Research and Experiment on Affinity Propagation Clustering Algorithm

Huan Zhang
College of Mechanical and Electronic Engineering
Qingdao Agriculture University
Qingdao, P. R. China
huan0804@163.com

Kun Song
College of Automation and Electrical Engineering
Nanjing University of Technology
Nanjing, P. R. China
automationsk@163.com

***Abstract*—This paper introduces Affinity Propagation (AP) clustering algorithm, which is intensively researched by some scholars owing to its advantage of fast speed and no need of setting the initial clusters manually. Mainly analyzed the characteristics of Affinity Propagation clustering algorithm at first, and then compared several principle similarity calculating methods based on Euclidean distance and Mahalanobis distance and etc. Experiment on AP clustering algorithm were done with the parts of the UCI data sets, thus the effectiveness of this algorithm was verified. Finally, the experimental results were analyzed in general.**

Keywords-affinity propagation clustering; similarity; Euclidean distance; Mahalanobis Distance; UCI data set

I. 引言

作为数据挖掘（Data Mining）的一个重要分支，聚类方法引起了人们的关注，广泛地应用于模式识别、图像处理、系统建模、信息处理和人工智能等学科中。聚类是根据相似性的原则，将研究对象进行归类。使同一类的元素相似性最大、不同类的元素差异性最大。常用的聚类方法有 k 均值（k-means）、模糊 c 均值（Fuzzy c-means，FCM）等。这些聚类方法需要人工设置初始的聚类中心和数目。如果初始聚类中心选择不当，容易陷入局部极值。在某些应用领域，由于事先不知道这些对象的工况数目(即聚类数目)，所以这些聚类方法的应用受到一定的限制。

仿射传播（Affinity Propagation，AP）聚类是由 Frey 等人于 2007 年在 SCIENCE 杂志上提出的一种快速、有效的聚类算法，是一种基于划分的聚类方法[1]。该算法目前在人脸图像的聚类、“基因外显子”发现、搜索最优航线等领域得到了应用。它不需要事先确定聚类中心的数目，各样本点通过迭代竞争聚类中心。在事先未知聚类数目情况下，能够自动准确的聚类。为了提高聚类的准确性，有的学者还对算法进行了改进[2, 3]。本文主要介绍了仿射传播聚类算法及其聚类效果评价指标，针对不同数据特征，介绍了几种相似性计算方法。最后利用 UCI 数据集的部分数据进行了聚类实验，并取得了较好的实验结果。

II. 仿射传播聚类概述

A. 相似性计算方法

基于划分的聚类方法聚类的依据是数据点之间的相似度，这个相似度的计算方法一般包括：欧氏距离、明氏距离、马氏距离和切比雪夫距离等。

设 m 个属性 n 个数据点组成的 $m\times n$ 对象矩阵 x_{ij}(i=1，2，…，n；j=1，2，…，m)。d_{ij}表示对象 x_i和 x_j之间的距离。

（a）欧氏距离(Euclidean distance)，即各个属性之差的平方和的平方根。

$$d(x_i, x_j) = \sqrt{\sum_{t=1}^{m} \left| x_{it} - x_{jt} \right|^2} \tag{1}$$

（b）明氏距离(Minkowski distance)。

$$d_{ij} = \left(\sum_{t=1}^{m} \left| x_{it} - x_{jt} \right|^p \right)^{\frac{1}{p}} \tag{2}$$

（c）马氏距离(Mahalanobis distance)。

$$d(x_i, x_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)} \tag{3}$$

其中，S 是由样品 x_1，x_2，…，x_n算得的协方差矩阵：

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^{\mathrm{T}} \text{，} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{。}$$

（d）切比雪夫距离(Chebyshev distance)，即各个属性之差的绝对值的最大值。

$$d_{ij} = \max_{1 \le t \le n} \left| x_{it} - x_{jt} \right| \quad (i, j=1,2,\ldots,n) \tag{4}$$

B. 仿射传播聚类算法

假设样本数目为 N 的数据样本，仿射传播聚类一开始是将所有 N 个样本都看成潜在的聚类中心。根据每个样本与其他样本吸引度的信息，各样本点竞争最终的聚类中

978-1-4244-9439-2/11/$26.00 ©2011 IEEE

心。设样本数为 N 的集合 $X=(x_1, x_2, \cdots, x_N)$，基于欧氏距离[公式(1)]的点 x_i 和 x_j 的相似度矩阵为：

$$S(i,j)=\begin{bmatrix} p & d_{12} & \cdots & d_{1N} \\ d_{21} & p & \cdots & d_{2N} \\ \vdots & \vdots & & \vdots \\ d_{N1} & d_{N2} & \cdots & p \end{bmatrix} \quad (5)$$

其中 p 为偏向参数，可取 $S(i,j)$的中值。p 越大表示各个样本点成为聚类中心的倾向性越大，在迭代中表现为竞争越激烈，竞争激烈的结果是产生的聚类中心数多。反之，偏向参数 p 较小时，各个样本点成为聚类中心的倾向性变小。迭代缺乏竞争，则导致聚类数目变少。$S(i,j)$的大小表示样本点 j 对样本点 i 吸引程度的大小，也就是说样本点 i 认为样本点 j 为聚类中心的归属感强弱。传统的相似度计算一般是基于欧式距离，但是根据数据特征的不同，也可以采用别的相似性计算方法，如马氏距离等。这样，处于聚类中心处的样本点到其他点之间的吸引度之和较大，成为聚类中心的可能性就大。相反，处于聚类边缘的样本点到其他样本点的吸引度之和较小，成为聚类中心的可能性就小。为了找到合适的聚类中心，就需要不断地从数据样本中搜集相关证据，证明可以作为聚类中心的样本点。相似度 $S(i,j)$越大表示样本点 k 对样本点 i 的吸引程度（Responsibility）越大，也可以说样本点 i 认为样本点 k 为聚类中心的归属感（Availability）越强。相反，处于聚类边缘的样本点到其他点的吸引度之和较小，成为聚类中心的可能性就小。为了找到合适的聚类中心就需要从数据样本中不断的搜集相关证据。这个证据就是 $R(i,j)$和 $A(i,j)$。证据越强，即 $R(i,j)$和 $A(i,j)$之和越大，样本点 k 作为聚类中心的可能性就越大。$R(i,j)$和 $A(i,j)$的迭代公式为：

$$R(i,k)=S(i,k)-\max_{j\neq k}\{A(i,j)+S(i,j)\} \quad (6)$$

$$A(i,k)=\min\left\{0, R(k,k)+\sum_{j\notin\{i,k\}}\max\{0, R(j,k)\}\right\} \quad (7)$$

更新证据公式为：

$$R^{new}(i,k)=(1-\lambda)R(i,k)+\lambda R^{old}(i,k) \quad (8)$$

$$A^{new}(i,k)=(1-\lambda)A(i,k)+\lambda A^{old}(i,k) \quad (9)$$

其中 λ 为是 $R(i,j)$和 $A(i,j)$迭代的阻尼系数，调整 λ 可以改变迭代的快慢。λ 一般取值在 0～1 之间，λ 越大迭代越慢，但会提高聚类精度。λ 越小迭代加快，但可能发生震荡。算法的输入是聚类样本，输出时聚类中心和各个样本点的类属情况。

算法的主要步骤如下[4]：

（a）参数的初始化。根据相似性计算公式(1～4)计算相似度矩阵 S，见公式(5)。设置初始化参数 p 和阻尼系数 λ。即 p 可取 $S(i,j)$的中值，λ 可取值 0.6。初始化 $R(i,j)$，$A(i,j)$，$(i=1, 2, \cdots, n; j=1, 2, \cdots, n)$。

（b）证据更新。由公式(6)和(8)更新证据矩阵 $R(i,k)$，$(i=1, 2, \ldots, n; k=1, 2, \cdots, n)$。根据公式(7)、(9)更新证据矩阵 $A(i,k)$，$(i=1, 2, \cdots, n; k=1, 2, \cdots, n)$。

（c）寻找每个样本点的聚类中心。聚类中心的寻找方法是对于数据点 i，若数据点 k 使得 $A(i,k)+R(i,k)$ 是 $\{A(i,1)+R(i,1),\ldots,A(i,k)+R(i,k), \cdots, A(i,N)+R(i,N)\}$中的最大值，那么数据点 k 就认为是数据点 i 的聚类中心。

（d）检查是否满足迭代终止条件。迭代终止条件一般设置为达到最大迭代次数或是算法收敛，即聚类中心经过 l 次循环后不再变化。若满足这个条件，则迭代结束，否则转步骤（b）。

（f）迭代过程结束。输出聚类中心和各个样本点的类属结果。

仿射传播聚类算法流程图，如图 1 所示。

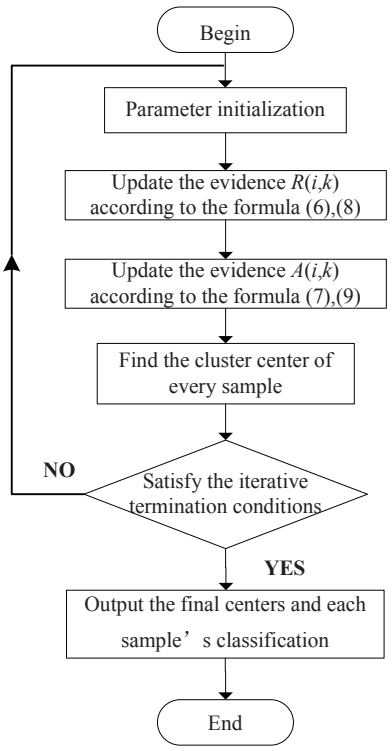

图 1 仿射传播聚类算法流程图

III. 算法的实验

A. 基于欧式距离的相似性计算

UCI 数据集(http://archive.ics.uci.edu/ml/)中的 ecoli、vehicle、ionosphere、libras、wine 数据进行算法的聚类实验。其中，ecoli 取 cp(cytoplasm)、im(inner membrane without signal sequence)、pp(perisplasm)三类，选取的各数据集合的信息，见表 1。

表 1 数据集信息

数据集	样本数	维数	类数
ecoli	272	7	3
vehicle	92	18	4
ionosphere	351	34	2
libras	72	90	3
wine	178	13	3

采用仿射传播聚类算法对各个数据集数据进行聚类实验，其中 ecoli 数据的聚类结果分别如图 2、图 3 所示。

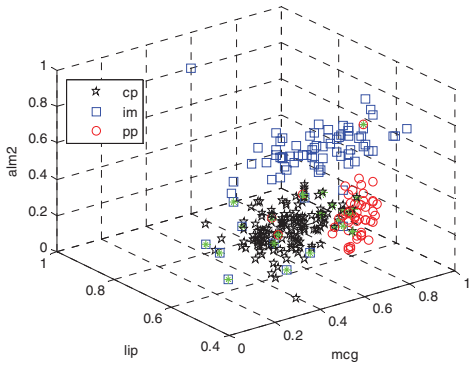

图 2 ecolis 数据聚类结果(a)

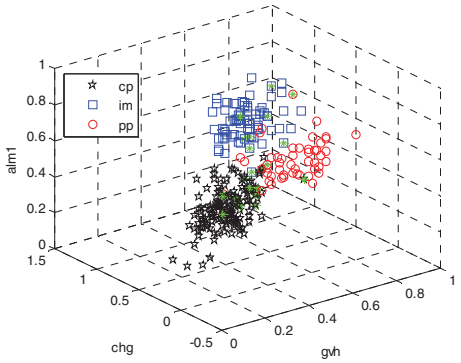

图 3 ecolis 数据聚类结果(b)

为了评价聚类的效果，引入 Silhouette 指标，它反映了聚类结构的类内紧密性和类间可分性[5]。设一个具有 n 个样本，k 个聚类 $Ci(i=1，2，\cdots，k)$的数据集，某一个样本 t 的 Silhouette 指标为：

$$S_{il}(t)=\frac{\min\{d(t,C_i)\}-a(t)}{\max\{a(t),\min\{d(t,C_i)\}\}} \tag{10}$$

其中，d(t,Ci)为 C_i 的样本 t 到另一个类 C_j 的所有样本的平均不相似度或距离，$a(t)$为聚类 C_i 中的样本 t 与 C_i 内所有其他样本的平均不相似度或距离。一个数据集中所有样本点 Silhouette 指标的平均值可以反映整个数据集聚类的质量。*Sil-av* 的值越大表示聚类质量越高，*Sil-av*>0.5 说明各个类能明显的分开，*Sil-av*<0.5 说明一些类有重叠的情况，*Sil-av*<0.2 说明缺乏实质性的聚类结构。各数据的聚类 Silhouette 指标，见表 2。

表 2 各个数据的 Silhouette 指标

数据集	AP	FCM	K-means
ecoli	0.614275	0.597412	0.608214
vehicle	0.703831	0.422121	0.522163
ionosphere	0.412664	0.409660	0.408062
libras	0.463452	0.401132	0.436976
wine	0.460790	0.537322	0.537959

实验分析：通过对不同数据集的聚类，可以看出 AP 算法能够较为准确的聚类。对于比较松散的数据对象，AP 算法倾向于产生较多的局部聚类来实现好的聚类效果。当为了达到标准分类数，需人为地调节偏向参数 p 时，过小的 p 会使聚类精度降低。这就是 wine 数据 AP 聚类 Silhouette 指标较差的原因。

B. 基于马氏距离的相似性计算

采用文献的数据进行聚类实验[6]。样本数据为为研究心肌梗塞的危险，考察两组人群，分别是心肌梗塞和正常组。一共考察两个指标 X_1：总胆固醇和 X_2：高密度脂蛋白胆固醇。现在就对这共 46 组 2 维数据集合进行聚类。聚类的结果，如图 4 和图 5 所示。定义聚类误差为以实际分类情况为标准，被错误分类的样本点占总样本点的百分比。基于欧式距离相似性计算的 AP 聚类误差为 0.26087，而基于马氏距离相似性计算的 AP 聚类误差为 0.23913。

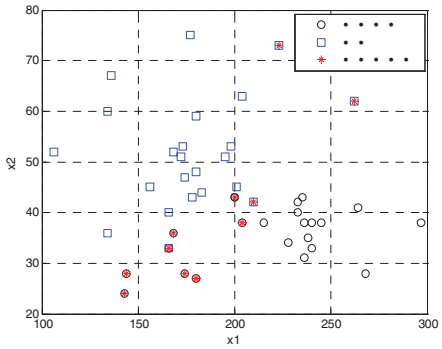

图 4 基于欧式距离的相似度计算

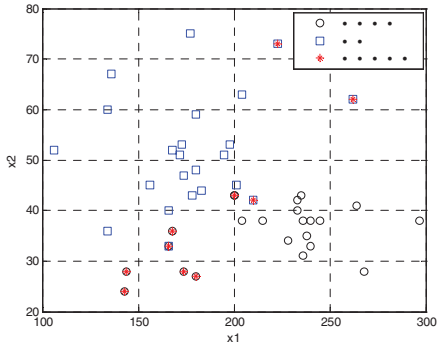

图 5 基于马氏距离的相似度计算

实验结果分析。每个数据对象都有自己的特征，有些数据对象属性之间是有联系的。基于马氏距离的相似性计算更能适应这种情况，能够产生较好的聚类结果。

IV. 结语

本文对仿射传播聚类算法进行了分析，对影响聚类效果的算法参数和相似性度量方法进行了讨论。最后利用 UCI 数据集的部分数据进行了聚类实验，通过实验比较了基于欧式距离和马氏距离相似度计算的仿射传播聚类，实验结果验证了仿射聚类算法的有效性和快速性。

致谢

本文的撰写得到了青岛农业大学机电工程学院领导和电气工程教研室老师们的帮助。

References

[1] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," Science. AAAS Washington, vol. 315, pp.972-976, February 2007.

[2] K. J. Wang, J. Li, and J. Y. Zhang et al, Self-supervised affinity propagation clustering. Computer Engineering, Vol 33, pp.197-201, December 2007 (In Chinese).

[3] K. J. Wang, J. Y. Zhang and D. Li et al, Self-adaptive affinity propagation. Acta Automatica Sinica, Vol 33, pp.1242-1246, December 2007 (In Chinese).

[4] Y. Q. Li and H. Z. Yang, Method for multi-model building based on affinity propagation clustering and Gauss process. Computers and Applied Chemistry,:Vol 27, pp.51-54, January 2010 (In Chinese).

[5] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: John Wiley and Sons, 1990.

[6] C. L. Mei and J. F. Fan, The Method for Data Analysis. Beijing: China Higher Education Press, 2006.