文章编号:1671-4598(2018)06-0149-03 DOI:10.16526/j.cnki.11—4762/tp.2018.06.038 中图分类号:TP301.6 文献标识码:A

基于相似性算法与蚁群算法的聚类算法

朱俚治
(南京航空航天大学 信息中心,南京 210016)

摘要:由于当今的网络数据是海量的,因此科研人员对某些问题进行研究时需要将不同属性的数据从中提取出来,然而在提取这些数据之前需要将相同数据进行聚类;数据聚类的过程,也就是寻找数据最优属性的过程,然而人工蚁群就是一种寻找问题最优解的算法,因此在本文中再次将蚁群算法在聚类中进行应用;提出的聚类算法可以分为两个部分,第一部分是:通过相似性算法来衡量数据之间的相似度,第二部分是:根据第一部分的计算结果,再采用蚁群算法为需要聚类的数据选择不同的聚类中心,从而对不同属性的数据进行聚类,经过以上两个过程的计算,可以实现对数据的聚类;在文中进行数据聚类时采用的相似性度量来代替距离的计算,是本文创新点之一,采用蚁群算法在聚类过程中来选择聚类中心也是本文的创新所在。

关键词:蚁群算法;聚类;相似性

Clustering Algorithm Based on Similarity Algorithm and Ant Colony Algorithm

Zhu Lizhi
(Information Center, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: as the network data today is massive, so the researchers on the study of some problems need to be different from the properties of the data extraction, however, these data will be required clustering before extracting the same data . The process of data clustering is the process of finding the optimal attributes of data. However, the artificial ant colony is an algorithm for finding the optimal solution of the problem. Therefore, the ant colony algorithm is applied to the clustering in this paper. The clustering algorithm proposed by this paper can be divided into two parts, the first part is: through the similarity algorithm to measure the similarity between the data, the second part is: according to the results of the first part, using ant colony algorithm to select different clustering in the heart for clustering data, and clustering of the different attributes of the data. Through the calculation of the above two processes, can realize the clustering of data. In this paper, the similarity measure used in data clustering is used to replace the distance calculation. It is one of the innovations of this paper. It is also the innovation of this paper that the ant colony algorithm is used to select the cluster center in the clustering process.

Keywords: ant colony algorithm; clustering; similarity

0 引言

当今已是大数据时代,随着信息技术的发展,人们积累的音频、视频、文本图片等越来越多,在某些时候需要将这些数据从海量的数据中提取出来,从而将这些数据提供给相关研究人员进行研究和分析。为了将这些不同属性的数据进行正确分类,必须采用聚类算法或分类算法,聚类和分类是机器学习中两种重要的理论和技术,原理上来看聚类技术和分类技术有明显的区别,聚类分析是一种无监督的学习,而分类技术一种有监督的学习[1-9],聚类分析的智能性要强于其它算法,基于上述理由可以知道聚类分析在大数据时代有着广泛的应用[10-13]。

聚类分析是一种寻找最优解的算法,而粒子群算法,人工蜂群算法以及人工蚁群算法都是一种寻找问题最优解的算法,然而这些仿生物学的智能算法在聚类过程中都有所应用。人工蚁群算法最早是由意大利学者 MDorigo 提出,该算法在工程领域中主要有如下应用:组合优化问题,网络优化,机器人优化等一系列方面[1-9]。人工蚁群在聚类技术中也有相当的应用:人工蚁群的觅食过程就是一个寻找问题最优解的过程,因此基于蚁群觅食行为的算法在聚类算法中是最早的应用[1-9]。在 2000 年 Monmarche 学者提出了一种混合型的蚁群聚类算法[3-9],该算法在聚类时容易产生早熟现象,在收敛之前过早的就终止了该算法,这样将造成算法的聚类效果较差等不足之处。人工蚁群在搜索食物源之时,人工蚂蚁都能挥发一种叫做信息素的物质,因此 Labroche 等相关科研人员根据人工蚁群的这一特征,在 2002 年提出了一种基于人工蚁群信息素浓度的聚类算法[2-9],在该算法中人工蚁群,在通过不同的标签来寻找巢穴这一过程,从而实现数据聚类的目的,此算法具有较强的鲁棒性性和较好的适应性,但在算法收敛性这一方面存在一定问题,值得广大科研人员研究。以上是蚁群算法在聚类中有不同的应用,这些算法存在一定的优点,同时也存在某些的不足之处,因此本文在参考了相关的蚁群算法和聚类算法后,提出了一种新的聚类算法,该算法有两部分组成即相似性计算与蚁群算法对聚类中心的选择。

1 什么是聚类分析

聚类分析是一种智能技术,在数据挖掘中是一个重要分支。聚类技术能够对未知数据进行有效的分类,强化机器的智能性,因此各种聚类算法在各个科学领域都有所应用[10-13]。

聚类算法的思想是使用一定的规则将最为相似的数据聚成一个族,然而要求不同族之间的差异性达到最大值,这就是数据的聚类[10-13]。聚类分析进行数据聚类时,在没有训练的条件下能够把对象划分为若干类,这样最相似的数据就聚成了族,因此聚类分析进行数据聚类时是一种无监督的学习。聚类分析中,同一个族中的数据具有较高的相似度,而不同的族数

收稿日期:2017-09-18; 修回日期:2017-10-18。

作者简介:朱俚治(1980-),男,江苏宜兴人,大学,工程师,主要从事计算机技术与信息安全方向的研究。

据之间的差异达到最大，这就是聚类的最大特点，在聚类的过程中相似度是聚类经常采用的度量方法。

2 蚁群算法的介绍

2.1 蚁群算法的特点

蚁群算法最初之目的是帮助人们去理解蚂蚁这类物种的复杂行为。蚁群算法出现后不久，便引起了数学家、计算机专家和工程师们的注意，他们把超越生物本身的模型，转换成为一项有用的优化和控制算法，该算法这就是蚁群算法[6-13]。

蚁群在寻找到达食物源的最短路径时，是群体的行为，这些蚁群在路径中都会留下一种叫做信息素的物质，信息素具有挥发性。蚁群之间通过信息素这一物质进行相互交流，从而实现蚁群之间的相互协作和竞争。在路径上的信息素浓度有强有弱，然而蚁群总是向着信息素浓度最高的方向前进。如果路径中的蚁群数量越多，则信息素的浓度就越高，信息素浓度越高，则吸引更多的蚂蚁来到这条路劲中，因此在信息素的协调和作用下，使得整个蚁群都能向着食物源的方向前进。

2.2 蚁群算法的简要分析

蚁群是一个种群，所以蚁群在寻找最短路径过程中是一个群体行为，蚂蚁之间通过相互和协作和竞争来完成共同的目标。当蚁群出发寻找最短路径时，各条路径中的蚂蚁可以在不同的时间点上出发，也可以在同一个时间点上，有若干蚂蚁同时出发。当蚁群搜索算法开始时，寻找最短路径时是蚂蚁的个体行为，但是通过蚁群之间的协调机制和信息素，最终能使得整个蚁群都能找到通往食物源的最短路径。

在蚁群算法中，蚁群之间采用一种分布式的合作机制，蚁群之间的协作性和并行性是其分布合作的具体表现，信息素浓度的大小指引着蚁群寻找通往食物源的最短路径，这是蚁群算法的主要特点。在蚁群算法中人工蚂蚁具有记忆力，因此缩短了蚁群寻找最短路径的时间，从而提高了算法的效率。由于人工蚂蚁具有记忆力，从而蚁群在寻找最短路径时，并不是盲目的搜索，而是有规律、有意识地寻找最优路径，即最短路径。

2.3 蚁群算法的重要参数

β 指的是期望启发因子，它反映的是启发式信息在影响蚁群搜索的过程中的相对重要度。β 值越大，蚁群就越容易选择局部较短路径，这时算法的收敛速度是加快了，但是随机性却不高，容易得到局部的相对最优。

Q 同样是蚁群算法中的一个重要参数，蚁群进行搜索时在路径中能留下不同强度的信息素，不同路径中蚁群挥发的信息素浓度是不同的，如果路径中信息素浓度越大，则 Q 的值就越大，事实可以证明 Q 值大小的不同，能对整个蚁群算法产生不同的正反馈功能。算法在正反馈和负反馈的影响下，能够有效地找到问题的最优解。

3 数据的聚类过程

3.1 聚类数据的相似度计算

$$f(x)=\frac{\text{需要聚类的数据属性值}}{\text{聚类中心的数据属性值}} \tag{1}$$

分析和讨论：

1）如果 $f(x)=\dfrac{\text{需要聚类的数据属性值}}{\text{聚类中心的数据属性值}}$ 的比值接近于 1 时，则表示该数据的属性与聚类中心十分接近，此时该数据的属性值与聚类中心的距离就相对较小，进行聚类的成功概率就越大。根据以上的叙述有以下的分析：

在数据相似性计算过程中，如果 $\dfrac{A'}{A}$ 的比值十分接近于 1 时，那么函数 $f(x)=\left|1-\dfrac{A'}{A}\right|$ 的值就十分接近于 0，则 A' 偏离已知对象 A 的程度就十分小，将趋向于 0。因此如果 $f(x)=\left|1-\dfrac{A'}{A}\right|$ 的值越小，则 A' 偏离 A 的程度就越小，数据之间的距离就越小，数据的相似性就越强。

2）如果 $f(x)=\dfrac{\text{需要聚类的数据属性值}}{\text{聚类中心的数据属性值}}$ 的比值大于 1 时，当比值大于 1 的程度越明显，则表明该数据与聚类中心的距离较大，此时数据之间的相似性较弱，进行聚类时成功的概率就较小。根据以上的叙述有以下的分析：

在数据相似性属性计算过程中，如果 A' 的属性值大于已知对象 A 时，那么 $\dfrac{A'}{A}$ 的比值将大于 1 时。当 $\dfrac{A'}{A}$ 的比值越大时，则函数 $f(x)=\left|1-\dfrac{A'}{A}\right|$ 的值大于 0 的程度将越明显。当对象 A' 的值大于已知对象 A 时，数据的相似性就越差。如果 A' 与 A 之间的差值就越大，数据之间的距离就越大，则此时数据聚类的成功概率就较差。

3）如果 $f(x)=\dfrac{\text{需要聚类的数据属性值}}{\text{聚类中心的数据属性值}}$ 的比值小于 1 时，当比值小于 1 的程度越明显，则表明该数据的属性值与聚类中心的距离较大，数据的属性值与聚类中心数据的相似性较弱，这时数据聚类的成功概率就越小。根据以上的叙述有以下的分析：

在数据相似性属性计算过程中，如果对象 A' 的属性值小于已知对象 A 时，那么 $\dfrac{A'}{A}$ 的比值将小于 1。当 $\dfrac{A'}{A}$ 的比值越小 1 时，则函数 $f(x)=\left|1-\dfrac{A'}{A}\right|$ 的值小于 1 的程度将越明显，当 A' 的值小于已知对象 A 时，数据的相似性就越差。如果 A' 与 A 之间的差值就越大，则此时数据之间的距离就越大，聚类的成功的概率就较差。

在 3.2 经过 3.1 的数据属性相似度计算后，以下用蚁群算法进行数据聚类中心选择。

3.2 聚类中心的选择与聚类的实现

1）函数：

$$h(x)=|1-f(x)| \tag{2}$$

以上函数讨论如下：

当公式（1）：$y_n=f(x)$ 比值越近于 1，则表明数据的相似性更优。数据相似越好，则此时 $h(x)=|1-f(x)|$ 的值就越小，在采用蚁群算法进行聚类时，聚类数据选择该聚类中心的概率也就越大，此时蚁群启发参数 β 期望值的相对重要程度也相对越大。

当公式（1）：$y_n=f(x)$ 比值偏离 1 的程度越明显，则表明数据的相似性更差。数据相似越差，则此时 $h(x)=|1-f(x)|$ 的值就越大，在采用蚁群算法进行聚类时，聚类数据选择该聚类中心的概率也就越小，此时蚁群启发参数 β 期望值的相对重要程度也相对越小。

2）在采用蚁群算法进行数据聚类时，聚类公式：$h(x)=|1-f(x)|$ 有以下的分析和结论：

（1）蚁群算法聚类时相关参数的说明：

符号 d_{ij} 代表着城市 i 转移到城市 j 的距离，每一个城市都是聚类过程中的一个聚类中心。

在选择聚类中心的算法中：设定 $d_{ij} = h(x) = |1 - f(x)|$。

在聚类中心选择算法中，有两个重要的蚁群算法参数：

参数 η_{ij} 表示某条路径的能见度程度，反映由城市 i 转移到城市 j 的启发程度，通常取值为 $1/d_{ij}$。参数 β 是蚁群算法中的一个重要因子，作为蚁群寻找下一条路径的期望启发因子。β 值越大，则人工蚂蚁选择这条路径的概率就也大，并且蚁群越倾向于能见程度高的路径上行走，因此 β 和 η_{ij} 的值越大，则人工蚁群选择该路径的概率就越大，该路径上的人工蚁群数量就也就越多。另外在蚁群算法中还以可以证明减小距离，参数 β 的值也可使该路径被选中的概率增加。

（2）蚁群算法聚类中心的选择：

在本文提出的算法中，一般将数据视为具有不同属性的人工蚂蚁，聚类中心就是蚁群所要寻找的食物源，数据聚类的过程可以被看作是蚂蚁寻找食物源的过程，在这一过程中每个城市可看作某个聚类中心，当蚁群从一个城市到达另一个城市时，这些聚类中心可以按照一定的规律变化。

以下是聚类过程中聚类中心选择的具体分析：

当 $d_{ij} = h(x) = |1 - f(x)|$ 越小，则 d_{ij} 的值也就越小。如果 d_{ij} 的值越小，则 $\eta_{ij} = \frac{1}{d_{ij}}$ 就越大，并且 β 之值就越大。当 d_{ij} 值越小，根据 3.2 节公式（2）可以知道，此时需要聚类的数据偏离聚类中心的程度越小，数据与聚类中心数据的相似程度就越高，因此当 β 值越大，则 η_{ij} 的值就越大，由于 η_{ij} 反映由城市 i 转移到城市 j 的启发程度，因此 η_{ij} 的值越大，则人工蚁群则由城市 i 转移到城市 j 的期望程度就越高，然而在蚁群算法中每一个城市相当于一个数据聚类中心。因此如果 $y_n = h(x)$ 之值越小，则数据之间的相似度就越高，数据选择该聚类中的概率就越大。

当 d_{ij} 值越大，则 $\eta_{ij} = \frac{1}{d_{ij}}$ 就越小，并且 d_{ij} 值越小，根据 3.2 节公式（1）可以知道，此时需要聚类的数据偏离聚类中心的程度越大，数据与聚类中心数据的相似程度就越小，因此当 β 值越小，则 η_{ij} 的值就越小，由于 η_{ij} 反映由城市 i 转移到城市 j 的启发程度，因此当 η_{ij} 的值越小，人工蚁群则由城市 i 转移到城市 j 的期望程度就越小，然而在蚁群算法中每一个城市相当于一个数据聚类中心。因此如果 $y_n = h(x)$ 之值越大，则数据之间的距离就越大，数据选择该聚类中心的概率就越小。

（3）数据聚类过程的分析：

当蚁群选择一个聚类中心后，这条路径中的信息素浓度会逐步的增大，这时蚁群算法中的参数 Q 值就会逐步变大。然而 Q 值越大，这样又会吸引更多的人工蚂蚁来到这个聚类中心。在本文中人工蚂蚁代表的含义需要聚类的数据，因此这些根据蚁群搜索食物的原理，相似属性的数据都能在无监督学习的条件下聚集到某个聚类中心。

当蚁群发现下一个更佳的聚类中心时，那么原来路径中人工蚁群的数量会逐步减少，并且蚁群挥发的信息素浓度也会逐步减小，此时蚁群算法的参数 Q 值就会逐步变小。然而在通往新聚类中心的路径上信息素浓度又会逐步增强，Q 值就会逐步变大。如果 Q 的值越大，这样又会吸引更多的人工蚂蚁来到这个新的聚类中心，此时可以达到数据聚类的新目的。

蚁群算法在搜索食物源的过程是反复迭代的过程，因此采用蚁群算法作为聚类的过程也是一个多次迭代的过程，经过若干次反复迭代过程，需要聚类的数据都找到最好的聚类中心，最终达到数据之目的。

在 3.1 节的算法中对需要聚类数据进行了相似的计算，而在 3.2 节中采用蚁群算法为相似的数据选择了相应的聚类中心，因此经过 3.1 与 3.2 两部分算法对数据属性的计算可以完成对未知数据的聚类。

4　结束语

为了对海量数据进行相似性计算，需要采用聚类算法，聚类算法是一种智能型算法，因此在工程领域中有广泛地应用价值。本文采用数据属性的相似性计算实现了数据的聚类，而经典型的聚类算法是通过计算数据与聚类中心的距离来实现聚类的，这是不同于其他聚类算法的特征之处，蚁群算法在聚类算法中的再次应用也体现了该算法是一种寻找最优解的算法，由本文提出的聚类算法有自己的特色和不同于其它算法的地方，该算法在某种程度上具有一定的智能性，能够对某些数据进行聚类。

参考文献：

[1] 李玲娟，李　冰. 一种基于特征加权的蚁群聚类新算法 [J]. 计算技术与发展，2010，20（8）：67－70.

[2] 张建华，江　贺，张宪超. 蚁群聚类算法综述 [J]. 计算机工程与应用，2006（6）：171－174.

[3] 李　振，贾瑞玉. 一种改进的 K－means 蚁群聚类算法 [J]. 计算机技术与发展，2015，25（12）：28－31.

[4] 王　鹤. 蚁群聚类算法 [J]. 中国科技信息，2007（15）：280－281.

[5] 马春英，曹安得，周允征. 蚁群聚类组合的改进算法 [J]. 沈阳建筑大学学报（自然科学版）. 2011，27（4）：798－803.

[6] 田力威，曹安得. 基于信息熵的蚁群聚类组合算法的研究 [J]. 计算机应用研究，2011，28（4）：1269－1271.

[7] 向培素. 聚类算法综述 [J]. 西南民族大学学报（自然科学版），2011，37（5）：112－114.

[8] 陈一昭，姜　麟. 蚁群算法参数分析 [J.]. 科学技术与工程，2011，11（36）：9080－9084.

[9] 徐红梅，陈义保，刘加光，等. 蚁群算法总参数设置的研究 [J]. 山东理工大学学报（自然科学版），2008，22（1）：7－11.

[10] 方　媛，车启凤. 数据挖掘之聚类算法综述 [J.] 河西学院学报，2012，28（5）：72－75.

[11] 伍育红. 聚类算法综述 [J]. 计算机科学，2015，42（6）：491－524.

[12] 李卫军. K－means 聚类算法的研究综述 [J]. 现代计算机，2014（8）：31－36.

[13] 海　沫. 大数据聚类算法综述 [J]. 计算机科学，2016，43（6）：380－383.