一种基于边缘特征的聚类学习新方法*)

A New Approach of Clustering Based on Edge Character

刘道海[1] 方 毅[2] 黄樟灿[1]

(武汉理工大学数理系 武汉430070)[1](武汉理工大学电信学院 武汉430070)[2]

Abstract According to mankind's actual classing procedure for substance, this paper puts forward a new Clustering Learning based on Edge Character (CLEC) approach that is different from some traditional methods. This approach does not depend on the number of class sorts that is very difficult to be gotten accurately and for which the result is rather sensitive. On the other hand, CLEC is fit in the processing to distribution data. The experiments sufficiently proved the efficacy of this approach.

Keywords Edge character, Clustering learning

人类认识世界的一种重要方法是将认识对象进行分类，分类可以凭借经验和专业知识来实现，而聚类分析作为一种定量方法，从数据分析的角度，给出了一个更准确、细致的分类工具。作为统计学的一个分支和一种无教师监督的学习方法，聚类分析已有几十年的研究历史，并取得了很多研究成果。目前，聚类学习的典型代表有 K-means 方法和 K-medoid 方法等[3]。

这些聚类分析方法的聚类前提是必须由用户预先确定分类的类别数 K，但事先确定 K 值是一件很困难的事情，并且由于分类结果对 K 很敏感，不同的 K 值往往会得到完全不同的结果，所以由用户给定的 K 所产生的误差将会使得聚类效果很差[2]。

研究人进行聚类分析的过程，我们发现人是通过类边缘的特征提取类的边缘，从而达到确定整个类的数据对象的目的。本文基于由类边缘的特征发现类边缘这一聚类过程，依据数据对象整体和局部分布所形成的反差对比，充分考虑了数据对象与其周围局部数据对象的关系，提出了一种不依赖于 K 值的基于边缘特征的聚类学习新方法(A Clustering Learning based on Edge Character，以下简称 CLEC 方法)。在对几组分布在二维空间的数据对象分析中，CLEC 得到了令人满意的效果。

1 聚类学习的任务及现有主要方法

聚类分析的主要任务是：对分布在 m 维空间的一组样本点 $X_1, X_2, \cdots, X_n$，根据这些样本点在样本空间的距离 d 将它们分成若干块。使其相应的统计误差 $\sum_{l}^{k}\sum_{i=1}^{n} d(X_i, Q_l)$ 取得极小。其中 X_i 为聚类学习的对象，Q_l 为各聚类对象的代表。记样本空间为 Ω，距离的定义是：设 $d(X,Y)$ 是 $\Omega\times\Omega\to R^+$ 的一个函数，它必须满足正定性、对称性和三角不等式。

典型聚类学习方法的主要聚类过程(图1)是：首先由用户给定的聚类个数 K 随机产生 K 个初始聚类，然后通过循环不断地检查聚类情况，通过某种修改准则确定新的聚类对象，当 $\sum_{l}^{k}\sum_{i=1}^{n} d(X_i, Q_l)$ 统计误差达到极小值时终止循环。

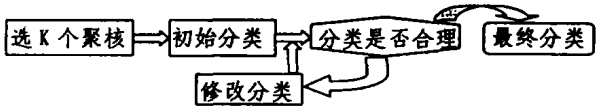

图1 典型聚类学习过程

传统方法最大的弊端是 K 值难确定及分类结果对其过于敏感，这使得算法的效率不高且得到的结果并不理想。针对人并不是由 $\sum_{l}^{k}\sum_{i=1}^{n} d(X_i, Q_l)$ 是否极小来进行分类，一些学者也提出了不同的聚类分析的思路。有人用对象间的平均距离 φ 作为固定的度量标准，由某一对象出发，计算其与其它各对象的距离。若两对象间的距离小于 φ，则两对象归为一类 C。新对象与类 C 中最近点的距离若小于 φ，则新对象归类；否则，由这点出发重新聚类[4]。

用固定的平均距离作为聚类学习的标准会使得算法在处理图2和图3的数据对象时会出现异常情况。图2所示，是一组二维数据的分布。很明显，数据应分为两组。但若根据文[4]的聚类方法，由于图右下部的一类数据高度聚集，使得数据元素间的平均距离 φ 很小，以至在度量图左上部数据的间距时无法使之聚类。而同一分类系统中，元素间的距离在不同类别彼此相差悬殊的情况是常见的[2]。对图3所示的二维数据分布，人们会很自然地分为三类。但由于受一些边缘点或噪声的影响(如图黑圆圈所示)，算法[4]根据聚类标准 φ 将本不属于本类的临界点归类，而由这些点出发将会使另一类整体聚类，最后导致整个数据都聚为一类。这些异常情况都不是我们所希望看到的，究其原因，是因为算法将数据都孤立了，没有考虑一个数据对象与其它所有数据对象之间的联系。

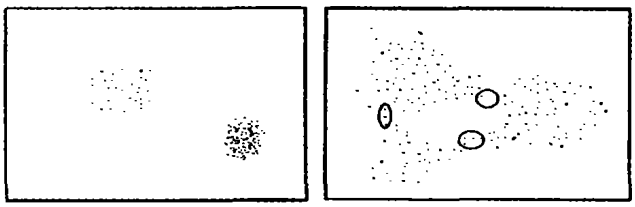

图2 一种二维数据对象分布示意图

图3

*)国家自然科学基金资助项目(编号：70071042，60073043)

2 CLEC 方法

在研究人进行聚类分析的主观思维过程中，我们发现人分类决策的过程极其简单。主要是根据类边缘数据的特征抽取类边缘，确定了类边缘自然就确定了整个类的数据对象。基于人不是直接考察统计误差 $\sum_{j}^{K}\sum_{i=1}^{n}d(X_i,Q_j)$ 来进行聚类判断，这里提出了一种基于边缘特征的新方法，即 CLEC 方法。CLEC 是基于人进行聚类学习的基本原则而设计的，根据类边缘的特征抽取类边缘数据成为了 CLEC 的关键。注意到类边缘这一部分数据对象周围的非数据对象较多，反映在图上就是黑点与白点的数目之比较小。因此，本文类边缘数据的特征主要由数据对象在空间分布的疏密程度来考察。

该方法首先点定义了距离和点的邻域密度的概念，然后提出了聚类搜索的约束条件，最后根据这一条件完成聚类搜索的工作。

1)CLEC 方法采用规格化的 Euclidean 计算公式来定义两 m 维数据对象之间的距离。具体定义如下：

$$d(X,Y)=\left[\sum_{k=1}^{m}(x_k-y_k)^2\right]^{1/2} \tag{1}$$

2)记 $\rho(X)$ 为样本点 X 的邻域密度。邻域是以样本点 X 为圆心，r 为半径的一个球域。若落在这个球内的样品数为 m，则

$$\rho(X)=m \tag{2}$$

3)根据数据对象在 m 维样本空间 Ω 的分布情况，这里采用样本分布平均密度 σ 来描述空间中元素的整体分布情况，以此作为聚类的约束条件 μ。

$$\sigma=\lambda\cdot(1/n) \tag{3}$$

这里样本空间 Ω 已作过归一化处理，n 表示样本空间中样本点的数目，参数 λ 有助于使算法适应不同的分类问题。

4)由样本空间 Ω 中的某一点出发，按照聚类约束条件 μ 进行聚类搜索。若对某一数据对象 X_i，存在 $\rho(X_i)\leqslant\sigma$，我们则称 X_i 满足 μ。这时将 X_i 归入集合 $A(X_j)$。其中：

$$A(X_j)=\{X_k\mid\rho(X_k)\leqslant\sigma\} \tag{4}$$

$A(X_j)$ 表示以数据点 X_j 为初始聚类点出发的集合。

由于类中心部分数据对象的 $\rho(X_j)$ 与类边缘数据的 $\rho(X_j)$ 相差很大，所以 $\rho(X)$ 就可作为类边缘的特征度量。而在从一个类向另一类的空间过渡过程中 $\rho(X)$ 必然发生突变，于是我们可以根据这个突变信号的产生来作为一个聚类过程的完结。

3 算法流程

根据 CLEC 的主要思想，我们给出了 CLEC 方法的主要流程：

```
1. Object(X)=Sample();//从样本空间中取出数据对象集合 Object.
2. If(Object!=Null)
{
3. Seed=RandomSelect(Object);//从对象集合 Object 中任取一对象 Seed，产生初始聚类.
4. Clafy(SeedAround,Seed);i=0;//将 Seed 放入数据预取集合 SeedAround 中.
5. Density=ComputeDensity(SeedAround[i]);//计算当前数据对象 SeedAround[i]的邻域密度.
6. If(Density>D)
{
7. Classfy(SeedAround[i]);//若满足约束条件则聚类.
8. Minus(Object,SeedAround[i]);从对象集合 Object 中减去已聚类的个体.
9. Length = Refresh (SeedAround);//将当前数据对象 SeedAround[i]邻域内的点归入集合 SeedAround 中.
   i++;}
10. If(i<Length)Goto 5;
    Else Goto 2;//若集合数组已到最后一个元素则继续产生新类.
}
```

该算法有两个主要控制参数：r、λ。这两个参数与实际问题紧密相关，我们称之为聚类强度。γ 越小，λ 越大，表明问题对聚类约束条件要求较高，反之则较低。故 CLEC 方法可以适应不同的实际问题需要，算法的鲁棒性较强。

CLEC 在处理数据的过程中，时间上先后顺序与聚类结果无关，也就是该算法很适合作分布式计算，这在处理超大规模数据(千万级及亿级)时是相当有实用价值的。

4 实验结果

我们用 CLEC 对图2和图3的数据进行了处理(图4、图5)，从处理结果可以比较明显地看出 CLEC 的有效性。

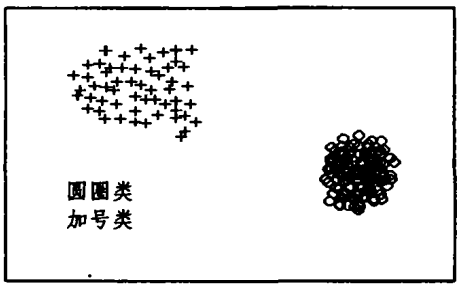

图4 CLEC 对图2数据处理的结果

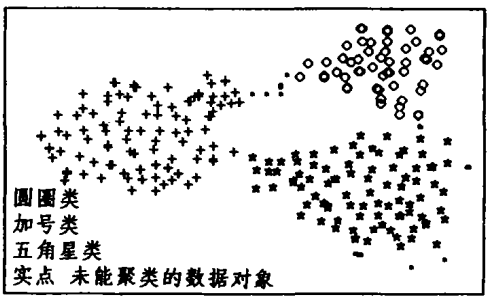

图5 CLEC 对图3数据处理的结果

为了检验 CLEC 在处理大规模数据时的运行能力，这里分别选取从5000到50000个二维数据对象的实际分类问题。适当选择聚类参数，得到了 CLEC 最后的分类结果。表中的聚类质量用平均聚类数/实际类数来衡量。

表1 CLEC 方法对不同数目对象的分类结果

样本点10^6	5	10	20	30	40	50
聚类种数	411	804	1527	2178	3659	4982
聚类质量			0.95			

实验中的参数 r、λ 均分别取0.05(规格化后)。实验结果表明 CLEC 的准确性较高。

结束语 本文基于人在聚类判断过程中所遵循的基本原则，提出了一种新的聚类分析方法——CLEC。与目前大多数聚类分析算法相比，CLEC 具有如下几个特点：

1)无须用户指定类别个数 K，它主要根据人进行聚类决策的过程，依据数据对象整体和局部分布所形成的反差对比，

(下转第129页)

if($l_i=-1$ or $l_i=c_k$ but $|c_k|\leqslant\delta$)
then 报道元组 t_i 是一个例外。

由于波聚类算法具有很强的去噪声能力，能识别任意形状的聚类族，因此数据集中的小比例元组要么被当成噪声去掉，其标识 l_i 的初值不会被改变($l_i=-1$)；要么其所属聚类族 c_k 的族码 $|c_k|$ 小于阈值 δ。而大比例元组所属聚类族的族码大于阈值 δ。由于波聚类算法的时间复杂度为 O(N)，遍历数据库的次数为1次(读入元组并将其分配到相应格)，因此整个算法的时间复杂度为 O(N)，遍历数据库的次数为1次。文[2]中挖掘 DB(P，D)例外的 NL 算法的时间复杂度为 O(kN^2)(k 是特征维数)，遍历数据库的次数≥n−2(n=⌈200/B⌉，B 是缓冲区的百分比)；CS 算法的时间复杂度与 N 成线性关系，遍历数据库的次数最多为3次。由此可见，本文算法 FindCL 优于 NL 算法及 CS 算法。

5 实验

图2是对一个实际时间序列应用算法 FindTS 的实验结果，图中的时间序列 X 为一股票的日成交量。实验中采用了双正交样条小波滤波器，例外数据在高频信号中的能量比 ω=80%。结果表明，原序列 X 中的局部偏离被准确定位。又由于算法只需进行一次小波分解就可将所有 TS(ω)例外同时挖掘出来，故算法具有高效、正确的特点。

随着分解层数的加深，低频部分的频率分辨率增大，使得高频部分中的能量加大，可以根据需要在不同的分解层上挖掘例外(图2给出了在尺度3和尺度5上挖掘到的例外)，因此算法具有多解的特性。

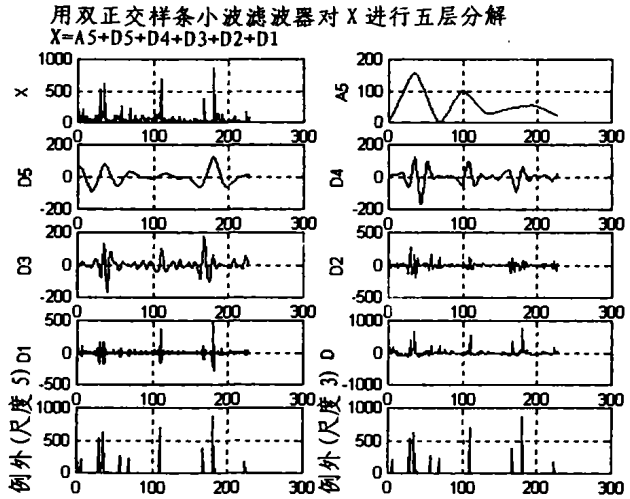

图2 用小波变换对时序进行分解的结果及挖掘到的例外

为便于可视化，图3利用算法 FindCL 对2维图像进行例外挖掘，图像上的每一个像素对应于波聚类算法中的一个格。图中显示了波聚类的结果及算法 FindCL 挖掘到的例外。结果表明，波聚类算法具有很好的聚类结果，基于波聚类算法挖掘到的例外与原图中的例外非常吻合。

图3 波聚类的结果及基于波聚类挖掘到的例外

结束语 例外挖掘是一项重要的、有意义的工作。本文提出了 TS(ω)例外和 CL(δ)例外的概念，并用小波变换的多分辨率分析性质对它们进行挖掘。实验表明，无论对时序数据还是高维数据，基于小波变换的例外挖掘方法都具有较高的效率和较好的结果。

例外数据的挖掘包括例外数据的发现以及例外数据的分析[8]，本文仅仅完成了例外数据的发现，因此下一步的主要工作是进行例外数据的分析。

参考文献

1 Jagadish H V，et al. Mining Deviants in a Time Series Database. In：Proc. of the 25th VLDB Conf. Edinburgh，Scotland，1999. 102～113

2 Knorr E，Ng R T. Algorithms for Mining Distance Based Outliers in Large Databases. In：Proc. of the 24th VLDB Conf. New York：USA，1998. 392～403

3 Knorr E M，Ng R T. Finding Intentional Knowledge of Distance-Based Outliers. In：Proc. of the 25th VLDB Conf. Edinburgh：Scotland，1999. 211～222

4 Sheikholeslami G，et al. WaveCluster：A Multi-Resolution Clustering Approach for Very Large Spatial Databases. In：Proc. of the 24th VLDB Conf. New York：USA，1998. 428～439

5 胡昌华，张军波，等. 基于 MATLAB 的系统分析与设计——小波分析. 西安：西安电子科技大学出版社，2000

6 徐佩霞，孙功宪. 小波分析与应用实例. 合肥：中国科学技术大学出版社，1996

7 Arning A，et al. A Linear Method for Deviation Detection in Large Databases. KDD，1995

8 史东辉，蔡庆生，等. 基于规则的分类数据离群挖掘方法研究. 计算机研究与发展，2000，37(9)：1094～1100

(上接第139页)

通过自适应搜索自动确定聚类个数，而不再是直接考察统计误差，$\sum_{i}^{K}\sum_{i=1}^{n}d(X_i,Q_i)$从而可以实现对数据的智能分类。

2)数据对象的处理过程在时间上无先后顺序之分，这一特性使得算法在处理分布式数据时效率很高。

3)算法实现简单，时间复杂度为 $o(n^2)$，使之能有效地完成大规模数据的聚类分析工作。

CLEC 方法的提出为从其它角度重新审视传统的聚类分析方法提供了新的思路。

参考文献

1 Bobrowski L，Bezdek J C. c-Means Clustering with the l_1 and l_∞ Norms. IEEE Transation on Systems，Man and Cybernetics，1991，21(3)：545～554

2 Milligan G W，Cooper M C. An Examination of Procedure for Detecting the Number of Clusters in a Data Set. Psychometrika，1985，50：159～179

3 Leouski A V，Croft W B. An Evaluation of Techniques for Clustering Search Results：[Technical Report IR-76]. Deparment of Computer Science，University of Massachusetts. 1996

4 Zhu Ming ，Wang Jun Pu. A New Approach of Clustering. Patten Recognizationand Artificial Intelligence. 262～265