DOI:10.16644/j.cnki.cn33-1094/tp.2017.09.001

离差最大化赋权的蚁群聚类算法

张新建

(中国人民银行合肥中心支行，安徽 合肥 230091)

摘　要：受蚂蚁觅食过程启发的聚类算法又被称为蚁群聚类算法，把觅食行为分为搜索食物和搬运食物两个环节，把数据对象视为蚂蚁，把聚类中心视为“食物源”，这样数据对象的聚类过程就可以转化为蚂蚁觅食过程，但在该算法中没有区分数据对象不同属性的重要性，通过采用离差最大化方法，根据每个属性的重要性赋予它一个权值，从而改进了原算法中的距离计算，使得相似的数据对象能快速的聚集到一起，提高了算法的运行效率。

关键词：聚类算法；蚁群算法；离差；权值

中图分类号：TP301.6　　**文献标志码：**A　　**文章编号：**1006-8228(2017)09-01-04

Ant colony clustering algorithm with maximizing weight of deviation

Zhang Xinjian

(*The People's Bank Of China Hefei Central Sub-branch, Heifei, Anhui 230091, China*)

Abstract: Clustering algorithm inspired by the foraging process called ant colony clustering algorithm, the foraging behavior is divided into two aspects, food searching and food handling, the data object as an ant, the cluster center as a "food source", so the clustering process of data objects can be converted to the ant foraging process, but did not distinguish between the importance of the different attributes of the data objects in the algorithm, this paper uses the maximum deviation method for each attribute according to its importance as it gives a weight, which improves the original algorithm in the distance calculation, makes similar objects fast together, and improves the efficiency of the algorithm.

Key words: clustering algorithm; ant colony algorithm; deviation; weigh

0 引言

通过对自然界中蚂蚁寻找食物过程的观察，学者们发现实际上整个寻找食物的过程可以简单地分为两个环节：搜索食物和搬运食物。蚂蚁在寻找食物时不论是在搜索食物环节还是在搬运食物环节，都会在它所经过的路径上释放一定量的信息素，这种信息素的强度可以被其他蚂蚁所感知到，同时信息素本身也具有一定的挥发性，即它的强度会随着时间的推移而慢慢减弱以至消失。自然界中蚂蚁不仅可以感知到信息素的强弱，也具有追逐信息素的倾向，即如果某条路径上信息素的强度很高，那么蚂蚁在选择路径时，选择这条路径的概率就很大。信息素对蚂蚁选择路径行为的影响通过蚂蚁群体行为的放大就可以表现出一种正反馈现象，即如果某条路径上信息素强度高于其他路径，那么蚂蚁就会以较高的概率选择此路径，同时鉴于蚂蚁在运动时也会在路径上释放一定量的信息素，因此该路径上的信息素强度会逐渐增强，而随着信息素强度的增强，它又会对其他蚂蚁散发出更大的吸引力，会吸引更多的蚂蚁通过此路径；而其他路径则因为只有较少蚂蚁通过，信息素强度得不到增强，同时又因为空气挥发作用使得信息素强度逐渐降低，使得该路径对蚂蚁的吸引力愈加低下，经过一段时间之后蚂蚁甚至会“忘记”该路径的存在。蚂蚁的这种通过信息素在彼此之间进行信息交流的群体行为可以应用在聚类算法之中。下面对基于蚂蚁觅食原理的聚类算法的基本思想[1]进行简单的介绍。

如果将待聚类的数据对象看成是蚂蚁，而算法所要寻求的聚类中心看成是蚂蚁所要寻找的“食物源”，

收稿日期：2017-07-04

作者简介：张新建(1985-)，男，安徽合肥人，硕士，工程师，主要研究方向：智能算法。

那么就可以把数据聚类过程转化为蚂蚁寻找食物源的过程。假设待聚类的数据对象为:X={X|X_i(x_{i1},x_{i2},…,x_{im}), i=1,2,…,N},对算法进行初始化,将各条路径上的信息素初始化为0,即$\tau_{ij}(0)=0$,设置聚类簇的半径r、统计误差ε、概率阈值P_0,以及α、β等参数。计算对象X_i到X_j的欧式距离d_{ij},有:

$$d_{ij}=\left(\sum_{k=1}^{m}|x_{ik}-x_{jk}|^2\right)^{\frac{1}{2}} \quad k\in\{1,2,...,m\} \tag{1}$$

在算法运行过程计算各路径上的信息素$\tau_{ij}(t)$:

$$\tau_{ij}(t)=\begin{cases}1, d_{ij}\le r\\ 0, d_{ij}>r\end{cases} \tag{2}$$

在算法运行过程计算数据对象X_i和数据对象X_j属于同一类簇的概率为:

$$P_{ij}(t)=\frac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)}{\sum_{s\in S}\tau_{sj}^{\alpha}(t)\eta_{sj}^{\beta}(t)} \tag{3}$$

其中,$S=\{X_s\,|\,d_{sj}\le r, s=1,2,...,j,j+1,...N\}$,如果$P_{ij}(t)$大于阈值$P_0$,表明$X_i$和$X_j$是属于同一类簇,那么就将$X_i$合并到$X_j$的领域内。这里$\eta_{ij}$是$d_{ij}$的倒数,它表示的是数据对象之间的欧氏距离对聚类概率的影响。

1 离差最大化赋权算法

1.1 多属性决策

多属性决策是多目标方案决策的一种,又称有限方案多目标决策,它是对具有多个属性的有限方案,按照某种决定准则进行多方案选择和排序。其理论方法已被广泛地应用于社会、经济、管理、军事等领域,其求解方法和属性权重有密切的关系。因为它的合理性直接影响着多属性决策排序的准确性,所以在多属性决策中,权重问题的研究占有重要的地位。

1.2 离差最大化赋权算法

离差最大化赋权法是王应明1998年在文献[2]中提出的,到目前为止,在多属性决策模型中它的应用已经比较广泛了[3]。它是从对各方案排序的角度出发,认为若各个方案的某个属性值没有差别,则该属性对于方案排序将不起作用,在多属性决定中该属性的意义就不大。所以,属性对于各个方案而言差异越大,则该属性在方案排序过程中的区分度越大,属性越重要,应该赋予该属性较大的权重。

文献[4]给出了离差最大化赋权法的计算过程。首先对样本集的全体X作如下表示,即$X=(x_{jt})_{n\times m}$,其中$x_{jt}(1\le j\le n, 1\le t\le m)$是第j个样本的第t个特征的赋值。

设特征的权向量为$\omega=(\omega_1,w_2,....,w_m)$,$w_t\ge 0, 0<i\le n$并满足$\sum_{i=1}^{m}w_i^2=1$。

通常来说,需要进行聚类处理的数据对象都包含两个或者多个属性,数据对象正是由对这些属性进行取值形成的,这些属性反映了数据对象在某些方面的特征,而属性的取值则是数据对象的本身特征的量化表示。因此对数据对象进行聚类处理也就是对数据对象的属性进行处理,也就是说聚类处理的结果是由数据对象的属性所决定的。数据对象具有多个属性,每个属性反映的是某方面的特征的信息,就属性本身而言所有的属性都是平等的,没有主次之分,它们都是数据对象本身信息的客观反映。然而每个属性的取值范围又是不同的,也就是说不同数据对象在同一个属性上的取值,差异性大小是不同的,差异越小,表明数据对象之间在该属性下的相异度较小,差异越大,则表明数据对象之间在该属性下相异度较大,因此影响样本X_j属于某一类簇的概率主要取决于每个样本在同一属性下赋值上的差异程度。

样本x_j在特征t下与其他样本之间的离差用用$H_{jt}(w)$表示,而$H_i(w)$表示在特征t下所有样本之间的总离差,H(w)表示在所有特征下所有样本之间的总离差,则有如下定义:

$$H_{jt}(w)=\sum_{k=1}^{n}\left|x_{jt}w_t-x_{kt}w_t\right|, 1\le j\le n, 1\le t\le m \tag{4}$$

根据上述定义,易得

$$H_t(w)=\sum_{j=1}^{n}\sum_{k=1}^{n}\left|x_{jt}w_t-x_{kt}w_t\right|, 1\le t\le m \tag{5}$$

$$H(w)=\sum_{t=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n}\left|x_{jt}w_t-x_{kt}w_t\right| \tag{6}$$

由上分析,加权向量ω的选择,应使所有特征下所有样本之间的总离差最大。于是求解权向量ω等价于求解如下的最优化模型:

$$\begin{cases}\max H(w)=\sum_{t=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n}\left|x_{jt}w_t-x_{kt}w_t\right|\\ \sum_{t=1}^{m}w_t^2=1\\ 0\le w_t\le 1, 1\le t\le m\end{cases} \tag{7}$$

通过作Lagrange函数解此最优化模型:

$$L(w,\lambda)=\sum_{t=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n}\left|x_{jt}w_t-x_{kt}w_t\right|+\frac{1}{2}\lambda\left(\sum_{t=1}^{m}w_t^2-1\right) \quad (8)$$

对它求偏导数，并令

$$\begin{cases}\dfrac{\partial L}{\partial \omega_t}=\sum\limits_{j=1}^{n}\sum\limits_{k=1}^{n}\left|x_{jt}-x_{kt}\right|+\lambda w_t=0,1\le t\le m\\ \dfrac{\partial L}{\partial \lambda}=\sum\limits_{t=1}^{m}\omega_t^2-1=0\end{cases} \quad (9)$$

求得最优解并进行归一化处理后得出的结果为：

$$\omega_t=\frac{\sum\limits_{j=1}^{n}\sum\limits_{k=1}^{n}\left|x_{jt}-x_{kt}\right|}{\sum\limits_{t=1}^{m}\sum\limits_{j=1}^{n}\sum\limits_{k=1}^{n}\left|x_{jt}-x_{kt}\right|},1\le t\le m \quad (10)$$

由(10)式可知，数据对象的每个属性的权重是在这个属性下样本之间的离差与所有属性下样本之间的总离差的比值。因此如果在某个属性下样本之间的离差越大，表明这类数据对象在这个属性上的差异性很大，则该属性对聚类结果的影响就越大，即它的权重就大，反之则小。由(10)式给出的权重的计算公式，容易计算，所得到的权重也能客观真实的反映每个样本属性在聚类中贡献。

2 基于离差最大化赋权的蚁群聚类算法

2.1 属性权重对算法聚类结果的影响

2.1.1 对特征属性进行赋权的必要性分析

在聚类算法中经常被使用的数据对象间的距离表示的是数据对象之间的相近程度，而事实上，相似不仅依赖于对象间的相近程度，还依赖于对象内在的性质，而对象内在的性质是通过它的属性表示出来的，因此对象中每个属性变量的重要性是不同的，因此在多属性数据对象之间的距离计算中，不同的属性很显然对数据对象的内在性质有不同的贡献，有的属性很重要，而有的属性则并不重要，甚至可有可无，它表明数据的各个不同的特征属性对数据性质的影响程度即对聚类结果的贡献程度是不同，因此这需要算法在计算的时候体现出来，即在可以通过对不同的属性变量赋予不同权重的方式来解决，即对每个变量根据其重要程度赋一个权重，因此公式(1)在增加了权重之后，可以更改为(11)

$$d_{ij}=\left(\sum_{k=1}^{m}\omega_k\left|x_{ik}-x_{jk}\right|^2\right)^{\frac{1}{2}} \quad (11)$$

$\omega_k(k=1,2,...,m)$表示每个属性变量的权重。

在算法对数据对象进行聚类分析时，数据对象属性个数的增加会使算法的计算量急剧膨胀，从而降低算法运行的效率。因此在进行聚类时合理地运用加权欧氏距离，根据每个属性对聚类结果贡献的不同，给每个属性赋一个权值，这样既可以充分利用数据的分布特征，从而加快某些聚类算法的速度，同时又可以更准确的反映数据对象的内在性质，进而提高聚类结果的准确性，对改进聚类结果能起到较好的效果。

2.1.2 权重的设置方法

较常使用的加权方法有以下几种：德尔菲（Delphi）法、层次分析（Analytic Hierarchy Process, AHP）法以及模糊聚类分析法。德尔菲法和AHP法都是基于专家群体的知识、经验和价值判断。尽管AHP法中对专家的主观判断做了数学处理，但专家知识的局限性并未消除，这两种方法都存在一定的主观性。模糊聚类分析法是基于样本模糊数据的相似性对评价指标群体做出相对重要程度分类，但该方法不能确定单个属性的权重。

数据对象的属性对于聚类任务非常重要。数据集用可分性越好的属性子集来描述，具有相同类别的数据对象越集中，而不同类别的数据对象之间则相距越远。表现在数据分布图上就是同类的数据对分布较为集中，而类与类之间的距离则比较大。

在多属性数据对象的距离计算中，不同的属性很显然对数据对象的性质有不同的影响。在本文第1章中介绍的离差最大化赋权算法，可以根据数据对象各属性重要性的不同，计算出不同的权值，从而能够客观的反映数据对象的情况，这正好满足了聚类运算的目的，即客观地反映出数据集中所隐藏的信息。

2.2 改进后的算法

基于觅食的蚁群聚类算法利用了蚁群的分布式搜索的特性，因此相比于经典的k–means算法，它改善了算法过早的陷入局部最优的缺陷，但是在蚁群聚类算法进行计算的时候，并没有对各个特征属性的重要性加以区分，因而不能有差异的反映各个属性对聚类结果的不同影响。

本文将离差最大化赋权算法应用于基于蚂蚁觅食原理的聚类算法中对数据对象的属性的权值的计算中，从而给不同的属性赋予不同的权重，突出重要属性的影响，同时弱化非重要属性的影响，从而更快

更好的获得聚类结果。

2.2.1 改进后的算法流程

Step 1 初始化聚类中心，设定参数N，m，r，ε_0，α，β，ρ_0

Step 2 求出上文介绍的加权向量ω_k(k=1,2,…,m)

Step 3 计算样本X_i到X_j之间的加权欧氏距离

$$d_{ij}=\sqrt{\sum_{k=1}^{m}\omega_k\left|x_{ik}-x_{jk}\right|^2}$$

Step 4 计算各路径上的信息量：$\tau_{ij}(t)=\begin{cases}1, d_{ij}\le r\\ 0, d_{ij}>r\end{cases}$

Step 5 对象X_i合并到X_j的概率为：

$$P_{ij}(t)=\frac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)}{\sum_{s\in S}\tau_{sj}^{\alpha}(t)\eta_{sj}^{\beta}(t)}$$

Step 6 判断$\rho_{ij}\ge\rho_0$是否成立，若成立则将X_i合并到X_j的邻域。

Step 7 计算归并X_j领域的数据集合的聚类中心$\bar{c}_j=\frac{1}{J}\sum_{k=1}^{J}X_k$。

Step 8 计算第j个聚类的偏离误差：

$$D_j=\sum_{k=1}^{J}\sqrt{\sum_{i=1}^{m}(x_{ki}-c_{ji})^2}$$

其中c_{ji}表示第j个聚类中心的第i个分量。

Step 9 计算总体误差$\varepsilon=\sum_{j=1}^{k}D_j$

Step 10 判断$\varepsilon\le\varepsilon_0$若成立，则停止，并输出聚类个数；否则，转步骤Step 3继续迭代。

2.2.2 仿真实验及分析

为了验证改进后的算法的有效性，将使用UCI机器学习数据集中的Iris(150,4)和Wine(178,13)数据集来进行仿真实验，并对和原算法的实验结果进行对比分析。实验中设置的参数如下：ant=5,r(Iris)=1.5,r(Wine)=10,p_0=0.000005，鉴于参数ε_0的设定有太大的主观性，根据离差最大化赋权法计算样本Iris的4个属性的权值分别为(0.1967,0.4507,0.5785,0.1798)。样本Wine中的13个属性权值为(0.1415,0.1063,0.1130,0.09014,0.1346,0.0814,0.0187,0.0457,0.0603,0.0559,0.0546,0.0809,0.5788)。结束条件设定为算法循环NC=200次。表1的数据是算法运行50次，取每次运行中的最佳聚类结果，取平均值得出。

表1 实验结果对比

数据集	Iris		Wine	
算法	原算法	改进算法	原算法	改进算法
平均错误	11.2	8.4	44	32
错误率(%)	7.46	5.6	4.7	17.9

通过表1可以看到改进后的蚁群聚类算法相比较于原算法在聚类的准确度上有了一定的改进。这主要是因为改进后的算法根据数据的各个特征属性的重要程度而赋予不同的权值，对于聚类贡献较大的特征属性赋予较大的权值，而对于聚类贡献相对较小的特征属性则赋予较小的权值，进而突出了重要属性的作用，弱化了非重要属性对聚类结果的干扰，实验证明了，改进后的算法取得了较好的效果。

3 结束语

本文研究了蚁群算法在数据挖掘聚类方法中的一个应用，改进了基于蚂蚁觅食原理的聚类算法中的距离计算，采用离差最大化赋权算法给数据对象的属性赋予一定的权值，从而使得数据对象属性的重要程度得到了区分，利于相似的数据对象能快速的聚集到一起，并且弱化了非重要属性对聚类结果的干扰，减少了无效的相似度计算，提高了聚类的准确率，但是基于觅食的蚁群聚类算法受初始聚类中心的影响较大，而初始聚类中心的选取，在目前为止并没有一个较为完善的方法，并且算法在运行过程中需要设置的重要参数较多，如聚类半径r，统计误差ε，蚂蚁数量m等，都需要根据实际情况及经验作出确定，带有一定的主观性，因此，如何找到一个科学的参数设定方法将是今后研究工作的重点。

参考文献(References):

[1] 高新波，谢维信.模糊聚类理论发展及应用的研究发展[J].科学通报，1999.44(21).

[2] 王应明.运用离差最大化方法进行多指标决策与排序[J].系统工程与电子技术，1998.20(7):24-26

[3] 王坚强.基于离差优化的信息不完全确定的多准则分类方法[J].控制与决策，2006.21(5):513-516

[4] 李正义，曾雪兰，覃菊莹.离差最大化特征加权模糊C-划分的聚类分析[J].模糊系统与数学，2008.22(4):171-172