一种新的中心对称聚类算法*)

林嘉宜　许剑峰　彭　宏
（华南理工大学计算机科学与工程学院　广州510641）

A Novel Clustering Algorithm Based on Central Symmetry

LIN Jia-Yi　XU Jian-Feng　PENG Hong
(College of Computer Science and Engineering, South China University of Technology, Guangzhou, 510641)

Abstract　Data clustering is an important reserch field in data mining. The key of the clustering algorithm is the distance measure. In this paper, we put forward a new distance measure based on central symmetry. Then we apply it to data clustering. The experimental studies prove the feasibility of this algorithm and get a satisfied result in face detection.

Keywords　Clustering, Central symmetry, Face detection

1　引言

聚类分析是数据挖掘中的一个重要研究方向。聚类将数据对象分组成多个类，在同一个类中的对象之间的相似度较高，而不同类中的对象之间的相似度较低。目前，聚类分析技术已经广泛应用于许多领域，包括数据挖掘，统计学，生物学，机器学习等。

聚类算法对数据对象的相似性度量是通过对象的属性值来计算的，在对象的属性上定义不同的相似性度量将会得到不同的聚类结果。目前主要的聚类方法有划分方法，层次方法，基于密度的方法，基于网络的方法和基于模型的方法等。

在长期的研究中，人们已经提出了许多有效且具扩展性的聚类算法[1,2]，其中k-平均算法是最常用的聚类算法之一。本文的第二部分在介绍了传统的相似性度量算法后，提出了一种新的基于中心对称的相似性度量；在第三部分将k-平均算法和新的基于中心对称的度量结合起来，提出一种新的中心对称聚类算法，并应用到数据聚类中；最后，用实验验证了新算法的有效性，并且将其应用到于人脸检测。

2　基于中心对称的相似性度量

2.1　传统的相似性度量

不同的聚类算法，最重要的区别是相似性度量的不同。用得最多的相似性度量是基于距离的度量，其中包括欧几里得距离，定义如下：

$$d(i,j)=\left(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^2\right)^{\frac{1}{2}} \tag{1}$$

其中　$i=(x_{i1},x_{i2},\cdots x_{ip})$和$j=(x_{j1},x_{j2},\cdots x_{jp})$是两个p维的数据对象。

用欧几里得距离作为相似性量度，通常能找到超球面的聚类，而另一个著名的度量方法曼哈坦距离能找到超椭球聚类，其定义为：

$$d(i,j)=\sum_{k=1}^{p}|x_{ik}-x_{jk}| \tag{2}$$

以上两者的概化是明考斯基距离，定义为：

$$d(i,j)=\left(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^q\right)^{\frac{1}{q}} \tag{3}$$

其中q是一个正整数。当q=1时，它表示曼哈坦距离，当q=2时，它表示欧几里得距离。

传统的k-平均算法[3,4]一般采用欧几里得距离来找到超球面的聚类，但是要找到其他形状的聚类就要定义更加灵活的相似性度量，本文提出了一种基于中心对称的量度，用于检测具有中心对称性的数据模式。

2.2　新的基于中心对称的相似性度量

给定N个模式$\vec{x}_i$，$(i=1,2,\cdots,N)$，和一个中心向量$\vec{c}$，则任两个模式$\vec{x}_i$，$\vec{x}_j$之间的距离定义为：

$$a_{ij}=\|\vec{x}_i-\vec{x}_j\| \tag{3}$$

模式$\vec{x}_i$和中心向量$\vec{c}$的距离定义为：

$$a_{ic}=\|\vec{x}_i-\vec{c}\| \tag{4}$$

那么模式$\vec{x}_i$和中心向量$\vec{c}$的中心对称度量可以定义为：

$$d(\vec{x}_i,\vec{c})=(\cos\alpha+k)\cdot\frac{\max(a_{ic},a_{jc})}{a_{ic}+a_{jc}} \tag{5}$$

其中α是向量$\vec{a}_{ic}=\vec{x}_i-\vec{c}$和向量$\vec{a}_{jc}=\vec{x}_j-\vec{c}$之间的夹角。根据余弦定理，有

$$\cos\alpha=\frac{a_{ic}^2+a_{jc}^2-a_{ij}^2}{2a_{ic}\cdot a_{jc}} \tag{6}$$

(5)式中k是一个正整数，且$k\geqslant 2$。当$d(\vec{x}_i,\vec{c})$取得最小值，则称$\vec{x}_j$是$\vec{x}_i$关于$\vec{c}$的中心对称模式。

下面对公式(5)做简要分析。

判断两个数据模式$\vec{x}_i$，$\vec{x}_j$是否关于中心向量$\vec{c}$对称，首先考虑向量$\vec{a}_{ic}$和向量$\vec{a}_{jc}$之间的夹角α的大小，α的值域是$[0,\pi]$，显然，α越大，则$\vec{x}_i$，$\vec{x}_j$两模式关于中心向量$\vec{c}$的对称度越好。$\cos\alpha$的值域为$[-1,1]$，再加上一个正整数$k(k\geqslant 2)$，使其值域落在正整数范围$[k-1,k+1]$内，可以根据不同的应用设定不同的k值。显然，$\cos\alpha+k$的值越小，则对称度越好。

其次，还要考虑a_{ic}和a_{jc}的大小，两者大小越接近，则$\vec{x}_i$，$\vec{x}_j$两模式关于中心向量$\vec{c}$的对称度越好。(5)式引入了系数$\lambda=\frac{\max(a_{ic},a_{jc})}{a_{ic}+a_{jc}}$作为度量。$\lambda$的值域为$[\frac{1}{2},1]$。$\lambda$的值越小，则对称度越好。

*)基金项目：广东省自然科学基金(990582)和广州市科委基金(2000-J-006-01)。林嘉宜　博士研究生，主要研究方向：数据挖掘技术。许剑峰　博士研究生。彭　宏　教授，博导。

综上，$d(\vec{x}_i,\vec{c})$值域为$[\frac{1}{2}(k-1),k+1]$。不难证明，当 $\alpha=\pi$ 且 $a_{ic}=a_{jc}$时，$d(\vec{x}_i,\vec{c})$取得最小值$\frac{1}{2}(k-1)$。

下面对该中心对称度量举例说明：

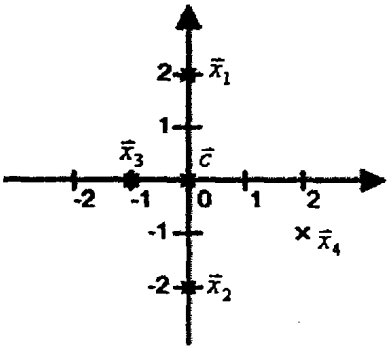

图1　中心对称量度的例子

如图1所示，有四个模式 $\vec{x}_1=(0,2)^T,\vec{x}_2=(0,-2)^T,\vec{x}_3=(-1,0)^T,\vec{x}_4=(2,-1)^T$ 和中心点 $\vec{c}=(0,0)^T$。把数据代入(5)，结合(6)得(k 取2)：

$$d(\vec{x}_1,\vec{c})=(\frac{a_{1c}^2+a_{2c}^2-a_{12}^2}{2a_{1c}\cdot a_{2c}}+2)\cdot\frac{\max(a_{1c},a_{2c})}{a_{1c}+a_{2c}}=0.5$$

$$d(\vec{x}_2,\vec{c})=(\frac{a_{2c}^2+a_{1c}^2-a_{21}^2}{2a_{2c}\cdot a_{1c}}+2)\cdot\frac{\max(a_{2c},a_{1c})}{a_{2c}+a_{1c}}=0.5$$

$$d(\vec{x}_3,\vec{c})=(\frac{a_{3c}^2+a_{4c}^2-a_{34}^2}{2a_{3c}\cdot a_{4c}}+2)\cdot\frac{\max(a_{3c},a_{4c})}{a_{3c}+a_{4c}}\approx0.764$$

$$d(\vec{x}_4,\vec{c})=(\frac{a_{4c}^2+a_{3c}^2-a_{43}^2}{2a_{4c}\cdot a_{3c}}+2)\cdot\frac{\max(a_{4c},a_{3c})}{a_{4c}+a_{3c}}\approx0.764$$

计算结果表明，$\vec{x}_1$和 $\vec{x}_2$是关于中心点 $\vec{c}$的一对对称点，两者的对称度量 $d(\vec{x}_1,\vec{c})$和 $d(\vec{x}_2,\vec{c})$都取得最小值；$\vec{x}_3$和 $\vec{x}_4$关于中心点 $\vec{c}$的对称度量约为0.764，并不是关于中心点 $\vec{c}$的严格对称。这与观测结果是一致的。

3　聚类算法

以(5)式作为相似性度量，本文提出了一种新的中心对称聚类算法。算法的主要思想是：先用 k 平均算法定出类的初始中心点，然后再根据(5)式对数据点以及中心点不断调整，最终得到 k 个聚类。

聚类算法用计算机实现的步骤如下：

(1)初始化中心点。先用 k 平均算法进行聚类，得到 K 个聚类 $C_1,C_2,\cdots,C_K$ 以及每个聚类的心点 $\vec{c}_1,\vec{c}_2,\cdots,\vec{c}_K$，中心点计算公式如下：

$$\vec{c}_j=\frac{1}{|C_j|}\sum_{\vec{x}_i\in C_j}\vec{x}_i \tag{8}$$

其中$|C_j|$表示类 C_j 的元素个数；清空 K 个聚类 $C_j(j=1,2,\cdots,K)$中的元素，仅保留中心点的信息；

(2)用上述的中心对称度量进一步优化聚类。对于数据中的每个模式 $\vec{x}_i$，用公式(5)分别计算与每个类中心点 $\vec{c}_j$ 的中心对称相似性度量。当度量取得最小值时，有

$$j=Arg\min_{j=1,\cdots,K}d(\vec{x}_i,\vec{c}) \tag{7}$$

表明 $\vec{x}_i$ 相对于 $\vec{c}_j$ 具有比其他中心点更好的中心对称性，因此把 $\vec{x}_i$ 归到 $\vec{c}_j$ 所属的聚类 C_j 中；

(3)重新计算中心点。对于每一个聚类 C_j，用(8)式重新计算中心点 $\vec{c}_j$；

(4)如果聚类结果收敛或者达到预定的循环次数，则退出。否则，清空 K 个聚类中的元素，跳回步骤(2)。

4　实验结果

通过实验，发现本文的聚类算法确实具有从数据模式中找到具有中心对称性质的聚类的能力。对一组二维空间上的模式点进行聚类的结果如图2所示。

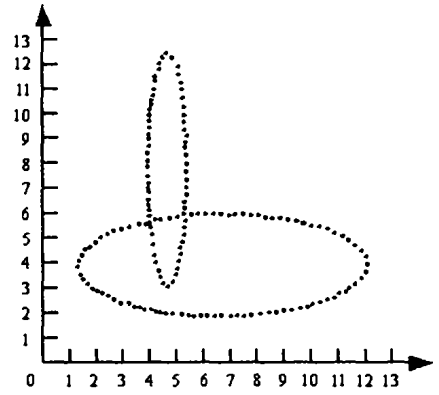

图2(a)

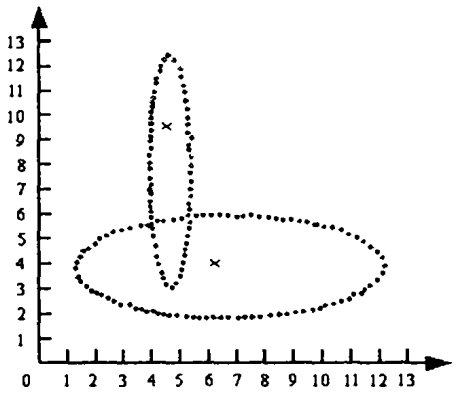

图2(b)

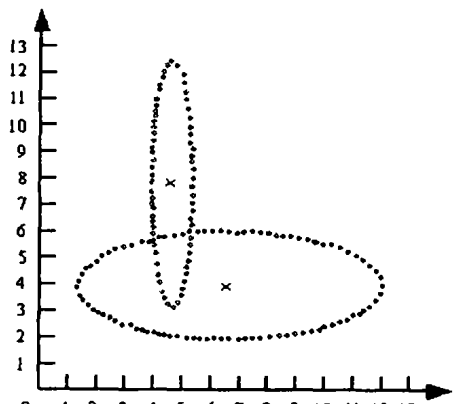

图2(c)

图2　(a)包含两个交错椭圆的数据点分布图(b)用 K 平均算法初始化的聚类图(c)用基于中心对称聚类算法得到的结果

图2(a)是数据模式点在二维空间上的分布图，这些数据模式点排列成两个交错的椭圆；图2(b)是用 k 平均算法聚类后得到的结果，分别用实心圆点和空心圆点表示不同的聚类，交叉点表示类的中心点；图2(c)是用本文提出的中心对称聚类算法得到的结果(K 取2)，已经将两个椭圆基本分开。

图3是另外一组数据模式聚类的结果，图中的两个近似椭圆共有4个交点。

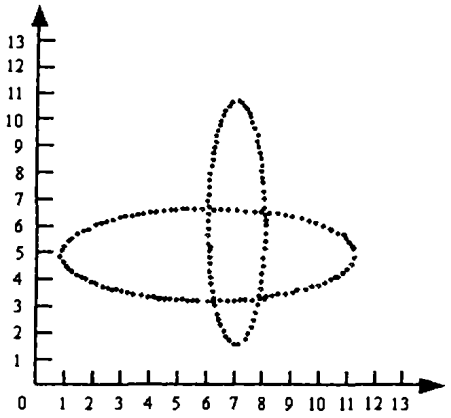

图3(a)

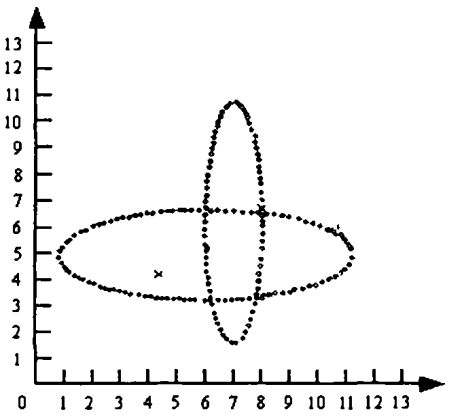

图3(b)

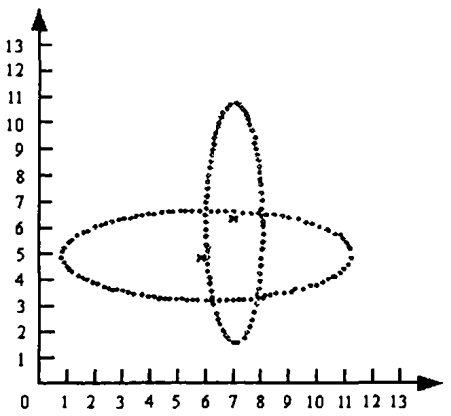

图3(c)

图3 (a)包含两个交错椭圆(四个交点)的数据点分布图(b)用k平均算法初始化的聚类图.(c)用中心对称聚类算法得到的结果

图4(a) 图4(b)

图4(c)

图4 (a)原图片.(b)经过边缘提取和二值化处理，除噪声处理后的图.(c)经过聚类算法后的人脸定位

人脸的形状也是近似椭圆形，所以本文的聚类算法也可用于人脸识别中的人脸检测。如图4所示，图4(a)是一幅包含人脸的图片；图4(b)是经过边缘增强以及二值化后的图，其边缘由一系列数据点组成；图4(c)是用本文算法聚类后检测出的人脸位置。

总结 本文提出的中心对称聚类算法，综合了k平均算法和基于对称的相似性度量算法。实验结果表明，该聚类算法确实能有效地找到有中心对称性质的数据聚类，可以用于人脸识别中的人脸检测等方面。

参考文献

1 Jain A K, Dubes R C. Algorithms for clustering. Englewood Cliffs, N. J. Prentice Hall, 1988

2 Jain A K, Murty M N, Flynn P J. Data clustering: A survey. ACM Comput. Surv., 1999, 31: 264～323

3 MacQueen J. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Statist, Prob., 1967, 1: 281～297

4 Kaufman L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons, 1990

(上接第104页)

7 Fiolhais, Carlos, Trindade, Alberto J. Virtual water, a virtual reality project for learning physics and chemistry. In: Proc. of the 1998 Europhysics Conf. on Computational Physics (CCP 1998)

8 Gourlay D, Lun K C, Guan L. Virtual reality and telemedicine for home health care. Computers and Graphics (Pergamon), 2000, 24 (5): 695～699

9 Werkhven P. Virtual Environment Essential for Designing ship. Computer Graphics, Nov. 1996

10 Xiao tianyuan, et al. Next Generation Manufacturing-Distributed Interactive Virtual Product Development. System Simulation and Scientific. Computing [C]. Beijing China, Oct. 1999

11 Armbruster H. The Flexibility of ATM: Supporting Future Multimedia and Mobile Communications. IEEE Communications Magazine 1995. 2

12 Netmeeting of windows 98, Microsoft Inc