

文章编号 :0253-9721(2007)02-0100-04

## 二维非接触人体测量中体型的模糊聚类分析

王玉秀,李晓久,刘皓

(天津工业大学 艺术与服装学院,天津 300160)

**摘 要** 二维非接触人体测量系统中,围度尺寸需要通过二维至三维的数据转换间接获得。为增强围度预测的精度,应按不同体型类别分别建立预测模型。为避免指标间的信息重复,进行主成分分析,将原有的 24 项指标归纳为 4 项综合指标,对人体体型进行客观的综合评价。应用模糊 C 均值聚类分析进行人体体型分类,分别建立预测模型,以最大隶属度原则对样本进行分类。应用 F 统计量和聚类有效性函数检验聚类的有效性,并结合实际应用结果确定最佳分类数。

**关键词** 主成分分析; 体型分类; 聚类分析

中图分类号: TS941.17 文献标识码: A

### Fuzzy clustering analysis of body type in 2D non contacted body measurement system

WANG Yuxiu, LI Xiaojiu, LIU Hao

(School of Art and Clothing, Tianjin Polytechnic University, Tianjin 300160, China)

**Abstract** In the 2D non-contacted body measurement, the size of body girths can only be acquired indirectly by data transform from 2D to 3D. In order to increase the precision of the prediction of girth size, it is necessary to develop different prediction models in accordance with body types. To avoid the information overlapping, the principal component analysis is undertaken to sum up the initial 24 independent variables and reduce them to 4 independent variables for objective and comprehensive evaluation of body types. The function of fuzzy clustering analysis is used to make classification of body types whose prediction models are established respectively. The samples are classified on the principle of the maximum membership and the classification cluster validity are verified by F-statistic and the value of clustering validity function. And the optimal number of classification is determined through practical application.

**Key words** principal component analysis; classification of body types; clustering analysis

在二维非接触人体测量系统的开发过程中,对人体尺寸数据的客观分类具有重要的意义。首先,人体测量系统通过对数字图像的处理直接获取人体高度、宽度、厚度的尺寸;但围度方面的尺寸,如胸围、腰围、臀围等则需要通过二维至三维的数据转换间接获得,因此,如何通过人体的宽度、厚度和高度等已知信息,准确地推算出人体的三围尺寸是二维非接触人体测量系统开发的关键<sup>[1]</sup>。为增强围度尺寸预测的精确度,减少模型内部不确定因素,可将人

体体型分类,分别建立预测模型。其次,系统功能的进一步延伸是建立个性化数字人台,为实现网络化服装量体定制的实用性功能创造条件。研究人体体型的客观生长规律,进行体型的无监督分类是建立个性化数字人台的基础。本文以在校女大学生为测量对象,获取并选定 24 项原始数据,通过主成分分析对这些原始数据进行综合整理,使用模糊划分方法对样本进行无监督的聚类分析。实验验证,该体型聚类结果可有效提高围度的测量精度。

收稿日期:2006-06-20 修回日期:2006-09-22

作者简介:王玉秀(1976—),女,讲师,博士生。研究方向为数字化服装技术。李晓久,通讯作者,E-mail:lixiaojie@tjpu.edu.cn。

## 1 聚类综合指标的选取

测量对象选择 101 名年龄在 18 ~ 25 岁的在校女大学生。实验选定与人体体型相关的 24 项数据指标:体重、身高、颈椎点高、肩宽、乳点高、胸宽、背宽、肩高、背长、胸厚、腰厚、臀厚、颈厚、胸围、腰围、臀围、颈围、臀高、臂长、下裆高、腰高、肩胸距、胸腰距、腰臀距。由于指标较多,而人体不同部位尺寸之间存在着内在的联系,为避免以上指标间的信息重复,尽量减小由于多重共线性引起的误差,首先对变量进行主成分分析,使变量简化降维,使问题更简单直观。对 24 个指标数据进行主成分分析,分析结果见表 1。表 2 为 4 个主成分的系数矩阵。

表 1 4 个主成分的特征根、贡献率与累计贡献率

Tab.1 Characteristic root, contribution rate and cumulative contribution rate of 4 principal components

项目	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分
特征根	5.542	2.827	1.200	1.002
贡献率	46.180	23.558	9.996	8.352
累计贡献率	46.180	69.739	79.735	88.087

表 2 4 个主成分的系数矩阵

Tab.2 Coefficient matrix of 4 principal components

部位	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分
体重	0.838	0.211	0.105	-0.042
身高	0.790	-0.541	0.044	0.080
颈椎点高	0.827	-0.522	0.025	0.094
肩宽	0.363	0.213	-0.010	0.501
乳点高	0.628	-0.630	0.265	-0.164
胸宽	0.582	0.044	0.114	0.083
背宽	0.219	0.570	-0.207	0.208
肩高	0.820	-0.505	-0.002	0.118
背长	0.368	-0.305	0.203	0.595
胸厚	0.270	0.575	-0.081	-0.229
腰厚	0.581	0.600	0.192	-0.294
臀厚	0.525	0.651	0.166	-0.192
颈厚	0.552	0.360	0.102	0.134
胸围	0.639	0.594	-0.148	-0.068
腰围	0.666	0.582	0.000	-0.131
臀围	0.698	0.553	0.039	0.057
颈围	0.580	0.395	-0.025	0.010
臀高	0.705	-0.276	-0.364	-0.270
臂长	0.711	-0.318	-0.262	-0.034
下裆高	0.608	-0.594	-0.371	-0.220
腰高	0.794	-0.296	-0.082	-0.215
肩胸距	0.375	0.149	0.563	0.106
胸腰距	-0.026	-0.352	0.581	0.036
腰臀距	0.278	-0.108	0.509	0.070

由表 2 可知,第 1 主成分均匀地综合了除肩宽、

背宽、背长、胸厚、肩胸距、胸腰距、腰臀距以外的 17 个指标的信息,这 17 个指标变量在第 1 主成分上的载荷都超过了 50%,反映了人体高度、围度与体重的综合信息;第 2 主成分中反映人体的围度、厚度方面的指标以较高的正值出现,而身高、颈椎点高、乳点高、下裆高这 4 个反映人体高度信息的指标则以较凸出的负值出现,由此可见体型的变化呈现出 2 种不同的倾向,一方面表现为细长型,另一方面表现为矮胖型;第 3 主成分中肩胸距、胸腰距、腰臀距 3 个指标值凸出,集中体现人体躯干部位的各段长度;第 4 主成分中只有肩宽、背长 2 个指标值凸出,集中体现人体腰节以上躯干部位的宽度与长度信息。由于前 4 个主成分的累计贡献率为 88.087%,而具体的含义较为清晰,因此可以将以上 4 个综合指标作为体型分类的变量做进一步的聚类分析。

## 2 改进的模糊 C 均值体型聚类分析

### 2.1 模糊 C 均值聚类分析原理

在模糊 C 均值聚类方法中,每一个数据点按照一定的模糊隶属度隶属于某一聚类中心。该方法首先随机选取若干聚类中心作为初始聚类中心,所有数据点都被赋予对聚类中心一定的模糊隶属度,然后通过迭代方法不断修正聚类中心。迭代过程中以极小化所有数据点到各个聚类中心距离与隶属度值的加权和为优化目标。

对于经过标准化处理后形成的原始数据,设  $X \in R^p$ ;  $X = \{x_1, x_2, \dots, x_n\}$ ,其中  $n$  为数据项数;整数  $C$  为类别数,且  $2 \leq C \leq n$ ;隶属度矩阵  $U = [u_{ik}]_{C \times n}$ ,具有以下性质:

$$u_{ik} \in [0, 1]$$

$$\sum_{i=1}^C u_{ik} = 1, \quad \forall k \quad (1)$$

$$0 < \sum_{i=1}^C u_{ik} < n, \quad \forall i$$

$V = \{v_1, v_2, \dots, v_c\}$  为  $C$  个聚类中心集合,  $v_i \in R^p$ , 是第  $i$  类的聚类中心。定义目标函数为:

$$J(U, V) = \sum_{k=1}^n \sum_{i=1}^C (u_{ik})^m (d_{ik})^2 \quad (2)$$

$m \in [1, \infty]$  是加权指数。  $J(U, V)$  表示各类中样本到聚类中心的加权距离平方和,权重是样本  $x_k$  对第  $i$  类隶属度的  $m$  次方。设  $x_k$  和  $v_i$  都是  $p$  维向量,即  $\forall x_k, v_i \in R^p$ ,  $A$  为  $p \times p$  矩阵,则  $(d_{ik})^2$  的一般表达式为:

$$(d_{ik})^2 = \|x_k - v_i\|^2 = (x_k - v_i)^T A(x_k - v_i) \quad (3)$$

当取  $A = I$  时,为欧氏距离。聚类准则为求  $J(U, V)$  的极小值  $\min\{J(U, V)\}$ 。对于  $\forall k$ , 定义集合  $I_k$  和  $\bar{I}_k$  为:  $I_k = \{i | 1 \leq i \leq C, d_{ik} = 0\}$ ,  $\bar{I}_k = \{1, 2, \dots, C\} - I_k$ 。

使得  $J(U, V)$  为最小的  $u_{ik}$  值为:

$$\begin{cases} u_{ik} = 1 / \sum_{j=1}^C (d_{ik}/d_{jk})^{2(m-1)}, \text{当 } I_k = \phi \\ u_{ik} = 0, \forall i \in \bar{I}_k, \sum_{i \in I_k} u_{ik} = 1, \text{当 } I_k \neq \phi \end{cases} \quad (4)$$

由此可获得  $J(U, V)$  为最小时  $v_i$  的值

$$v_i = \frac{\sum_{k=1}^n (u_{ik})^m x_k}{\sum_{k=1}^n (u_{ik})^m}, \quad i = 1, 2, \dots, C \quad (5)$$

### 2.2 改进的模糊 C 均值方法

在模糊 C 均值算法中,由于引入了式(1)的归一化条件,在样本集不理想的情况下可能导致结果不好。比如,如果某个野值样本远离各类的聚类中心,本来它严格属于各类的隶属度都很小,但由于式(1)条件的要求,将会使它对各类都有较大的隶属度(比如 2 类情况下各类的隶属度都是 0.5),这种野值的存在将影响迭代的最终结果<sup>[21]</sup>。为了克服这种缺陷,放松归一化条件,使所有样本对各类的隶属度总和为  $n$ ,即

$$\sum_{i=1}^C \sum_{k=1}^n u_i(x_k) = n$$

因此,式(4)变换为

$$u_i(x_k) = \frac{n(1/\|x_k - v_i\|^2)^{1/(m-1)}}{\sum_{i=1}^C \sum_{k=1}^n (1/\|x_k - v_i\|^2)^{1/(m-1)}} \quad (6)$$

### 2.3 改进的模糊 C 均值聚类计算方法

1) 确定分类数  $C$ ,理论上  $C$  值的取值范围是

表 3 相对于不同分类数的 *Mixed-F* 统计量值和聚类有效值

Tab.3 *Mixed-F* statistic value and clustering validity value of different number of classification

	2	3	4	5	6	7	8
<i>Mixed-F</i>	10.587 5	35.205 6	28.791 9	26.766 7	24.555 6	25.326 1	23.550 1
$E(D^*; C)$	0.627 2	0.528 2	0.688 3	0.646 4	0.607 1	0.571 7	0.540 9

分析表 3 数据,根据 *Mixed-F* 的值可以认为 3 类为样本的最佳分类数,并且样本分为 3 类时的聚类有效值也相对较小,因此表 3 数据显示样本分为 3 类是最合理的。

在实际应用中,聚类结果的“好坏”往往是由结果的“有用性”来决定的<sup>[41]</sup>。由于此次实验的目的是

$2 \leq C \leq n$ ,  $n$  是数据个数,但此次聚类作为计算三维数据的预聚类,分类数过多会使计算复杂化,因此,取值  $2 \leq C \leq 8$ 。2) 求初始聚类中心。使用硬 C 均值(Hard C-means Cluster)方法确定聚类中心并作为初始聚类中心。3) 确定加权指数  $m$ 。 $m$  值的大小与最终分类的模糊程度有关,当  $m \rightarrow 1$  时,分类具有较小模糊性;当  $m \rightarrow \infty$  时,  $U$  中所有元素均接近为  $1/C$ ,对  $m$  的选取尚无理论依据可循,一般的经验值为  $1.1 \leq m \leq 5$ ,此次实验取经验值  $m = 2$ 。4) 任意制定模糊分类矩阵  $U^{(0)}$ 。5) 依次取  $b = 0, 1, 2, \dots$ ,由式(6)计算  $U^{(b+1)}$ ,由式(5)计算  $v_i^{(b)}$ ;以一个合适的矩阵范数来比较  $U^{(b)}$  和  $U^{(b+1)}$ ,若  $\|U^{(b)} - U^{(b+1)}\| < \epsilon$ ,停止,否则,重复步骤 5),  $\epsilon$  是一个预置的允许误差。

### 2.4 确定最佳分类数

聚类分析是一种无监督的分类方法,其分类过程是客观的,但是人为确定的分类数存在主观性,为了减少主观因素对客观评价的影响,应用混合  $F$  统计量来确定一个最优分类数,应用基于模糊熵的聚类有效性函数值来检验分类的有效性。单个指标的  $F$  统计量是平均组间离差平方和与平均组内离差平方和之比,因此  $F$  统计量的值越大表示类间离散程度越大,类内的离散度越小,分类效果越好。模糊熵表示各类中样本的不确定性,其值越小则表明分类的效果越好<sup>[31]</sup>。对于给定的聚类数  $C$  和最优模糊划分系数矩阵  $D$ ,根据式(4)、(5)能够得到式(7)所示的聚类有效性函数。

$$E(D^*; C) = \sum_{i=1}^C \left| 1 - \frac{1}{n} \sum_{j=1}^n |1 - 2d_{ij}^*| \right| \quad (7)$$

确定最佳分类值,其对应  $E(D^*; C)$  应是各类别中最小的。选取迭代精度为 0.01,计算 2~8 类 *Mixed-F* 统计量值和聚类有效值,其结果如表 3 所示。

增强围度计算的精确度,因此,应进一步计算围度以验证分类的有效性与可行性。

将样本分为 3 类,具体分类结果见表 4。通过对数据的进一步分析,发现第 1 类人群体型纤细偏瘦,第 2 类人群体型胖瘦比较均匀,基本属于正常体型,第 3 类人群体型粗壮偏胖。从图像上可以比较

明显地观察出 3 类人群的体型差异,特别是第 1 类与第 3 类区分更加明显。分类结果显示,这种无监督的聚类方法能够与客观规律相吻合。

表 4 各类别包含样本量

Tab.4 Sample size of different classification

类别	样本量	所占比例/ %
1	28	27.7
2	53	52.5
3	20	19.8
总计	101	100.0

采用灰色系统理论,建立臀围尺寸的计算模型。首先,使用所有样本建立统一的灰色模型进行计算,具体方法详见文献[5],再将分类后 3 类样本分别建立模型并计算臀围尺寸。从人体绘制臀围曲线,精确测量其曲线长度,作为臀围的基准数据,将以上 2 种计算结果与基准数据进行对比。结果显示,分类后误差平均值降低 0.22 cm,误差值分布趋势普遍向中心收敛,误差峰值明显降低。由于样本量大,随机抽取 21 个样本的分类前后误差量作折线图,如图 1 所示。由图可以看出这种对体型的分类可增强围度测量的精度与稳定性。

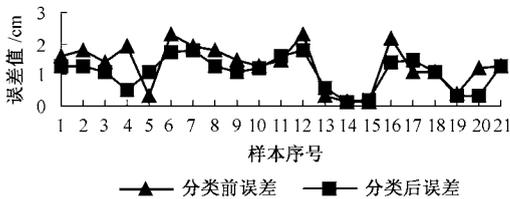


图 1 分类前后误差对比

Fig.1 Comparison between the error of classified and unclassified

### 3 结 论

1) 人体体型是一个复杂的多变量系统,各变量间彼此相关<sup>[6]</sup>。为避免以上指标间的信息重复,减小由于多重共线性引起的误差,使用主成分分析方法将原有 24 项原始数据指标综合为 4 项综合指标,使变量合理简化降维。

2) 使用 4 个综合指标作为体型分类变量,进行改进的模糊 C 均值聚类分析,并通过实际应用确定最佳分类数为 3。这样既提高了预测精度,又符合人体实际生长规律。

该分类方法可用于二维非接触人体测量系统进行人体体型的分类,亦可为进一步研究体型规律、建立体型自动判别系统提供依据。

FZXB

#### 参考文献:

- [1] 李晓久,王玉秀.二维非接触式人体测量系统中臀围的计算[J].纺织学报,2004,25(2):98-100.
- [2] 边肇祺,张学工.模式识别[M].北京:清华大学出版社,2000:281-282.
- [3] 孙才志,王敬东,潘俊.模糊聚类分析最佳聚类数的确定方法研究[J].模糊系统与数学,2001,15(1):89-92.
- [4] 张文彤.SPSS 统计分析高级教程[M].北京:高等教育出版社,2004:258-259.
- [5] 李晓久,朱广舟.人体胸围尺寸灰色模型建立[J].纺织学报,2005,26(5):77-79.
- [6] 陈文飞,潘箐.人体体型灰色模型建立[J].中国纺织大学学报,2000,26(4):14-16.