面向限制 K-means 算法的迭代学习分配次序策略

邱 烨 何振峰
(福州大学数学与计算机科学学院 福州 350108)

摘 要 结合关联限制 K-means 算法能有效地提高聚类结果,但对数据对象分配次序却非常敏感。为获得一个好的分配次序,提出了一种基于分配次序聚类不稳定性的迭代学习算法。根据 Cop-Kmeans 算法的稳定性特点,采用迭代思想,逐步确定数据对象的稳定性,进而确定分配次序。实验结果表明,基于分配次序聚类不稳定性迭代学习算法有效地提高了 Cop-Kmeans 算法的准确率。

关键词 聚类分析,半监督聚类,K-means,关联限制
中图法分类号 TP181 **文献标识码** A

Iterative Learning Assignment Order for Constrained K-means Algorithm

QIU Ye HE Zhen-feng
(School of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

Abstract Constrained K-means algorithm often improves clustering accuracy, but sensitive to the assignment order of instances. A clustering uncertainty based assignment order Iterative Learning Algorithm(UAILA) was proposed to gain a good assignment order. The instances stability was gradually confirmed by iterative thought according to the characteristics of Cop-Kmeans algorithm stability, and then assignment order was confirmed. The experiment demonstrates that the algorithm effectively improves the accuracy of Cop-Kmeans algorithm.

Keywords Clustering analysis, Semi-supervise clustering, K-means, Instance-level constraints

1 引言

传统地,聚类被视为一种无监督的数据分析方法,聚类算法主要用于数据对象的分组,使得同组数据对象相似度高,不同组的数据对象相似度低[1,2]。在不同的科学领域内,聚类都有着悠久而丰富的历史,而在1955年提出的K-means则是最受欢迎、最简单的聚类算法之一。尽管K-means的提出已超过50年,且有成千上万的聚类算法已被提出,但K-means仍然被广泛应用[3]。随着研究的深入,研究者们意识到聚类本质上是主观的,对于同样一个数据集,不同的主观要求得到的结果不同。例如,对于鲸鱼、大象、金枪鱼等,如果按照是否为哺乳类动物进行聚类,则鲸鱼和大象应该聚为一类;而根据是否在水中生活为标准,则鲸鱼和金枪鱼应聚为一类[4]。在现实应用领域里,实验者发现把这样一些主观的用户倾向应用到数据聚类中,可以有效地提高聚类效果。因此,如何把这些用户倾向结合到聚类分析中成为了一个非常重要的问题,而半监督聚类算法就是针对此问题提出的[5]。半监督聚类算法在无监督聚类算法的基础上,结合了一些关于数据对象或聚类目标的有限的背景知识。用户倾向实际上就是背景知识,Wagstaff研究了一类特殊的背景知识:数据对象间的关联限制[6]。这些背景知识通常表示为两类限制,即两个数据对象必须在同一类或者不在同一类。Wagstaff最早提出的一种结合限制的聚类算法是结合限制的K-means算法,即Cop-Kmeans(下面简称CKM)算法。CKM算法的聚类效果普遍好于传统的K-means算法,但CKM算法对聚类过程中的数据分配次序是敏感的,一个好的分配次序可以更有效地提高CKM聚类效果[2]。为此,Yi Hong等人针对CKM算法,提出了一种新的学习数据对象分配次序,称之为基于分配次序聚类不稳定性学习算法(Clustering Uncertainty Based Assignment Order Learning Algorithm),即UALA[2]。UALA利用传统的K-means算法对数据集进行多次聚类,根据聚类结果一次性确定数据对象的稳定性,并按其稳定性对数据对象分配次序进行排序。但是UALA没有考虑到CKM算法中数据对象的稳定性会随着限制的加入而改变这一问题,因此,本文在UALA+CKM的基础上,提出一种新的聚类分配次序学习算法,即基于分配次序聚类不稳定性迭代学习算法。该算法采用迭代思想,利用CKM算法进行多次聚类,并逐步确定其稳定性,进而确定分配次序,从而更有效地提高了CKM算法的准确率。

2 Cop-Kmeans 算法

2.1 K-means 算法

K-means是一种典型的基于划分且常用于自动地把一个规模为 N 的数据集划分为 K 类的无监督聚类算法。算法描

到稿日期:2011-11-23 返修日期:2012-02-15 本文受国家自然科学基金项目(60805042),福建省教育厅科技项目(JA11015)资助。
邱 烨(1988—),女,硕士生,主要研究方向为数据挖掘、机器学习,E-mail:jxgcqe@126.com;何振峰(1971—),男,博士,副教授,主要研究方向为机器学习及其农业应用。

述如下：

输入：数据集 DS=〈X_1，X_2，…，X_N〉，类个数 K；

输出：K 个类中心点；

Step1　从训练集中随机产生 K 个中心点；

Step2　采用欧式距离计算每个数据对象与 K 个类中心的距离，把每个数据对象分配到与其距离最近的类中；

Step3　根据类中数据对象的值，计算它们的平均值从而进行类中心点的更新（当某类为空时，它的中心点从训练集中随机地抽取，而其他的中心点不变）；

Step4　迭代 Step2、Step3，直到类中心点不再发生变化，输出 K 个类中心，算法结束。

2.2　限制描述

虽然目前存在很多限制描述，但应用最广泛的还是 Wagstaff 提出的数据对象间的关联限制。数据对象间的关联限制用于表示关于一对数据对象是否为一类的背景知识。一般表示为两类限制：正关联限制表示两个数据对象必须在同一类；负关联限制表示两个数据对象不能在同一类[7]。

两类限制都具有一定的传递性[6]。一般来说，在结合限制的聚类算法施加限制之前，均依据以上性质来扩展限制集，这样，算法在执行时所增加的限制数会多于最初给定的限制数。因此，结合限制的聚类算法在输入部分需要多输入一个限制集。

2.3　COP-Kmeans 算法（CKM）

CKM 不仅通过距离进行数据对象的聚类，而且在此基础上加入了限制条件检查，即在聚类过程中必须满足给定的限制条件。在 K-means 算法的 Step2 中，加入相应的关联限制，在保证数据对象满足关联限制的情况下，分配数据对象到与其距离最近的类。限制集是根据所规定的限制个数，用随机函数随机产生的。关于限制集的描述如下：

$$Con(i,j)=\begin{cases}1, & x_i \text{ and } x_j \text{ are } Must-link\\ -1, & x_i \text{ and } x_j \text{ are } Connot-link\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

式中，$Con(i,j)$表示数据对象 x_i 和 x_j 的关联限制关系，该算法在每次分配数据对象时，加入相应的关联限制。所有数据对象被分配到最近的适合类，必须满足任何数据对象间的关联限制条件。如果某一个数据对象无法加入到任何一个类中，那么就需重新初始化各个类，并重新划分所有数据对象。

3　基于分配次序的学习算法

3.1　基于分配次序聚类不稳定性学习算法

虽然 CKM 算法比传统的 K-means 算法效果要好，但是 CKM 在数据聚类时，具有数据对象的分配次序敏感性。CKM 算法在数据对象划分的过程中，稳定性较低的数据对象的错误聚类率很高，一旦它们错误聚类，那么跟它们限制相关的数据对象也会被错误聚类。所以如果先让那些稳定性高的聚类，那么错误聚类率就相对较低，与之限制相关的数据对象正确聚类率就高，这样聚类效果肯定会更好。基于这种思想，Yi Hong 和 Sam Kwong 提出了 UALA，并将其运用到 CKM 中。

UALA 的主要思想是首先用 K-means 算法对同一个数据集进行 M 次聚类，计算所有聚类结果的平均值，再根据所得到的平均值计算出每个数据对象的不稳定性，并按升序排序。如果多次聚类对于同一个数据对象的结果是一致的，那么就说这个数据对象不稳定性低，反之其不稳定性高。

3.1.1　计算 M 次聚类结果

为了描述不同样本在聚类过程中的一致性，把每次的聚类结果转换到 $N*N$ 的矩阵 S 中。每次聚类，对于一对数据对象 x_i 和 x_j，结果聚为一类的 $S_{ij}=S_{ji}=1$，反之 $S_{ij}=S_{ji}=0$。一共进行 M 次聚类，最后对 M 次聚类结果求平均值，得到最后的稳定矩阵 S。

S_{ij} 是指 S 矩阵的第 i 行和第 j 列，反映的是数据对象 x_i 和 x_j 被划分到同一类的不确定性值。根据 S_{ij} 计算可知，$0\leqslant S_{ij}\leqslant 1$，如果 S_{ij} 的值趋于 0 或 1，说明数据对象 x_i 和 x_j 不被划分在同一类或者被划分在同一类的可能性非常高。如果 S_{ij} 的值趋于 0.5，那么数据对象 x_i 和 x_j 的聚类结果就有很高的随机性。

3.1.2　计算不稳定性 U 值

每个数据对象 x_i 的不稳定性值 $U(x_i)$ 是指 x_i 聚类的不确定性的信息熵，表示为：

$$U(x_i)=\frac{-\sum_{i=1,j\neq i}^{N}[S_{ij}\cdot\log(S_{ij})+(1-S_{ij})\cdot\log(1-S_{ij})]}{N-1} \tag{2}$$

式中，U 值代表数据对象的不稳定性，U 值越低，说明不稳定性越低；反之，不稳定性越高。然后将所有数据对象{x_1，x_2，…，x_N}的不稳定性值按升序排序：

$$U(x_1)\leqslant U(x_2)\leqslant\cdots\leqslant U(x_N) \tag{3}$$

最后，按上述排好的顺序，用 CKM 算法对其进行聚类。

3.2　UALA 存在的问题

该算法虽然有效地降低了数据对象聚类顺序的敏感性，但是仍然存在一些问题。UALA 采用无监督思想对数据对象稳定性进行一次性确定，可是数据对象稳定性会随着限制的加入而随之发生改变。

在本文的算法提出之前，先通过几个图来描述关于数据对象稳定性会随着限制的加入而随之改变的现象。图 1 表示的是一个两类的数据集。图 2 和图 3 则是基于不同聚类算法对数据对象稳定性的分析。

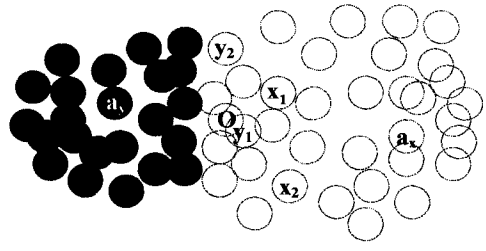


图 1　两类的数据集

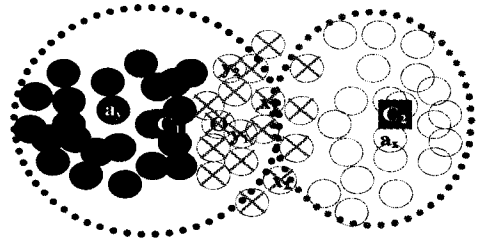
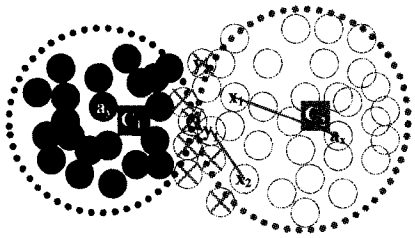
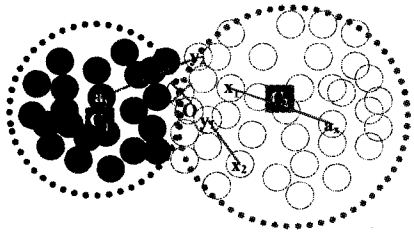


图 2　基于 K-means 稳定性分析



(a)



(b)

图 3　基于 CKM 稳定性分析

从图 2 可以看出，如果确定稳定性的聚类过程中采用 K-means，a_x 和 a_y 这种中心周围的点都具有高稳定性，而图中用带有×表示的数据对象都是很不稳定的点。假设不稳定点 O 稳定性高于不稳定点 y_1。从图 2 可以看出，O 会错误聚到 C_1 中，如果在聚类过程中，加入了 O 和 y_1 的正关联限制，则 y_1 也会随之错误聚到 C_1 中，而且它们周围这些不稳定点也会很大概率地错误聚类，很显然会使类的范围远离我们所预期的。

图 3 采用 CKM 算法确定数据对象的稳定性。在确定稳定性聚类过程中结合了关联限制，使得较稳定的点会增加。如图 3(a)所示，加入了 2 个限制，分别是 x_1 和 a_x 的正关联限制，x_2 和 y_1 的正关联限制。从图中可看出，a_x 是具有高稳定性的点，在正关联的作用下，x_1 的稳定性明显提高，可以稳定地聚类到 C_2 中，与其类似的不稳定点的基于关联限制的加入，使类的范围更加符合预期，并使一些不稳定点得以进入稳定点行列，如 x_2。在正关联限制的指导下，随着 x_2 的加入，又可能使得 y_1 的稳定性大大提高，且高于 O，但 O 也会随着与的正关联关系而提高稳定性，这样会使得不稳定点越来越少。

图 3(b)是基于图 3(a)显示的样本不稳定性进行排序，再一次执行 CKM 得到的稳定性。可以看到，这次多加了 1 对限制：y_2 和 a_y 的负关联限制。我们开始已经确定 a_y 是稳定性很高的点，在负关联限制的作用下，y_2 列入稳定性点的行列，同样，与之类似的不稳定点也会随着关联限制的加入，使其类的范围朝期望方向发展。

从上述分析可以知道，在限制的加入过程中，像 x_1，x_2，y_1，y_2 这些不稳定点会基于限制的加入而变成稳定点，从而改变聚类顺序。从上图可知，按照无监督思想，O 的稳定性高于 y_1，但在限制的加入过程中 y_1 的稳定性已经高于 O，这样它们之间的分配次序应该改变。

因此，不能用无监督思想来一次性确定 CKM 中数据对象的稳定性，而是应该随着限制的加入，迭代地进行数据对象稳定性的确定，这样才能更加确切地反映数据对象的稳定性，从而得到一个好的分配次序。因为只有建立在迭代出的稳定性基础上的聚类顺序才符合实际的需求。

3.3　分配次序的迭代学习算法

针对上述问题，提出了一种基于分配次序聚类不稳定性迭代学习算法。迭代排序是每次只确定 $1/n$ 个样本稳定性，经过 m 次逐步确定数据对象 m/n 个样本稳定性，再一次确定分配次序，从而解决 UALA 一次性确定数据对象分配次序的问题。

UAILA 主要思想：先用 K-means 进行 M 次聚类，得到聚类结果计算不稳定性，并确定稳定性最好的前 $1/n$ 个样本分配次序，后面的 $(n-1)/n$ 个样本分配次序按照最初的数据对象编号；然后用 CKM 进行 M 次聚类，确定稳定性最好的前 $2/n$ 个样本分配次序，此后一直用 CKM 进行聚类来计算稳定性，每次都比前一次多确定 $1/n$ 个样本分配次序，一直重复 m ($m \leqslant n$)次，直到确定所有的分配次序；最后按照确定的分配次序进行 CKM 聚类。整个算法描述如下：

输入：数据集 DS=〈X_1，X_2，…，X_N〉，类个数 K；

输出：K 个类中心点；

Step1　训练集中随机产生 K 个中心和限制集 Con；

Step2　对训练集进行 M 次聚类(用 K-means 算法，训练集分成 p 块，每次只用其中 q 块进行聚类)，计算稳定矩阵 S，且迭代次数 $m_i=1$；

Step3　迭代排序：

1) 根据式(2)，由 S 矩阵计算 U 值；

2) 确定分配次序，稳定性排前 m/n 个样本，按稳定性从高到低确定分配次序，首先分配，剩余样本按原先顺序分配；

3) 对训练集进行 M 次聚类(用 CKM 算法，训练集分成 p 块，每次只用其中 q 块进行聚类，聚类时采用上一步确定的分配次序)，并计算稳定矩阵 S；

4) $m_i=m_i+1$，如果 m_i 达到规定次数 m，则执行 1)、2)，确定最后分配次序，停止迭代。否则转 1)；

Step4　根据上述已经排好的顺序，按升序用 CKM 算法进行聚类。

4　实验结果与分析

4.1　实验数据

本文实验采用 5 个 UCI 数据集，如表 1 所列。

表 1　实验数据

Data Set	Instances	Attributes	Class
IRIS	150	4	3
WINE	178	13	3
Lung Cancer	32	56	3
Soybean(small)	47	35	4
Zoo	101	16	7

4.2　限制产生及结果评价方法

在设置限制时，采用随机发生器，每次随机产生一对数字，这些数字代表数据对象的编号。根据它们的真实划分，如果这对数字之间的关系未曾加入限制集，那么把它们加入限制集，重复以上过程，直至达到所需的限制数。

本文实验结果评价方法用的是 Rand 系数，即评价标准采用正确的样本对数与总样本对数之比，总样本对数 $L=N*(N-1)/2$，N 为总样本数。正确的样本对数分别为 a 和 b，其中，a 表示某样本对本应该属于同一类，结果也确实归为一类；b 表示的是某样本对本不应该属于同一类，结果确实不归为同一类，则准确率为 $\frac{a+b}{L}$。

(下转第 209 页)

daptive local search[J]. IEEE Trans, Evolut. Comput., 2008, 12:107-125

[10] Neri F, Tirronen V. Recent advances in differential evolution: a survey and experimental analysis[J]. Artif Intell Rev, 2010, 33: 61-106

[11] Deb K. Multi-objective optimization[M]. Berlin: Springer, 2005: 273-316

[12] Deb K, Pratap A, Agarwal S, et al. A Fast and Elitist Multi-objective Genetic Algorithm: NSGA-II[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182-197

[13] Storn R, Price K. Differential Evolution-A Simple and Efficient Heuristic for global Optimization over Continuous Spaces[J]. Journal of Global Optimization, 1997, 11(4): 341-359

[14] Brest J, Greiner S. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems[J]. IEEE Transactions on Evolutionary Computation, 2006, 10(6): 646-657

[15] Fang C, Wang L. An effective shuffled frog-leaping algorithm for resource-constrained project scheduling problem[J]. Computers & Operations Research, 2011, 39(5): 890-901

[16] 田菁, 沈林成. 多基地多无人机协同侦察问题研究[J]. 航空学报, 2007, 28(4): 913-921

(上接第 198 页)

4.3 实验参数设置

实验一共采用了两种限制数，分别为 100、200。参数 $M=100, m=2, n=3, p=20, q=18$。且在计算 U 值时，每次对矩阵值加 0.5 进行数据处理，因为计算 U 值的公式中出现了 $\log S_{ij}$ 和 $\log(1-S_{ij})$，而在矩阵 S_{ij} 值为 0 或者是 1 时，log 运算会出错。本文实验采用的是十指交叉验证，实验次数为 100 次，结果为 100 次的平均值。

4.4 算法收敛性分析

本算法是基于结合限制的 K-means 算法的改进，通过聚类稳定性来寻找较优的分配次序，这种分配次序的修改一般不影响其收敛性。而 CKM 等一系列基于划分的限制 K-means 算法已经被文献[6]证明不收敛，本算法亦不收敛，为此，规定了最大迭代次数以保证算法的终止。

4.5 实验结果

采用了 UAILA+CKM 与 UALA+CKM，CKM 及 K-means 进行比较，100 和 200 限制数的实验结果分别如表 2 和表 3 所列。

表 2 算法准确率：限制数 100

Data Set	K-means	CKM	UAUL+CKM	UAILA+CKM
Iris	84.93%	89.78%	89.09%	**90.05%**
Wine	64.83%	71.20%	71.28%	**71.33%**
Lung Cancer	52.80%	58.77%	**61.33%**	61.00%
Soybean(small)	83.97%	92.73%	94.83%	**96.53%**
Zoo	84.68%	88.61%	86.25%	**86.93%**

表 3 算法准确率：限制数 200

Data Set	K-means	CKM	UAUL+CKM	UAILA+CKM
Iris	84.93%	89.95%	91.15%	**91.15%**
Wine	64.83%	71.49%	**71.65%**	71.40%
Lung Cancer	52.80%	58.67%	58.83%	**58.83%**
Soybean(small)	83.97%	93.33%	93.90%	**97.50%**
Zoo	84.68%	92.03%	89.64%	**91.22%**

从表 2 和表 3 可以看出，所有 CKM 算法的效果都比传统 K-means 算法的好，且 UAILA 的效果基本都不低于 UALA。

对于数据集 Iris，限制数为 100 时，UALA 准确率低于 CKM 算法，而 UAILA 的准确率都高于 UALA 算法和 CKM 算法。限制数为 200 时，UAILA 和 UALA 算法准确率都高于 CKM 算法，且 UAILA 也不低于 UALA 算法。

对于数据集 Wine、Lung Cancer 和 Soybean，限制数为 100 和 200，UAILA 和 UALA 准确率都高于 CKM 算法，除了数据集 Wine 在限制数为 200 和数据集 Lung Cancer 在限制数为 100 时，UAILA 比 UALA 低一点，其它都高于 UALA。而且数据集 Soybean 的准确率有大幅度的提高，限制数为 100 时，数据集 Soybean 的准确率提高了 1.7%；限制数为 200 时，准确率提高了 3.6%。

对于数据集 Zoo，限制数为 100 和 200 时，UAILA 的准确率都要高于 UALA，但这两种算法的准确率都低于 CKM 算法。这说明 UAILA 虽然比 UALA 更有效地提高了准确率，但还是有缺陷。

我们可以看到，具有样本少、数据高维性等特点的数据集 Lung Cancer 在限制数为 200 时的准确率都低于限制数为 100 时的准确率。因此对于高维数据的数据集的限制聚类，还有待进一步的研究。

从上面的实验结果可以知道，迭代排序比一次排序的聚类效果好，这也说明 CKM 算法中数据对象稳定性会随着限制的加入而改变。

结束语 虽然 CKM 算法较传统的 K-means 在准确率上得到了一定的提高，但是 CKM 算法具有分配次序的敏感性，而 UALA 对于学习一个好的分配次序起了一定作用。分配次序随着聚类过程中限制的逐渐加入而随之变化，但 UALA 采用无监督思想进行一次性确定来分配次序。针对 UALA 存在的问题，本文在 UALA+CKM 的基础上，采用迭代思想，利用 CKM 算法逐步确定稳定性，提出了一种基于分配次序聚类不稳定性迭代学习算法。上述实验结果表明，该算法有效地改善了聚类效果。

参 考 文 献

[1] Wagstaff K, Cardie C, Rogers S, et al. Constrained K-means clustering with background knowledge[C]// Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001). 2001: 577-584

[2] Hong Yi, Kwong S. Learning Assignment Order of Instances for Constrained K-means Clustering Algorithm[J]. IEEE Systems, Man and Cybernetics Society, 2009, 39(2): 568-574

[3] Jain A. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31: 651-666

[4] Jain A K, Murty M N, Flynn P J. Data Clustering: A Review [J]. ACM Computing Surveys, 1999, 31(3): 265-323

[5] 肖宇, 于剑. 基于近邻传播算法的半监督聚类[J]. 软件学报, 2008, 19(11): 2803-2813

[6] 何振峰, 熊范纶. 结合限制的分隔模型及 K-Means 算法[J]. 软件学报, 2005, 16(5): 799-809

[7] Wagstaff K, Cardie C. Clustering with Instance-level Constraints [C]// Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000). 2000: 1103-1110