高层结构方案设计的 K-Means 聚类分析法

张世海 张世忠 段慧杰

(南阳理工学院, 南阳 473004)

【摘 要】首先,在分析了聚类分析及特征的基础上,给出了k-均值聚类算法的基本思想、算法流程、准则函数及算法步骤等;其次,将聚类分析理论和方法引入高层结构智能方案设计,建立了基于K-Means聚类分析方法的高层结构智能方案设计方法,并给出了工程应用实例,及该实例的聚类结果及聚类过程的空间分布图、评价函数与迭代次数及聚类数间的关系曲线。实践表明:k-means聚类分析方法能有效地用于高层结构智能方案设计,为高层建筑结构智能方案设计开拓了崭新的途径和方法。

【关键词】高层建筑; 结构方案设计; 聚类分析; k-means 算法

【中图分类号】TU973 【文献标识码】A 【文章编号】1674-7461(2013)02-0041-05

1 引言

实际工程中大部分高层建筑的结构方案设计都是在已有相似工程实例结构方案基础上的整合和改进,若干相似实例的快速获取是高质与高效进行结构方案设计的基础和关键。聚类是一种按照对象间相似性进行无监督分类(或分簇)的过程[1],而非监督的聚类是根据实际数据的特征,按照以某种度量为标准的数据之间的相似性,把一组没有划分类的对象集划分成一系列有意义的不同的类,把特征属性相似的归为一类,不相似的作为另一类,使同一类之间相似性最小化,不同类之间相似性最大化,即聚类具有分组数未知、没有关于聚类的任何先验性知识、不需要用训练样本进行学习和训练、聚类结果动态、不同相似性度量和不同的目的要求将产生不同的聚类结果等特征。而工程实例的结构方案千变万化,很难对其结构方案进行确切的分类,显然,利用聚类分析的方法可以帮助设计者从大量没有结构方案分类的工程实例库中快速获取若干相似实例,据此即可进行当前结构的方案设计。聚类分析的算法较多,而k-means算法是一种应用最广泛的方法[2-4],为此,本文将探索利用基于k-means的聚类方法,来进行高层建筑结构智能方案设计。

2 k-means 算法

2.1 k-均值算法的基本思想

k-均值算法以最终分类个数k为参数,把 n 个数据对象 $\{x_j\}_n$ 分为 k 个聚类 $\{c_i\}_k$,以使聚类内有较高的相似度,相似度根据一个聚类中数据对象的平均值(被看做聚类的重心)来进行计算。

2.2 k-均值算法的流程

首先从 n 个数据对象中随机地选择 k 个对象,作为初始的聚类中心,对剩余的每个对象,根据其与各个聚类中心的距离或相似度,分别将它们赋予与它们最近或最相似的聚类;然后,重新计算每个聚类的平均值作为新的类心并调整各样本的类别;不断重复上述过程,直到各样本到其判属类心的距离平方之和最小或评价函数(或目标函数、准则函数、标准测度函数)收敛为止。

2.3 k-均值算法的准则函数 J_W

准则函数 J_W 定义为各聚类内所有对象的平均误差之和,即计算类内的每个点到它所属类中心的

【基金项目】 国家自然科学基金(61040031);河南省重点科技攻关(092102310157)与"河南省工程结构性能监测与控制创新型科技团队"

【作者简介】 张世海(1966-),男,博士,教授。主要从事结构智能设计与结构性能监测与控制等方面研究。

距离平方和。设有待分类样本集 $x = \{x_1, x_2, \cdots, x_n\}$，在某种相似性测度基础上被分划为 c 类 $\{x_i^{(j)}; j = 1,2,\cdots,c; i = 1,2,\cdots,n_j\}$，其中上角标 j 表示类别，下角标 i 表示类内模式的序号，$\Sigma n_j = n$，类内距离准则函数 J_w 定义为：

$$J_w = \sum_{j=1}^{c} \sum_{j=1}^{n_j} \| x_i^{(j)} - m_j \|^2 \Rightarrow \min \quad (1)$$

式中，m_j 表示 ω_j 类的中心或模式均值向量，按下式确定。

$$m_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i^{(j)} \qquad j = 1,2,\cdots,c \quad (2)$$

公式(1)表征了各样本到其所属类中心距离的平方和。聚类的目标是使 J_w 取最小，即 $J_w \to \min$，因 J_w 值越大，说明某些样本没有就近分类，在此意义上聚类效果不好，应重新调整分划。这种准则也称为误差平方和准则。

显然，J_w 是各样本 $x_i(i \in [1,n])$ 和类心 $m_j(j \in [1,c])$ 的函数，在样本集 $\{x_i\}n$ 给定条件下，J_w 的值取决于类心集 $\{m_j\}c$ 的选取，类心集的确定相应于样本类别的分划。该准则适用于同类样本比较密集，且各类别样本分布区域体积差别不大的情况，否则采用上述准则可能是不适宜的。例如，当某一类样本数目较多而另一类样本较少，两类样本所占空间大小明显不同，两类间的距离又不足够大时，样本较多的那一类中一些边缘处的样本可能距离另一类的类心更近一些。

2.4 k-均值算法步骤描述

输入：包含 n 个对象的数据库 $D = X = \{x_j\}_n$ 及期望聚类的簇数目 k。

输出：k 个簇，使平方误差准则最小。

k-均值算法：

(1) assign initial value for means $m_{1s}, m_{2s}, \cdots, m_{ks}$;//随机选择 k 个对象作为初始的聚类中心：$m_{1s}, m_{2s}, \cdots, m_{ks}$，置迭代步数 $s = 0$

(2) repeat

(3) For $j = 1$ to n Do assign each x_j to the cluster which has the closest center(mean);//将待分类的每个对象 $x_j \in \{x_j\}_n$ 按最小距离原则赋给 k 个初始的聚类中心中的某一类，或根据聚类中数值对象的平均值，将每个数据对象重新赋给最相似的簇。即如果 $d_{jl}(s) = \min[d_{ji}(s)], j = 1,2,\cdots,n$，则判 $x_j \in c_l(s+1)$。其中，$d_{ji}(s)$ 表示 x_j 和类 $c_i(s)$ 的中心 $m_i(s)$ 间的距离。于是产生了新的聚类 $c_i(s+1)(i = 1,2,\cdots,k)$。

(4) For $i = 1$ to k Do calculate new center for each cluster;//按公式 3 计算重新分类后每个聚类中数据对象的平均值或类中心，更新聚类平均值。其中，$n_i(s+1)$ 为 $c_i(s+1)$ 类中所含样本数。

$$m_i(s+1) = \frac{1}{n_i(s+1)} \sum_{j=1}^{n_i(s+1)} x_j \quad i = 1,2,\cdots,c \quad (3)$$

因该步采用了平均的方法计算调整后 k 个聚类的中心，故称该方法为 k-均值法。

(5) Compute J_W;// 按公式 4 计算评价函数 J_W。

$$J_W(s+1) = \sum_{i=1}^{k} \sum_{j=1}^{n_j(s+1)} \| x_j - m_i(s+1) \|^2 \quad (4)$$

(6) UNTIL convergence criteria is met//平均误差 $J_W \leqslant \varepsilon$ 或者 J_W 不在明显地变化或者 $m_i(s+1) = m_i(s)(i = 1,2,\cdots,c)$ 则结束，否则，$s = s+1$，转3)。

3 基于 k-means 聚类分析的高层结构智能方案设计

在高层建筑结构方案设计的聚类分析过程中，存在多种类型的数据，而 k-means 算法能有效地对数值属性进行聚类分析，因此，可利用 k-means 算法的这一特征，通过对工程实例的结构高度、长宽比、高宽比、场地类别、设防烈度等数值型属性信息的聚类分析，来进行高层建筑结构方案设计。以下给出基于 k-means 算法聚类的高层结构方案设计实例检索方法和工程实例。

3.1 输入样本与聚类数确定

采用表 1 中给出的 26 个工程实例数据中的高度、高宽比、长宽比 3 个数值型属性为聚类和实例检索依据，其中，利用前 20 个数据进行聚类，利用后 6 个数据进行实例检索，确定的聚类数目 $k = 4$。为解决属性间的不可公度性，需对各属性进行标准化或归一化处理，通过标准化处理后将各个属性值转化为[0,1]区间上的数值[5]。标准化处理后的样本输入矩阵为 X，聚类后的待检索输入样本矩阵为 Y。

$X = \{x_{ij}\}_{20\times3} = \{0.377\ 3 \quad 0.658\ 3 \quad 0.750\ 9$
$1.000\ 0 \quad 0.130\ 6 \quad 0.674\ 4 \quad 0.921\ 0 \quad 0.117\ 9$
$0.421\ 3 \quad 0.416\ 8 \quad 0.848\ 1 \quad 0.178\ 3 \quad 0.558\ 3$
$0.638\ 5 \quad 0.502\ 4 \quad 0.497\ 6 \quad 0.467\ 2 \quad 0.353\ 0$
$0.350\ 6 \quad 0.350\ 6;$ $0.573\ 8 \quad 0.694\ 9 \quad 0.572\ 1$

0.865 7 0.235 5 0.598 7 1.000 0 0.421 2
0.243 8 0.419 6 0.205 6 0.000 0 0.630 2
0.867 3 0.560 5 0.150 9 0.265 3 0.595 4
0.515 8 0.240 5;0.005 1 0.312 0 0.000 0
0.005 1 0.168 8 0.110 0 0.235 3 1.000 0
0.250 6 0.102 3 0.038 4 0.035 8 0.110 0
0.245 5 0.000 0 0.005 1 0.040 9 0.409 2
0.099 7 0.399 0}

$$Y = \{y_{ij}\}_{6\times 3} = \{0.331\ 7 \quad 0.137\ 3 \quad 0.188\ 9$$
0.445 9 0.026 7 0.000 0;0.731 3 0.262 0
0.754 6 0.255 4 0.784 4 0.7844 0.081 8
0.196 9 0.470 6 0.150 9 0.767 3 0.404 1}

3.2 样本分类原则与评价函数确定

按最小距离原则将每个数据样本赋给最相似的簇,按公式(4)给出的平均误差公式计算评价函数 J_W 值。

3.3 聚类过程与结果

按前述 k-均值算法步骤进行聚类分析,聚类结果见表2。图2给出了第1-6步聚类结果,图3给出了评价函数 J_W 随迭代次数增加的变化曲线,图4给出了聚类数 k 由2变化到10时 J_W 随 k 单调减小变化曲线,显然,当 $k = 4$ 时 J_W 的曲率变化最大,此时的分类数是比较接近从样本几何分布上看最优的类数。

表1 高层建筑工程实例属性信息(部分)

编号	工程名称	建筑高度	长度	宽度	高宽比	长宽比	结构型式
1	南京电信鼓楼多媒体综合楼	140.5	26.9	28.9	4.73	1.02	框架+筒体
2	深圳特区报业大厦	186.75	68.2	30.66	5.46	2.22	框架+筒体
3	上海森茂国际大厦	202	42.8	42.8	4.72	1.00	框架+筒体
4	大连世界贸易大厦	243	38.3	37.4	6.49	1.02	钢框架-筒
5	光大银行长春分行大厦	99.9	61.8	37.2	2.69	1.66	钢框架
6	陕西信息大厦	189.4	55.4	38.8	4.88	1.43	钢框架-筒
7	上海浦东国际金融大厦	230	60.67	31.5	7.30	1.92	钢-混凝土
8	深圳天安数码时代大厦	97.8	24	64	3.81	4.91	框剪
9	上实南洋广场	147.75	107	54	2.74	1.98	框架+筒体
10	南京体仁大厦	147	54.2	38.7	3.80	1.40	框架+筒体
11	深圳鸿昌广场	218	100	87	2.51	1.15	钢筋砼框筒
12	北京数码大厦	107.75	97	85	1.27	1.14	钢筋砼框筒
13	深圳罗湖商务大厦	170.3	43.5	30.4	5.07	1.43	钢筋砼框筒
14	京城大厦	183.5	113	57.6	6.50	1.96	框剪
15	深圳国际贸易中心大厦	161.1	34.6	34.6	4.65	1.00	筒体
16	深圳发展中心大厦	160.3	49.5	48.6	2.18	1.02	框剪
17	中国国际贸易中心办公楼	155.3	63	54	2.87	1.16	筒体
18	珠江帆影4号主楼	136.5	73.2	28.1	4.86	2.60	框剪筒
19	深圳外贸中心大厦	136.0	43.05	31.04	4.38	1.39	筒体
20	中国银行大厦	136.1	128	50	2.72	2.56	筒体
21	深圳航空大厦	133	30.9	23.4	5.68	1.32	框剪
22	国际大厦	101	62.7	35.4	2.85	1.77	框剪
23	广东省人民银行	109.5	53.4	18.8	5.82	2.84	框剪
24	上海电信大楼	151.8	54	34	2.81	1.59	框架+筒体
25	香格里拉饭店	82.8	54.4	13.6	6.00	4.00	框剪
26	广州远洋大厦	78.4	33.6	13	6.00	2.58	板柱

表 2　k-均值算法聚类结果

迭代次数	K 个类中心			产生的新类	K 个新类中心		
1	0.377 3	0.573 8	0.005 1	{1,5,8,9,10,12,13,15 –20}	0.294 1	0.573 8	0.005 1
	0.658 3	0.694 9	0.312 0	{2,14}	0.612 6	0.694 9	0.312 0
	0.750 9	0.572 1	0.000 0	{3,6,11}	0.717 6	0.572 1	0.000 0
	1.000 0	0.865 7	0.005 1	{4,7}	1.000 0	0.865 7	0.005 1
2	0.294 1	0.573 8	0.005 1	{1,5,8,9,10,12,13,15 –20}	0.266 5	0.265 3	0.102 3
	0.612 6	0.694 9	0.312 0	{2,14}	0.601 4	0.781 1	0.278 8
	0.717 6	0.572 1	0.000 0	{3,6,11}	0.630 9	0.572 1	0.038 4
	1.000 0	0.865 7	0.005 1	{4,7}	0.955 2	0.932 8	0.120 2
3	0.266 5	0.265 3	0.102 3	{5,8,9,10,12,16 –20}	0.265 2	0.254 6	0.135 6
	0.601 4	0.781 1	0.278 8	{2,14}	0.601 4	0.781 1	0.278 8
	0.630 9	0.572 1	0.038 4	{1,3,6,11,13,15}	0.565 1	0.573 0	0.021 7
	0.955 2	0.932 8	0.120 2	{4,7}	0.955 2	0.932 8	0.120 2
4	0.265 2	0.254 6	0.135 6	{5,8 –12,16,17,20}	0.338 8	0.240 5	0.102 3
	0.601 4	0.781 1	0.278 8	{2,6,14}	0.612 6	0.694 9	0.245 5
	0.565 1	0.573 0	0.021 7	{1,3,13,15,18,19}	0.365 0	0.573 0	0.052 4
	0.955 2	0.932 8	0.120 2	{4,7}	0.955 2	0.932 8	0.120 2
5	0.338 8	0.240 5	0.102 3	{5,9,11,12,16,17,20}	0.344 0	0.235 5	0.040 9
	0.612 6	0.694 9	0.245 5	{2,3,6,13}	0.621 7	0.614 4	0.110 0
	0.365 0	0.573 0	0.052 4	{1,8,10,15,18,19}	0.280 3	0.538 1	0.101 0
	0.955 2	0.932 8	0.120 2	{4,7,14}	0.910 5	0.867 3	0.235 3
6	0.344 0	0.235 5	0.040 9	{5,9,11,12,16,17,20}	0.344 0	0.235 5	0.040 9
	0.621 7	0.614 4	0.110 0	{2,3,6,13}	0.621 7	0.614 4	0.110 0
	0.280 3	0.538 1	0.101 0	{1,8,10,15,18,19}	0.280 3	0.538 1	0.101 0
	0.910 5	0.867 3	0.235 3	{4,7,14}	0.910 5	0.867 3	0.235 3

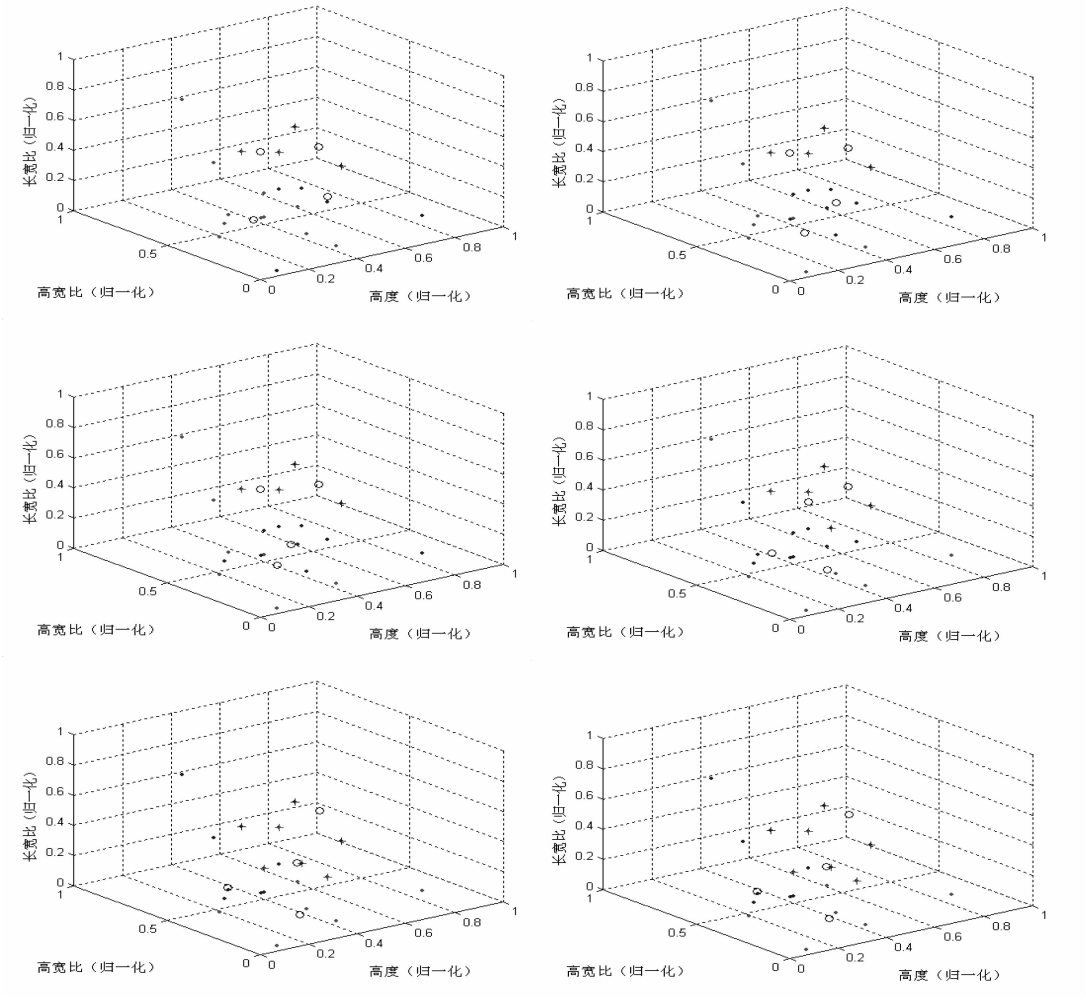

图 1　k-means 聚类结果(4 类,第 1 –6 步)

3.4 聚类结果检验

根据 4 个中心及其相应的聚类结果,即可利用待输入样本矩阵 Y 进行其相似实例聚类,以确定与当前方案相似的工程实例,据此就能确定结构型式及其结构方案。首先,可确定样本矩阵中每个待输入样本与各个聚类中心的距离;然后,根据最小距离原则确定其所属的类别及其相似的工程实例;最后,再根据相似工程实例方案的类别或相似实例中出现频次最高的结构方案类别作为当前的结构方案设计依据[6]。下式给出了 6 个待输入样本与 4 个聚类中心间的距离矩阵 D, 其中,d_{ij} 为样本 y_i 与聚类中心 c_j 之间的距离。

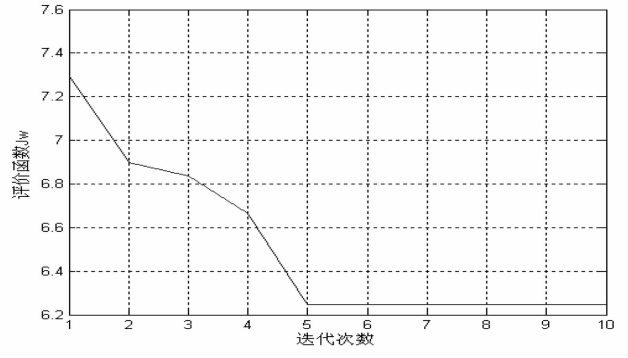

图 2 评价函数 J_W 与迭代次数关系曲线

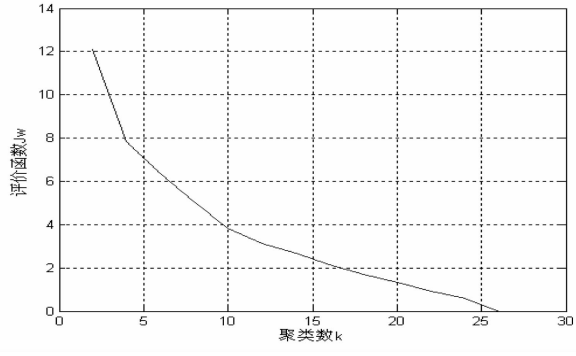

图 3 聚类数 k 与 J_W 的关系曲线

$$D = \{d_{ij}\}_{6\times4} = [0.275\ 1 \quad 0.040\ 5 \quad 0.050\ 8$$
$$0.753\ 8 \quad 0.000\ 6 \quad 0.562\ 3 \quad 0.104\ 5 \quad 2.007\ 4$$
$$0.629\ 9 \quad 0.004\ 6 \quad 0.244\ 7 \quad 0.358\ 8 \quad 0.053\ 7$$
$$0.244\ 0 \quad 0.004\ 5 \quad 1.347\ 6 \quad 0.917\ 7 \quad 0.054\ 0$$
$$0.434\ 2 \quad 0.189\ 0 \quad 0.322\ 7 \quad 0.024\ 8 \quad 0.072\ 4$$
$$0.679\ 9]$$

由上述距离矩阵,根据最小距离原则可确定 6 个待输入实例所属的类别分别为:2、1、2、3、2、2,各类的相似实例见表 2,由此即可根据所属类中的相似实例的结构方案进行当前结构方案的设计与创新。

4 结论

在给出了 k-均值算法的基本思想、准则函数、步骤流程等基础上,将具有无导师学习特征的聚类分析理论和方法引入高层结构智能方案设计,建立了基于 K-Means 聚类分析方法的高层结构智能方案设计实例获取方法,给出了工程应用实例:以表 1 中的 26 个工程实例数据为依据,对前 20 个工程实例数据进行了聚类分析,并给出了聚类结果及聚类过程的空间分布图、评价函数 J_W 随迭代次数增加的变化曲线、聚类数 k 由 2 变化到 10 时 J_W 随 k 单调减小变化曲线,并对后 6 个实例数据进行了实例聚类,给出了相似实例,为高层建筑结构方案智能设计开拓了崭新的途径和方法。

参考文献

[1] Jain A, Murty M, Flynn P. Data clustering: A review. ACM Computing Surveys (CSUR), 1999, 31(3): 264-323.

[2] Macqueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkely Symposium on Mathematical Statistics and Probability, Berkely, CA, 1967, vol. 1, 281-297.

[3] Huang J Z, Ng M K, Rong H-Q, Li Z-C. Automated variable Weighting in k-Means Type Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.

[4] Wagstaff K, Cardie C, Rogers S. Constrained k-means clusterring with background knowledge. In: Proceedings of the 8th International Conference on Machine Learning, Morgan, Kaufmann, 2001:577-584.

[5] 张世海. 高层建筑结构智能方案设计方法研究, 哈尔滨工业大学博士后研究工作报告, 2009.

[6] Shihai Zhang, Changyong Wang Shujun Liu. Intelligent scheme design of high-rise structure for K-means-based case retrieval. Proceedings of the 2010 Second WRI Global Congress on Intelligent Systems (GCIS'2010). Sponsored by Wuhan University of Technology and World Research Institutes. Los Almitos, California Washington · Tokyo GCIS'2010(vol. 3): 241-244.

(下转第 65 页)