一种基于熵的聚类算法*)

王洪春[1,2] 彭 宏[2]
(重庆师范大学数学与计算机科学学院 重庆 400047)[1]
(华南理工大学计算机科学与工程学院 广州 510641)[2]

摘 要 给出了一种以 Reny 熵为评价准则的聚类算法,通过非参数估计法估计密度函数,再利用类内熵和类间熵进行聚类和确定聚类的数目。这种算法不需要用户输入与聚类有关的参数,能根据由数据的分布的特性自动获取要聚类的数目,并能发现任意形状和任意大小的聚类。实验结果显示了算法的有效性和优越性。

关键词 数据挖掘,熵,聚类算法

A Clustering Algorithm Based on Entropy

WANG Hong-Chun[1,2] PENG Hong[2]
(Dept. of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047)[1]
(Dept. of Computer and Engeering , South China University of Technology, Guangdong 510641)[2]

Abstract A new clustering algorithm using Renyi's entropy as our similarity metric is presented. It estimates density function through the non-parameter estimation, cluster and find cluster number through within-cluster entropy and between-cluster entropy. The algorithm does not need the user input the parameters that related the cluster, can obtain the number automatically according to the distribution characteristic of data set, and can discover the arbitrary shape and the random size cluster. Experimental results show the validity and the superiority of the new algorithm.

Keywords Data mining, Entropy, Clustering algorithm

1 引言

聚类是一个将数据库中的数据划分成具有一定意义的子类,使得不同子类中的数据尽可能相异,而同一子类中的数据尽可能相同的过程。由于聚类技术无须任何应用领域知识就能发现数据中隐含的关系和模式,因此受到了数据挖掘研究人员的广泛重视,并被看作是数据挖掘的主要任务之一。迄今为止,人们提出了许多数据聚类的算法,像 CLARANS[1], BIRCH[2], DBSCAN[3], CURE[4] 等。所有这些算法在性能上各有所长,但都有一定的缺点。

在目前已有的聚类算法中,都是基于某种准则来评价一个已给定划分的特性的,但通常它们需要输入一些参数(如聚类的数目、聚类的密度等),并努力为这些参数定义一个最好的样本集的划分。可见聚类结果需要过多的领域知识,对非专业人员效果较差,而聚类结果通常都与输入参数密切相关;这些参数常常也很难决定,特别是包含高维对象的数据集。这不仅构成了用户的负担,也使得聚类质量难以控制。另外,有些聚类算法只能对某种分布数据聚类效果较好,对其它分布的数据聚类效果较差。许多聚类算法是根据欧氏距离和 Manhattan 距离来进行聚类的,基于这类距离的聚类方法一般只能发现具有类似大小和密度的圆形或球状聚类。比如比较流行的 C-Means 和模糊 C-Means 聚类算法,既需要提供参数——聚类数目,而且对非球型或椭球型分布的数据集聚类效果不理想。

因此,本文认为,要设计一种较好的聚类算法,需要解决:(1)如何提高算法的自主性,减少用户的参与,即需要(由用户)决定的输入参数要最少。(2)如何对任意大小和任意形状的聚类进行分析。实际上一个聚类是可以具有任意形状的,因此设计出能够发现任意形状类集的聚类算法是非常重要的。

针对以上问题,本文给出了一种基于熵的聚类算法,该算法能产生较好的聚类。该算法不需要用户输入任何与聚类模式相关的参数,可以智能地自动完成聚类过程。该算法可以对任意形状和大小的聚类进行分析。试验表明,它是一种较好的聚类算法。

2 信息熵

在物理学中,熵用来描述原子分布的无序程度。当某一系统越有序、越确定时,该系统的热熵越小;在信息论中,信息熵是一个信源发出某一消息所含信息量的度量,当某一信源发出的消息越确定时,该信源的信息熵越小。数据点的分布类似于原子的分布,当聚类的划分越合理,数据点在某一聚类上的归属越确定时,该聚类的信息熵值越小。在聚类分析中,由于在分组前数据点对某一聚类的归属在主观划分上是依赖于用户所选取的算法的,当用户采取不同的算法时,数据点的归属性就不同;而客观上来讲,数据点对某一聚类的归属又是确定的。因此,如果在主观上找到尽可能确定的数据点归属,即求得信息熵值最小的聚类结果,那么聚类的目的就达到了。

现在人们应用和研究的熵,一般指信息熵或由信息熵演化生成的其它熵。信息熵的概念是由信息论的创始人 Shan-

*)广东省科技攻关项目(2004A10202001)、广州市科攻关项目(2004Z2~D0091)。王洪春 副教授,博士,主要研究方向:人工智能、数据挖掘;
彭 宏 教授,博士生导师,主要研究方向:智能网络技术、数据挖掘。

non 于 1948 年提出的。信息理论是应用统计方法的延伸，它规定信息量等于消除的不确定性的数量；若所有不确定性都消除了，则信息量为最大。对于取值离散的样本空间（信源）[5]，有

$$[X \cdot P]:\{X:a_1,a_2,\cdots,a_r$$
$$P(X):p_1,p2,\cdots,p_r\}$$

其中 p_i 为事件 a_i 出现的概率，则事件 a_i 的所含有的信息量（自信息）用 $I(a_i)$ 表示，即

$$I(a_i)=\log\frac{1}{p_i}=-\log p_i$$

自信息 $I(a_i)$ 含有两种意义：在事件 a_i 发生以前，表示其不确定性；事件 a_i 发生以后，则表示其所提供的信息量。由于自信息 $I(a_i)$ 是一个随所发生消息变化的随机变量，不宜用作整个样本空间信息的度量，因此 Shannon 在信息论中定义自信息的数学期望为信源的信息熵，即

$$H(X)=E[-\log p_i]=-\sum_{i=1}^{n}p_i\log p_i$$

但是当随机变量 X 为连续型随机变量，且概率密度函数为 $f_X(x)$ 时，用 Shannon 定义的信息熵就不合适了。于是我们采用 Reny 熵[6]代替信息熵：

$$H_R(X)=\frac{1}{1-\alpha}\log\int f_X^{\alpha}(x)\mathrm{d}x,\alpha>0,\alpha\neq1$$

特别地，当 $\alpha=2$ 时，称为二次 Reny 熵：

$$H_R(X)=-\log\int f_X^2(x)\mathrm{d}x \qquad (1)$$

设 $X=\{x_1,x_2,\cdots,x_N\}$ 为 N 元数据集合，$x_i\in R^n$。如果数据集 X 可以聚类成 K 个子集 $C_1,C_2,\cdots,C_K$，其中 C_k 由数据点 $x_i,i=1,\cdots,N_k$ 组成，于是可采用非参数法估计出各聚类的概率密度函数，比如可以用 Parzen 窗方法估计出各数据集 C_k 的概率密度函数：

$$\hat{f}(x)=\frac{1}{N_k h^n}\sum_{i=1}^{N_k}\psi\left(\frac{x-x_i}{h}\right)$$

这里的 $\psi(x)$ 为核（Kernel）函数，h 为窗宽或平滑系数。在理论上，任何函数均可以用作核函数。但为了密度函数估计的方便性与合理性，通常要求核函数满足以下条件[7]：

(1) $\psi(-x)=\psi(x)$

(2) $\sup|\psi(x)|<\infty,\int_{-\infty}^{\infty}\psi(x)\mathrm{d}x=1$

窗宽 h 过大会使估计结果过于平滑，h 过小会产生过多的数据噪声。h 的值依赖于数据集的特性，窗宽的选择属于非参数密度估计的一般问题。

本文取核函数为辅对称的多变量 Gaussian 函数：

$$\psi(y)=\frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{\|y\|^2}{2}\right)$$

这里 $\|y\|=\sqrt{y^Ty}$

于是

$$\hat{f}_X(x)=\frac{1}{(2\pi)^{n/2}N_k h^n}\cdot\sum_{i=1}^{N_k}\exp\left(-\frac{(x-x_i)^T(x-x_i)}{2h^2}\right) \qquad (2)$$

用密度函数的估计式（2）代替式（1）中的密度函数 $f_X(x)$，可以得到数据集 C_k 熵的估计：

$$H(C_k)=-\log V(C_k) \qquad (3)$$

这里

$$V(C_k)=\int\frac{1}{(2\pi)^{n/2}N_k h^n}\sum_{i=1}^{N_k}\exp\left(-\frac{(x-x_i)^T(x-x_i)}{2h^2}\right)\cdot\frac{1}{(2\pi)^{n/2}N_k h^n}\sum_{j=1}^{N_k}\exp\left(-\frac{(x-x_j)^T(x-x_j)}{2h^2}\right)\mathrm{d}x$$
$$=\frac{1}{(2\pi)^n N_k^2 h^{2n}}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}\exp\left(-\frac{(x_i-x_j)^T(x_i-x_j)}{2h^2}\right)$$

3 聚类思想

我们用图 1 说明聚类的基本思想，如果一个数据集部分已聚成了两类 C_1 和 C_2，如图 1 用圆圈标示的部分。现在有一个新的数据 x，它到底属于哪一类呢？根据熵的物理意义，我们认为，如果把 x 加入到 C_1 后增加的熵小于把 x 加入 C_2 到后增加的熵，则 x 应该属于 C_1 这一类。一般地，如果一个数据集已初步聚成了 K 类：$C_1,C_2,\cdots,C_K$，若满足条件

$$H(C_i+x)-H(C_i)<H(C_k+x)-H(C_k)$$
$$k=1,2,\cdots,K\quad k\neq i \qquad (4)$$

则把数据 x 分配到 C_i，这里 $H(C_k)$ 表示 C_k 的熵。

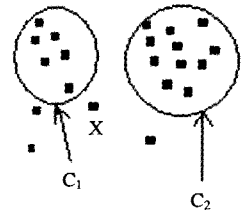

图 1 样本数据的分配

在此基础上，就可以开始初始的动态聚类了。聚类的过程如下：

①将数据集中的每个点单独作为一类并编号；

②任取一点，将其加入其它类中，用（3）式计算这点加入到这类后的熵，根据（4）式，把它分配到熵的增加量最小的类中；

③把分配点后的类重新编号；

④重复②、③，直到没有只由一个点组成的类为止。

从上面的聚类过程可以看出，该过程实际上就是一个循环，每次将一个点加入到一个类中，即每循环一次减少一个聚类数目（减少一个单点类），直到没有单个点组成的类为止。

4 聚类数目的确定

前面的熵的计算都是在各个类内进行的，我们称它为类内熵。由于基于距离的聚类要求聚类的结果应使得类内距离尽可能小，而类间距离要尽可能地大，与此相类似，基于熵的聚类要求应使得聚类后类内熵尽可能小，类间熵尽可能地大，这样才能使类内数据尽可能相似，类间数据差异尽可能地大。因此我们依照文[6]的（12）式再定义两个类的类间熵：

$$H(C_i,C_j)=-\log V(C_i,C_j)$$

其中

$$V(C_i,C_j)=\frac{1}{(2\pi)^n N_i^2 N_j^2 h^{2n}}\sum_{i=1}^{N}\sum_{j=1}^{N}M(x_i,x_j)\exp\left(-\frac{(x_i-x_j)^T(x_i-x_j)}{2h^2}\right)$$

$$M(x_i,x_j)=\begin{cases}1,x_i\in C_i,x_j\in C_j\text{ 或者 }x_i\in C_j,x_j\in C_i\\0,\text{其它}\end{cases}$$

如果 C_1 和 C_2 分得比较开，则 $V(C_i,C_j)$ 比较小，从而 $H(C_i,C_j)$ 比较大；反之 $V(C_i,C_j)$ 比较大，而 $H(C_i,C_j)$ 比较小。于是，这就给我们提供一个合并一些不应该分开的类，并能确定聚类数目的途径，具体步骤如下：

计算任意两个类间的类间熵，合并类间熵最小的两个类，类的总数减少一个，并重新标记每类的样本数。重复这个步骤，直到最后只剩两个类为止。如果在某个时刻计算的两个

（下转第 200 页）

用 Contourlet 变换取代小波变换的同时，虽然 CSPIHT 算法比传统 SPIHT 以更稀疏的方式表示了图像的边缘和纹理特征，但是以牺牲了少许的运算量做为代价。

结论 本文对图像 Contourlet 变换系数的分布特点进行了统计分析，根据其结构特性提出一种基于 Contourlet 变换的空间方向树结构，并通过统计结果验证了该空间方向树具有的"零树"特性。在此基础上，将该空间方向树结构与 SPIHT 算法相结合提出一种渐进式图像编码算法 CSPIHT，该算法在低码率下具有较 SPIHT 更加有效的编码效率；在中等码率下，尽管重构图像的 PNSR 略低于 SPIHT 算法，但重构图像中纹理和边缘区域的视觉效果要优于 SPIHT。

参 考 文 献

1 焦李成，谭山，刘芳. 脊波理论：从脊波变换到 Curvelet 变换. 工程数学学报，2005，22(5)：761～773
2 Do M N, Vetterli M. Contourlets: a new directional multiresolution image representation. Signals, Systems and Computers, 2002(1)：497～501
3 焦李成，孙强. 多尺度变换域图像的感知与识别：进展和展望. 计算机学报，2006，29(2)：177～193
4 Eslami R, Radha H. Wavelet-based contourlet transform and its application to image coding. IEEE Transactions on Image Processing, 2004, 5：3189～3192
5 Chappelier V, Guillemol C, Marinkovic S. Image Coding with Iterated Contourlet and Wavelet Transforms. IEEE Transactions on Image Processing, 2004, 5：3157～3160
6 Said A, Pearlman W A. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. IEEE Transactions on Circuits and Systems for Video Technology, 1996, 6(3)：243～250
7 Burt P J, Adelson E H. The Laplacian pyramid as a compact image code. IEEE Trans Commun, 1983, 31(4)：532～540
8 Do M N, Vetterli M. Framing pyramids. IEEE Trans Signal Proc Sep, 2003
9 Bamberger R H, Smith M J T. A filter bank for the directional decomposition of images: Theory and design. IEEE Trans Signal Proc, 1992, 40(4)：882～893
10 Do M N. Directional multiresolution image representations: [Ph D dissertation]. Swiss Federal Institute of Technology, Lausanne, Switzerland, December 2001. http://www.ifp.uiuc.edu/～minhdo/publications
11 Do M N, Vetterli M. The contourlet transform: an efficient directional multiresolution image representation. IEEE Transactions on Image Processing, 2005, 14(12)：2091～2106
12 沈兰荪，卓力. 小波编码与网络视频传输. 北京：科学出版社，2005

（上接第 179 页）

类最小的类间熵比以前的最小的类间熵显著改变时，根据熵的物理意义知道，这实际上就是一种由无序到有序以及由有序到无序的突变，这时各类分得最开。这时类的总数就是所要聚的类数，所得的结果就是所要聚类的结果。

这里采用的实际上是一种分层聚类方法，根据类减少时类间熵的改变量的突变来确定要聚类的数目。

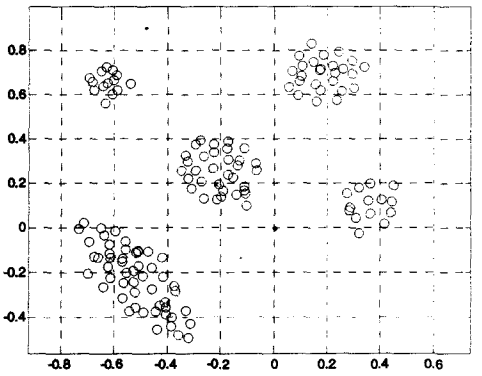

图 2 数据分散点图

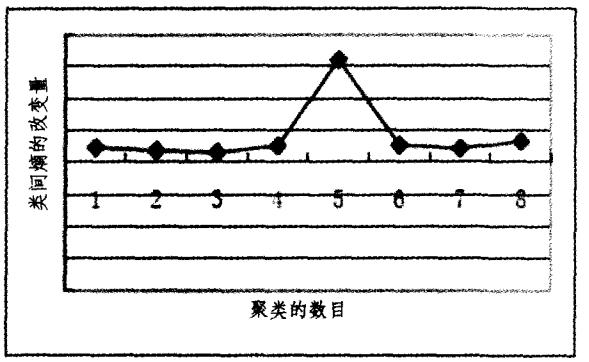

图 3 聚类数目与类间熵改变量间的关系

5 实验结果

以下实验中，计算熵时我们都取 $h=0.5$。首先，我们用人工的方法使用高斯分布随机器产生二维高斯型数据，如图 2 所示。首先初始聚类把它分成 8 类，然后再由类间熵来决定聚类数目。图 3 表示聚类数目与类间熵的改变量间的关系。可以看出，当聚成 5 类时，类间熵改变量异常，其它情况下都比较稳定，因此最终聚类的数目应为 5 类。从图 2 可以看出，虽然每个子类包含的数据的数目和密度不同，用本文的方法却能正确地确定其聚类数并进行聚类。

另外，我们针对于极不规则的类的情形，比如螺旋型、环型分布的数据也进行了试验，利用前面的方法也能比较准确地进行聚类（而传统的 C-Means 算法却办不到），而且聚类的数目也不需要事先指定。

结论 本文给出了一种基于熵的聚类算法，这种算法能够根据数据样本自身的分布特性进行聚类，能自动确定聚类的数目，减少了用户的参与，聚类的结果更具有客观性，同时能发现任意形状和大小的聚类，是一种聚类效果较好的算法。算法也有不足的地方，特别是初始点的选择以及噪音和离群点对结果影响比较大，这是以后的工作还需要进一步探讨的问题。不过可以通过样本筛选，对原始数据样本中的噪音和离群点进行处理，剔除那些假数据，提高数据的可靠性和可分性来消除噪音和离群点对结果的影响。

参 考 文 献

1 Raymond Hau N J. Efficient and effective clustering methods for spatial data mining. In: The 20th VLDB Conference, Santiago, Chile, 1994. 144～155
2 Zhang Tian, Ramakrishnan R, Livmy M. BIRCH: an efficient data clustering method for very large databases. In: Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, Canada, 1996. 103～114
3 Ester M, Kriegej H, Sander J, et al. A density—based algorithm for discovering clusters in large spatial database with noise. In: Proceedings of the 2nd Conference on Knowledge Discovering in Databases and Data Mining, Portland, USA, 1996. 226～231
4 Guha U, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases. Pergamon Information Systems, 2001, 26(1)：35～58
5 史玉峰，史文中，靳奉祥. 熵及其在空间数据不确定性研究中的应用. 计算机工程，2005，31(24)：36～37
6 Gokca E, PRINCIPE I C. Information theoretic clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(2)：158～171
7 李存华，孙志挥，陈耿，等. 核密度估计及其在聚类算法构造中的应用. 计算机研究与发展，2004，41(10)：1712～1719