(x) = (x - \bar{\mu})' \sum^{-1} (\mu_1 - \mu_2) \tag{10-4}$$

其中，$\bar{\mu} = (\mu_1 + \mu_2)/2$。

（2）当各总体的协方差阵不相等，即 $\sum_1 \neq \sum_2$ 时，整理后的判别函数 $W(x)$ 为：

$$W(x) = (x - \mu_2)' \sum_2^{-1} (x - \mu_2) - (x - \mu_1)' \sum_1^{-1} (x - \mu_1) \tag{10-5}$$

以上判别过程中，当两个总体的均值差异不显著时，判别分析的错判概率是很大的，所以，只有当两个总体的均值存在显著性差异时，距离判别才有意义。

三、Fisher 判别法

Fisher 判别法的核心思想是投影。将 k 组 p 维数据投影到某个方向，使得它们投影的组和组之间尽可能分开。测量组和组分开程度的方法借用了方差分析的思想。

投影就是将原来 p 维 X 空间样本点投影到 m 维 Y 空间中，Fisher 判别法的判别函数是判别变量的线性形式，即：

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p \tag{10-6}$$

其中，系数 a_i 为判别系数，表示各输入变量对于判别函数的影响，y 是样本在低维空间中的某个维度。

判别函数可能会有多个，所以得到低维空间中的多个维度：$y_1, y_2, \cdots, y_m$。并且，以上线性组合可能有多个解，为了达到好的分类效果，坐标变化的原则是应尽可能找到把来自不同类别的样本尽量分开的方向。

为此，首先，应在判别变量的 p 维空间中，找到某个线性组合，使各类别的平均值差异最大，作为第一维度，代表判别变量组间方差中的最大部分，得到第一判别函数。然后，按照相同办法依次找到相互独立的后续判别函数。

由于每个判别函数都代表判别变量组间方差的一部分，所有判别函数所表示的方差比例之和为 100%。显然，前面的判别函数代表的方差较大，对于分类相对重要，后面的判别函数只代表方差的很小部分，多数情况下可以被忽略。

下面对 Fisher 判别法的计算做简要的说明。

设从 k 个总体分别取得 k 组 p 维观察值如下：

$$\begin{cases} G_1: x_1^{(1)}, \cdots, x_{n_1}^{(1)} \\ \cdots\cdots \\ G_k: x_1^{(k)}, \cdots, x_{n_k}^{(k)} \end{cases} \qquad n = n_1 + \cdots + n_k$$

令 $y(x) = a'x$，为 x 向 y 的投影，上述数据变换为：

$$G_1: a'x_1^{(1)}, \cdots, a'x_{n_1}^{(1)}$$

$$\cdots\cdots$$

$$G_k: a'x_1^{(k)}, \cdots, a'x_{n_k}^{(k)}$$

于是，组间离差平方和为：

$$\mathrm{SSG} = \sum_{i=1}^{k} n_i (a'\bar{x}^{(i)} - a'\bar{x})^2 = a'\Big[\sum_{i=1}^{k} n_i (\bar{x}^{(i)} - \bar{x})(\bar{x}^{(i)} - \bar{x})'\Big]a = a'Ba \quad (10\text{-}7)$$

其中，B 为组间 SSCP 矩阵。

组内离差平方和为：

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (a'x_j^{(i)} - a'\bar{x}^{(i)})^2 = a'\Big[\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_j^{(i)} - \bar{x}^{(i)})(x_j^{(i)} - \bar{x}^{(i)})'\Big]a = a'Ea \quad (10\text{-}8)$$

其中，E 为组内 SSCP 矩阵。

于是，判别分析的目的就是寻找 a，使得 SSG 尽量大，SSE 尽量小，即 $\Delta(a) = \dfrac{a'Ba}{a'Ea} \to \max$。可以证明使得 $\Delta(a)$ 最大的值为方程 $|B - \lambda E| = 0$ 的最大特征值 λ_1。记方程 $|B - \lambda E| = 0$ 的全部特征值为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$，相应的特征向量为 $e_1, e_2, \cdots, e_r$，则判别函数即为 $y_i(x) = e_i'x = a'x$。

记 p_i 为第 i 个判别函数的判别能力，记为 $p_i = \dfrac{\lambda_i}{\sum_{l=1}^{r} \lambda_l}$。所以，前 m 个判别函数的总判别能力为 $\sum_{i=1}^{m} p_i = \dfrac{\sum_{i=1}^{m} \lambda_i}{\sum_{l=1}^{r} \lambda_l}$。

可根据两个标准来决定判别函数的个数：① 指定取特征值大于 1；② 前 m 个判别函数的判别能力达到指定要求。

进行判别时，首先计算 Y 空间中样本所属各类别的中心。对于要判别的对象，计算其 Fisher 判别函数值，以及 Y 空间中与各类别中心的距离。然后利用距离判别法确定其所属类别。

四、贝叶斯判别法

贝叶斯判别法属于贝叶斯方法的范畴。贝叶斯方法是一种研究不确定性的推理方法，其中贝叶斯概率就是用来表示不确定性的一种主观概率。贝叶斯概率的估计取决于

先验知识的正确性和后验知识的丰富性，随着人们主观意识的改变而改变。

贝叶斯判别法的主要思路是：利用样本所属分类的先验概率通过贝叶斯法则求出样本所属分类的后验概率，并依据该后验概率分布做出统计推断，将样本归为后验概率最大的类别。具体如下：

首先，计算样本点 X 属于总体 $G_i(i=1,2\cdots,k)$ 的概率，记为 $p(G_i|X)$。

然后，根据 k 个概率值的大小，将样本点 X 归为概率最大的类别。

下面将介绍贝叶斯判别法的计算过程，重点讨论后验概率 $p(G_i|X)$ 如何计算。

第一步，计算先验概率。**先验概率**是指随机抽取一个样本属于总体 $G_i(i=1,2,\cdots,k)$ 的概率，记为 $p(G_i)$，可视为先验知识。设 k 个总体的先验概率分别为 $q_1,q_2,\cdots,q_k$。先验概率可以通过样本直接获得。

第二步，计算样本似然。**样本似然**是指在总体 $G_i(i=1,2,\cdots,k)$ 中抽到样本 X 的概率或概率密度，记为 $p(X|G_i)$。

第三步，计算样本属于总体 $G_i(i=1,2,\cdots,k)$ 的概率 $p(G_i|X)$。

根据贝叶斯准则，用判别函数的信息调整先验概率，有

$$p(G_i \mid X) = \frac{q_i p(X \mid G_i)}{\sum_{j=1}^{k} q_j p(X \mid G_j)}, \quad i = 1,2,\cdots,k \tag{10-9}$$

样本 X 属于 $p(G_i|X)$ 最大的类别。

本章小结

本章主要介绍了聚类分析的基本思想与概念，并应用 SPSS 软件结合例题介绍了两种聚类分析方法的具体操作步骤，给出了实例分析，最后还介绍了判别分析的概念以及三种不同的判别分析的方法。

1. 聚类分析是一种根据研究对象特征对研究问题进行分类的多元分析方法，是依据样本间相似性的度量标准将数据集自动分成几个群组的一种方法。

2. 聚类分析有 Q 型聚类和 R 型聚类两大类方法，具体分析方法有以下五种：系统聚类法、快速聚类法、两步聚类法、有序样本聚类法（最优分割法）和模糊聚类法。

3. 数值变量相似性的测度分为点与点之间的距离测度和类与类之间的距离测度。前者包括绝对值距离、欧氏距离、平方欧氏距离、切比雪夫距离和明可夫斯基效力距离；后者的测度方法包括最短距离法、最长距离法、未加权（加权）的类间平均法、未加权（加权）的类间重心法和离差平方和法。

4. 非数值变量相似性的测度通常采用关联测度来完成，一般基于列联表计算。

5. 判别分析是一种典型的多元统计分析方法，主要用于根据一定的准则来确定未分类对象所属类别。

6. 判别分析根据判别准则，可分为距离判别法、Fisher 判别法和贝叶斯判别法。

本章习题

1. 简述聚类分析的基本思想。
2. 如何测量点与点之间的距离？
3. 类与类之间的测量方法有哪些？各方法有何优缺点？
4. 如何测量非数值变量之间的相似性？
5. 简述系统聚类法的步骤。
6. 根据《中国统计年鉴2012》的相关数据，从各地区城镇居民家庭平均每人全年消费性支出的角度对以下26个省（区、市）进行分类：北京、天津、河北、山西、辽宁、吉林、黑龙江、上海、江苏、浙江、安徽、福建、江西、山东、河南、湖北、湖南、广东、广西、海南、重庆、四川、陕西、甘肃、宁夏、新疆。

参考指标：食品、衣着、居住、家庭设备用品及服务、医疗保健、交通和通信、教育文化娱乐服务用品、其他商品和服务。

7. 下表给出了21个国家的森林、草原等生态数据，试根据下表数据对这些国家进行聚类分析，讨论这些国家的林木资源情况。

国别	森林面积（万公顷）	森林覆盖率（%）	林木蓄积量（亿立方米）	草原面积（万公顷）
中国	11 978	12.5	93.5	31 908
美国	28 446	30.4	202	23 754
日本	2 501	67.2	24.8	58
德国	1 028	28.4	14	599
英国	210	8.6	1.5	1 147
法国	1 458	26.7	16	1 288
意大利	635	21.1	3.6	514
加拿大	32 613	32.7	192.8	2 385
澳大利亚	10 700	13.9	10.5	45 190
苏联	92 000	41.1	841.5	37 370
捷克	458	35.8	8.9	168
波兰	868	27.8	11.4	405
匈牙利	161	17.4	2.5	129
南斯拉夫	929	36.3	11.4	640
罗马尼亚	634	26.7	11.3	447
保加利亚	385	34.7	2.5	200
印度	6 748	20.5	29	1 200
印度尼西亚	2 180	84	33.7	1 200
尼日利亚	1 490	16.1	0.8	2 090
墨西哥	4 850	24.6	32.6	7 450
巴西	57 500	67.6	238	15 900

8. 下表统计了28个省(区、市)的每月消费数据,试按下表数据进行聚类分析,掌握地区消费情况。

单位:元

地区	食品	衣着	燃料	住房	生活用品	文化生活
天津	135.20	36.40	10.47	44.16	36.40	3.94
辽宁	145.68	32.83	17.79	27.29	39.09	3.47
吉林	159.37	33.38	18.37	11.81	25.29	5.22
江苏	144.98	29.12	11.67	42.60	27.30	5.74
浙江	169.92	32.75	12.72	47.12	34.35	5.00
山东	115.84	30.76	12.20	33.61	33.77	3.85
黑龙江	116.22	29.57	13.24	13.76	21.75	6.04
安徽	153.11	23.09	15.62	23.54	18.18	6.39
福建	144.92	21.06	16.96	19.52	21.75	6.73
江西	140.54	21.59	17.64	19.19	15.97	4.94
湖北	140.64	28.26	12.35	18.53	21.95	6.23
湖南	164.02	24.74	13.63	22.20	18.06	6.04
广西	139.08	18.47	14.68	13.41	20.66	3.85
四川	137.80	20.74	11.07	17.74	16.49	4.39
贵州	121.67	21.53	12.58	14.49	12.18	4.57
新疆	123.24	38.00	13.72	4.64	17.77	5.75
河北	95.21	22.83	9.30	22.44	22.81	2.80
山西	104.78	25.11	6.46	9.89	18.17	3.25
内蒙古	128.41	27.63	8.94	12.58	23.99	3.27
河南	101.18	23.26	8.46	20.20	20.50	4.30
云南	124.27	19.81	8.89	14.22	15.53	3.03
陕西	106.02	20.56	10.94	10.11	18.00	3.29
甘肃	95.65	16.82	5.70	6.03	12.36	4.49
青海	107.12	16.45	8.98	5.40	8.78	5.93
宁夏	113.74	24.11	6.46	9.61	22.92	2.53

9. 以下数据是关于我国10个省(区、市)发展报告的部分数据,数据观测了出生时预期寿命、义务教育普及率和人均GDP等指标,根据上述指标将10个省(区、市)分为高发展水平和中等发展水平两类,分别用“1”和“2”表示。数据如下。

地区	出生预期寿命(年)	成人识字率(%)	人均GDP(元)	分组
北京	76.0	99.0	5 374	1
上海	79.5	99.0	5 359	1
浙江	78.0	99.0	5 372	1
河南	72.1	95.9	5 242	1
河北	73.8	77.7	5 370	1

（续表）

地区	出生预期寿命(年)	成人识字率(%)	人均 GDP(元)	分组
辽宁	71.2	93.0	4 250	2
吉林	75.3	94.9	3 412	2
江苏	70.0	91.2	3 990	2
安徽	72.8	99.0	2 300	2
福建	62.8	80.6	3 799	2

现在又增加了青海、湖北、山东、陕西的数据，但未对它们分组，我们希望将这几个地区归入上述两类。请建立标准判别函数对这四个地区进行分类。

地区	出生预期寿命(年)	成人识字率(%)	人均 GDP(元)
青海	68.5	79.3	1 950
湖北	69.9	96.9	2 840
山东	77.6	93.8	5 233
陕西	69.3	90.3	5 158