基于互近邻一致性的近邻传播算法*

邢 艳, 周 勇
(中国矿业大学 计算机学院, 江苏 徐州 221116)

摘 要: 近邻传播(AP)算法是一种新提出的聚类算法,是在数据点的相似度矩阵的基础上进行聚类,通过数据点之间交换信息,最后得到聚类结果。提出了基于互近邻一致性近邻传播算法,即 KMNC-AP 算法,该算法利用互近邻一致性调整数据点之间的相似度,进而提高聚类效率和精确度。实验结果表明,该算法在处理能力和运算速度上优于原算法。

关键词: 近邻传播算法; 互近邻一致性; 相似度; 数据挖掘
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)07-2524-03
doi:10.3969/j.issn.1001-3695.2012.07.032

Affinity propagation based on K-mutual nearest neighbor consistency

XING Yan, ZHOU Yong
(*School of Computer Science & Technology, China University of Mining & Technology, Xuzhou Jiangsu* 221116, *China*)

Abstract: Affinity propagation is a new clustering method. It based on the similarity between pairs of data points, through the exchange of information between data points, and finally obtained the final clustering results. This paper presented an improved AP clustering based on K-mutual nearest neighbor consistency KMNC-AP. The improved algorithm used the idea of K-mutual nearest neighbor consistency to adjust the similarity between data points. Experiments show that the improved algorithm is more accurate and faster than the original algorithm.

Key words: affinity propagation(AP) algorithm; K-mutual nearest neighbor consistency; similarity; data mining

0 引言

聚类分析是数据挖掘的一种有效方法。聚类是根据研究对象特征对研究对象进行分类的一种多元分析技术,把性质相似的个体归为一类,使得同一类中的个体都具有高度的同质性,不同类之间的个体具有高度的异质性。聚类的应用是非常广泛的,无论是在商务上,还是在市场分析生物学、Web 文档分类等领域中都得到了充分的应用。聚类分析是一个蓬勃发展、具有很强挑战性的领域,它的一些潜在应用对算法提出了特别的要求,如可扩展性、处理不同数据类型的能力、发现具有任意形状的聚类的能力、输入参数对领域知识的最小限度的依赖性、能够处理异常数据的能力、数据输入顺序对聚类结果的不敏感性、处理高维数据的能力、基于约束的聚类以及聚类结果的可解释性和可用性[1]。

迄今为止,研究人员已经提出了很多数据聚类算法,比较著名的有 K-means、K-medoids、CLARANS、BIRCH、DBSCAN 等算法。

近邻传播(AP)算法是由 Frey 等人(Science)提出来的一种新的聚类算法。与以往的聚类方法相比,近邻传播算法可以更快地处理大规模数据,得到较好的聚类结果。文献[2]中将近邻传播算法应用在人脸图像聚类、基因表达数据的基因识别、手写体字符识别、最优航空路线确定等问题上,实验结果表明, 近邻传播算法在很短的时间内就能得到 K-means 算法花费很长时间才能达到的聚类结果[2];并将所有的点作为候选的类代表点,不限制类代表点的数量,避免了聚类结果受初始聚类中心选取的影响[3]。近邻传播算法的另一个优点是,它对数据形成的相似矩阵的对称性没有任何要求[2],这就扩大了其应用范围。该算法快速、有效,引起了学者的广泛关注。

本文提出基于互近邻一致性的近邻传播聚类(AP based on K-mutual nearest neighbor consistency, KMNC-AP)算法。该算法能够充分利用数据的空间一致性,指导聚类过程,并能提高聚类的准确性和运行效率。

1 近邻传播算法

近邻传播算法根据 N 个数据点之间的相似度进行聚类,这些相似度可以是对称的,也可以是不对称的。这些相似度组成 $N \times N$ 的相似度矩阵 S。其中相似度 $S(i,k)$ 表示点 k 作为点 i 的聚类中心的合适程度。

AP 算法不需要事先指定聚类数目,它将所有的数据点都作为潜在的聚类中心,称之为 exemplar。以 S 矩阵的对角线上的数值 $s(k,k)$ 作为 k 点能否成为聚类中心的评判标准,该值越大,这个点成为聚类中心的可能性也就越大,这个值又称做参考度 p(preference)。聚类的数量受到参考度 p 的影响,如果每个数据点都有相同的可能作为聚类中心,那么 p 就应取相同的值。

AP 算法中传递两种类型的消息,即吸引度(responsibility)

收稿日期: 2011-12-30; **修回日期**: 2012-02-03 **基金项目**: 国家教育部博士点基金资助项目(20100095110003);国家博士后科学基金资助项目(20070421041);江苏省博士后科学基金资助项目(0701045B);中国矿业大学科技基金资助项目(2007B017)

作者简介:邢艳(1987-),女,硕士研究生,主要研究方向为半监督聚类分析(xingyan_cumt@163.com);周勇(1974-),男,副教授,硕导,博士,主要研究方向为数据挖掘、进化计算、计算机控制.

和归属度(availability)。$r(i,k)$表示从点 i 发送到候选聚类中心 k 的数值消息,反映 k 点是否适合作为 i 点的聚类中心;$a(i,k)$则表示从候选聚类中心 k 发送到 i 的数值消息,反映 i 点是否选择 k 作为其聚类中心。$r(i,k)$与 $a(i,k)$越大,则 k 点作为聚类中心的可能性就越大,并且 i 点隶属于以 k 点为聚类中心的聚类可能性也就越大。AP 算法通过迭代过程不断更新每一个点的吸引度和归属度值,直到迭代过程收敛,聚类中心也随之固定,同时将其余的数据点分配到相应的聚类中。

AP 算法的具体工作过程如下:先计算 N 个点之间的相似度值,将值放在 S 矩阵中,再选取 P 值(一般取 S 的中值)。设置一个最大迭代次数(如 1 000),迭代过程开始后,计算每一次的 R 和 A 值,根据 $R(k,k)+A(k,k)$值来判断是否为聚类中心(文献[2]中指定,当$(R(k,k)+A(k,k))>0$ 时,认为是一个聚类中心),当迭代次数超过最大值或者当聚类中心连续多少次迭代不发生改变时终止计算(文献[2]中设定连续 50 次迭代过程不发生改变时终止计算)。

2　互近邻一致性

KNN 算法即 K 最近邻法,是一个理论上比较成熟的方法,最初由 Cover 和 Hart 于 1968 年提出。KNN 算法的基本思想是:根据距离函数计算待分类样本 x 和每个训练样本的距离,选择与待分类样本距离最小的 K 个样本作为 x 的 K 个最近邻,最后根据 x 的 K 个最近邻判断 x 的类别。如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别,则该样本也属于这个类别。2004 年,Ding 等人[4]将这个概念扩展到数据聚类,提出 KNN 一致性和 KMN 一致性,对一个类中的任意数据点,要求它的 k 最近邻和 k 互最近邻都必须在该类中。

KMN 一致性是对 KNN 一致性的加强,它定义了对称的最近邻关系,是一种更强大更严格的近邻,能够更好地反映数据点之间的关系。

定义　互近邻(mutual nearest neighbor,MN)。如果点 x 的最近邻点是点 y,而且点 y 的最近邻点是点 x,那么点 x 的互最近邻是点 y,点 y 的互最近邻是点 x。

一般地,如果点 x 是点 y 的第 p 近邻点(p-thNN),而点 y 是点 x 的第 l 近邻点(l-thNN),令 k 等于 p 和 l 的最大值,那么点 x 是点 y 的 k 互近邻点(k-thMN),点 y 也是点 x 的 k 互近邻点。

由定义可以看出,一些数据的 k 互近邻点可能并不存在。如果点 x 的 1NN 是 y,而 y 的 1NN 是 z,那么点 x 的 1MN 就不存在。

3　基于互近邻一致性的近邻传播聚类

3.1　算法的基本思想

AP 算法中最重要的参数之一是相似度矩阵 S,如果相似度矩阵能够准确地给出数据之间的相似关系,则 AP 算法就能够得到很好的聚类结果。相似度矩阵的定义直接影响到基于相似度矩阵的聚类算法的性能。

本文提出了利用 KMN 一致性调整 AP 算法相似度的改进算法,即基于 KMN 一致性的近邻传播聚类(KMNC-AP)算法。算法认为 K 互近邻点对必须在同一个聚类中,即将 K 互近邻点的相似度设置为 0。

3.2　算法的基本步骤

a)初始化。计算相似度矩阵$[s(i,k)]_{n\times n}$,取相似度矩阵对角线元素 $s(k,k)$为一相同值 Pm,确定阻尼系数 lam、最大迭代次数 maxRun 及聚类划分连续不变次数 Convits。

b)调整相似度矩阵。将互为 K 近邻的数据点之间的相似度调整为 0,其中 K 值通过多次实验确定。调整步骤如下:

(a)计算每个数据点的 K 近邻点;判断数据点对是否互为 K 近邻点,调整相似度矩阵,将互为 K 近邻点的数据点对的相似度设置为 0。

(b)进一步调整相似度矩阵:如果 $s(i,j)=0$,$s(i,k)=0$,那么 $s(i,k)=0$。

c)信息传递。在调整后的相似度矩阵的基础上执行 AP 算法的信息传递过程。

(a)更新吸引度矩阵 R_{new}:

$$R(i,k)=S(i,k)-\max\{A(i,k')+S(i,k')\}$$
$$k'\in\{1,2,\cdots,N\}\text{ 且 }k'\neq k$$

(b) 更新归属度矩阵 A_{new}:

$$A(i,k)=\min\{0,R(k,k)+\sum_{i'}\{\max(0,R(i',k))\}\}$$
$$i'\in\{1,2,\cdots,N\}\text{ 且 }i'\neq i,i'\neq k$$
$$A(k,k)=\sum_{i'}\{\max(0,R(i',k))\}\quad i'\in\{1,2,\cdots,N\}\text{ 且 }i'\neq k$$

(c)$R_{new}=(1-\text{lam})\times R_{new}+\text{lam}\times R_{old}$

(d)$A_{new}=(1-\text{lam})\times A_{new}+\text{lam}\times A_{old}$

d)循环执行 c),直到满足终止条件。信息传递的终止条件为迭代次数超过设定的最大迭代数目 maxRun 或聚类划分连续 Convits 次不变。

e)输出聚类结果。输出类代表点及数据集中所有点的划分和聚类精度。

4　实验与分析

实验的计算机环境:处理器为 Pentium 2.6 GHz,内存 1 GB,硬盘为 80 GB,操作系统为 Windows XP,编程语言为 C#。

为了测试本文算法的有效性,实验选择 AP 算法作为对比算法。

4.1　实验数据集

本节选用 UCI 标准数据集中的 Iris、Glass 和 Wine 数据集作为测试数据集对本文提出的算法进行验证和比较。表 1 给出了实验数据集的大小、维数、包含类的数目以及各类的大小等数据集特性。

表 1　测试数据集描述

名称	大小	类簇数目	维数	各类的大小
Iris	150	3	4	50,50,50
Wine	178	3	13	59,71,48
Glass	214	6	9	70,76,17,13,9,29

4.2　实验评价标准

算法聚类的结果可以通过数据库提供的数据标签评估,本文采用聚类精度作为评价指标。聚类精度定义为

$$\text{accuracy}=TC/(TC+FC)$$

其中:TC 为正确聚类的样例数量,FC 为错误聚类的样例数量。

算法的好坏还要考虑其效率。本文从算法所用时间方面对改进算法进行验证。由于算法的主要时间用在消息迭代传播过程中,本文将调用消息传播的次数来比较两种算法在效率

方面的优劣。

4.3 实验

实验 1 研究 K 值对 KMNC-AP 算法的影响

由于 KMNC-AP 算法中 K 值的设置对聚类算法影响很大，实验设置不同的 K 值，在不同的数据集上运行 KMNC-AP 算法，研究不同的 K 值对 KMNC-AP 算法聚类结果的影响。

图 1 为不同 K 值对聚类精度的影响。图 2 为不同 K 值对迭代次数的影响。

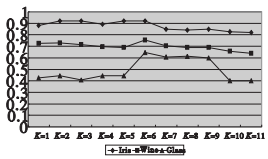

图1 不同K值对聚类精度的影响

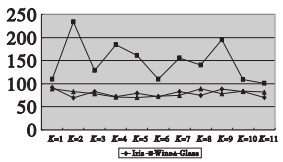

图2 不同K值对迭代次数的影响

实验结果表明，对于不同的数据集，最优 K 值的选择是不一样的，这与数据集的具体结构有关。从图 1 可以看出，当 K 值超过 6 时聚类结果的精确度都将下降。从图 2 可以看出，聚类趋于稳定所需的迭代次数并没有明显的上升或下降趋势，波动性比较大。

实验 2 比较 KMNC-AP 和 AP 算法

AP 和 KMNC-AP 两种算法均采用相同的初始参数，例如偏向参数 $p = Pm$（Pm 表示数据集全体样本相似度的中位数），阻尼系数 lam = 0.5。KMNC-AP 算法的 K 值取实验 1 中聚类结果最好的 K 值。数据集采用欧式距离的相反数作为样本之间的相似度。表 2 为 KMNC-AP 与 AP 算法在三个 UCI 数据集上的比较。

表 2 KMNC-AP 与 AP 算法在三个 UCI 数据集上的比较

数据集	KMNC-AP 聚类精度	AP 聚类精度	KMNC-AP 迭代次数	AP 迭代次数
Iris	0.92	0.91	69	76
Wine	0.72	0.71	110	146
Glass	0.64	0.62	73	79

在大量测试数据集上的实验结果表明，KMNC-AP 算法比 AP 算法具有较好的聚类性能，虽然在 Iris 数据集上，KMNC-AP 算法的聚类精度比 AP 算法低一些，但是在迭代次数上比 AP 算法要少很多，所有总体上 KMNC-AP 算法能给出令人满意的聚类结果。

5 结束语

本文将 KMN 一致性与近邻传播算法相结合，提出了基于 KMN 一致性的近邻传播聚类（KMNC-AP）算法。通过实验证明了改进算法的有效性，改进的算法不仅在聚类精度上优于原始的 AP 算法，而且在迭代次数上也比 AP 算法少。

参考文献：

[1] ZHOU Yong, XING Yan. Summary of affinity propagation[J]. Advanced Materials Research, 2011, 268-270: 811-816.

[2] FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[3] MEZARD M. Where are the exemplars[J]. Science, 2007, 315(5814): 949-951.

[4] DING C, HE Xiao-feng. K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization[C]// Proc of ACM Symposium on Applied computing. New York: ACM Press, 2004: 584-589.

[5] HART P E. The condensed nearest neighbor rule[J]. IEEE Trans on Information Theory, 1968, 14(3): 515-516.

[6] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61.

[7] 闭小梅，闭瑞华. KNN 算法综述[J]. 科技创新导报，2009(14)：31.

[8] 计华，张化祥，孙晓燕. 基于最近邻原则的半监督聚类算法[J]. 计算机工程与设计，2011，32(7)：2455-2458.

[9] 董俊，王锁萍，熊范纶. 可变相似性度量的近邻传播聚类[J]. 电子与信息学报，2010，33(3)：509-514.

[10] NAGENDRASWAMY H S, GURU D S. K-mutual nearest neighbour approach for clustering two-dimensional shapes described by fuzzy-symbolic features[J]. Engineering Letters, 2007, 14(1): 143-153.

[11] GOWDA K C, KRISHNA G. Agglomerative clustering using the concept of mutual nearest neighborhood[J]. Pattern Recognition, 1978, 10(2): 105-112.