DOI:10.19789/j.1004-9398.2015.06.006

基于人工蜂群算法的数据流聚类研究

程军锋
(陇南师范高等专科学校,甘肃 陇南 742500)

摘 要

聚类是数据挖掘领域非常重要技术之一,能发现数据中隐含的知识和价值,在数据流分析中有重要的意义.针对现有数据流聚类算法的不足,根据数据流特点提出了一种高效、可扩展的数据流聚类算法ABCCluStream.该算法借鉴了经典CluStream算法的在线—离线两阶段框架,根据蜂群算法的相关理论定义聚类过程中相关的参数、微簇聚类特征结构等,设计出了一种全新的算法.通过实验表明,本文所设计出的ABCCluStream算法具有聚类纯度高、可扩展性好的特点,能够产生高质量的聚类结果,适用于大规模的动态数据流聚类的分析和研究.

关键词: 数据流,聚类,蜂群,适应度,收益率.
中图分类号: TP311

1 引 言

近年来随着云计算、物联网等技术的发展,实时监控设备可以轻易获得大量的数据流数据,这些数据随着时间以不可预知的规模和速度连续到达.这些无限到达的数据由于数据量巨大,不可能把它们完整地存储起来按照传统的数据挖掘技术进行处理.

聚类是数据挖掘领域非常重要技术之一,能发现数据中隐含的知识和价值,是数据流分析中一种非常重要的技术.目前,已有一些针对数据流挖掘聚类算法,但这些算法大多都是对传统的数据聚类算法的改进,忽视了数据流的特点,不能适应数据流的动态变化.

针对上述问题,本文引入了人工蜂群算法的理论,针对数据流特点.根据蜂群算法的思想定义数据流聚类过程中相关的参数、微簇聚类特征等,以满足大规模的动态数据流聚类的分析和研究,提出了基于蜂群的数据流聚类算法ABCCluStream.

2 数据流聚类概述

聚类[1]是一个把数据对象集划分成多个组或者簇的过程,使得簇内数据对象具有很高的相似性,但与其他簇的对象很不相似,用于发现数据隐藏的模式.聚类方法作为数据挖掘的一种重要技术,已经被应用到数据流挖掘这个热点研究领域.

2.1 聚类的定义

聚类分析是一种无监督的机器学习分类方法,依据数据间的关联度量标准将数据集自动分成几个组群,使得簇中对象彼此相似,但与其他簇中对象不相似.通常以相似度和相异度作为相似度量标准.聚类就是将一组输入的有序数据集 $X=\{X_1,X_2,\cdots,X_m\}$,$X_i(i<m)$ 表示每个数据点,m 表示数据集中点的个数($m\geqslant 0$),根据样本间的相似和相异度量,生成一个分区$\{G_1,G_2,\cdots,G_n\}$,$G_j(j<n)$ 表示生成的聚类的每个分组,n 表示生成聚类的个数$0<n\ll m$),可表示如下:

$$G_1\cup G_2\cup,\cdots,\cup G_n=X \quad n\ll m$$
$$G_x\cap G_y=\varnothing \quad x\neq y,0<x\leqslant n,0<y\leqslant n$$

传统的聚类方法主要用于静态数据的聚类,常见的有:划分的聚类方法、层次的聚类方法、基于密度的聚类方法、基于模型的聚类方法、基于网格的方法等.

2.2 数据流聚类

与传统的静态数据相比,数据流是连续的、动态的流进流出系统的数据,主要应用场合包括网络监控、传感器系统、网络点击、股票市场、电力系统、流媒体等.这些数据源源不断地产生,并且随时间变

收稿日期:2015-01-03

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

化,我们把它们称为数据流. Henzinger 在 1998 年首次将数据流作为一种新的数据处理模型提出[2],他认为数据流是按照时间连续到达的数据序列. 数据流数据实时达到、数据达到次序不固定、数据量大且不可预知其最大值; 数据流一旦经过处理一般不保存且不能被再次取出处理.

数据流定义: 数据流是由一些带时间戳的多维记录 $X_1,X_2,\cdots,X_k,\cdots$ 所组成,它们所到达的时间戳依次为 $T_1,T_2,\cdots,T_k,\cdots$,表示数据点记录 X_i 在 T_i 时刻到达,同时约定 $T_i < T_j$ 时,数据点记录 X_i 比数据点记录 X_j 先到达. 每一个数据点记录 X_i 是一个 d 维向量,记作 $X_i = (x_{i_1},x_{i_2},\cdots,x_{i_d})$, x_{i_d} 分别代表数据点 X_i 的第 d 个属性.

数据流聚类的目标是将随时间而发生变化的流数据划分为若干个类,发现隐藏的数据模式和动态变化规律. 虽然在挖掘中底层数据流更加精确有效,但它们快速地动态变化,数据量大,数据流挖掘算法很难在较短的时间内对这些快速变化的数据进行处理. 数据流挖掘仅是得到这些数据的特性,而不是精确的结果. 由于最新到达的数据能反映数据流的变化特性,因此最重要. 所以数据流聚类算法要求对每个新到达的数据进行立即处理,之后这些数据或被保存,或被丢弃. 数据流自身的特性决定传统的静态数据聚类方法不可能直接用于数据流聚类. 与传统的聚类方法相比较,数据流聚类算法有以下要求: 在有限存储空间,单遍扫描,实时性处理最近到达的数据,并能够判断孤立点和高维数据. 数据流聚类就是要求在有限的存储空间内进行,研究设计出高效的大规模数据集的聚类算法.

虽然单层数据流聚类算法实现简单,但在算法设计中加入了访问的限制,使得它们在不同时间范围中进行聚类时非常困难,不能适应性数据流演绎变化特性. 而且单遍数据流聚类算法,如 Stream[3] 随着数据流变大计算开销增大,使得计算负担不断增加进而成为数据流在线应用的瓶颈. 另外,单层数据流聚类算法对于数据点到达顺序十分敏感,面对快速的数据流,它已经不能适应数据流的特点. 因此,通常对数据流的挖掘是在精度和数据量上折中,大多采用单遍的、联机的、多层的、多维的分析方法.

双层结构的数据流聚类框架考虑了数据流特点,被证明非常符合数据流聚类. 双层数据流聚类框架如图 1 所示,主要分为两个阶段: 在线层和离线层. 在线层算法负责对数据流新到达的数据做快速

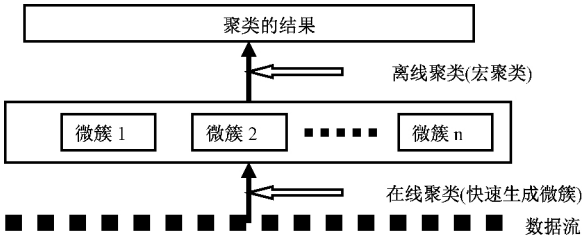


图 1　双层数据流挖掘框架

但较为粗糙的处理,将连续到达的数据流转化为概要数据,以中间结果的方式保存,为离线算法的输入准备,在线算法要求相对简单而且高效,追求时间效率,可以更好地适应较快的流速; 离线算法对离线阶段产生的概要数据进行处理,更加偏重于聚类结果的质量,对时间效率要求不高,所以采用较为复杂的处理方式,不必考虑数据流速的影响,通过更加复杂、更加精细的聚类算法得到最终的聚类结果. 双层算法框架将处理工作分配在整个数据流过程中,既考虑了底层数据的快速处理能力的特点,又兼顾上层用户的要求,有更高的算法效率.

双层结构算法在线层利用微聚类这种数据结构来保留数据的概要信息(梗概),当用户需要聚类查询,数据流聚类算法以概要数据结构中保存的微聚类作为挖掘对象,产生用户需要的结果. 微聚类就是通过聚类特征向量微簇结构(micro-cluster structure)来记录每个微聚类的信息. BIRCH 算法[4]中使用了聚类特征 CF 的概念,它用一个三维向量汇总了对象簇的信息. Aggarwal 等人在其基础之上,扩展了聚类特征的概念,提出了微簇的概念[5].

微簇使用一个 $(2d+3)$ 元组的特征向量 $(\overline{CF2^x},CF1^x,CF2^t,CF1^t,n)$ 来描述. 定义如下:

1) 对于 d 维数据的每个维度而言,数据值的平方和结果保存在矢量 $\overline{CF2^x}$ 中. 因此 $\overline{CF2^x}$ 一共包含 d 个值,$\overline{CF2^x}$ 的第 p 个条目可以表示为 $\sum_{j=1}^{n}(x_{ij}^p)^2$;

2) 对于 d 维数据的每个维度而言,数据值的求和结果保存在矢量 $\overline{CF1^x}$ 中. 因此 $\overline{CF1^x}$ 一共包含 d 个值,$\overline{CF1^x}$ 的第 p 个条目可以表示为 $\sum_{j=1}^{n}x_{ij}^p$;

3) 时间戳 $T_{i_1}\cdots T_{i_n}$ 的平方和结果保存在标量 $CF2^t$ 中;

4) 时间戳 $T_{i_1}\cdots T_{i_n}$ 的求和结果保存在标量 $CF1^t$ 中;

5) 数据点的个数保存在 n 中.

微簇继承了 CF 数据结构优异的拆分、合并特

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

性. 从结构来看,微簇基本上就是对聚类统计信息的总结. 它包含了聚类计算和空间存储利用所需的关键信息,可以用于进一步的聚类分析,在准确表示聚类的同时,降低了聚类分析的空间复杂度和时间复杂度. CluStream 算法[6]作为数据流聚类里程碑式的开创性算法,提出了在线—离线两阶段框架,有效地解决了数据流动态变化的问题.

双层算法在工作时,首先吸收足够的数据点,进行初始化时,在线算法利用进行微聚类得到微簇,产生每个微簇的五元组特征向量. 当新的数据点到达时,在线算法首先尝试找到离该数据点最近的微簇,如果有满足给定条件的微簇,则微聚类将该点吸收,修改代表微聚类的五元组的值;否则,以该数据点创建新的微簇. 为了创建微簇,有可能需要删除现有某个微簇或合并两个现有微聚类来释放空间.

3 蜂群算法的思想

蜂群算法建立在蜜蜂采蜜过程的模拟基础之上,属于智能群体优化算法[7],常见的群体智能算法有蚁群算法(ACO)[8]、粒子群算法(PSO)[9]、鱼群算法(AFSA)[10]、人工蜂群算法(ABC)[11]等. 从 2005 年 Karaboga 系统地提出了人工蜂群算法(Artificial Bee Colony, ABC)模型以来,由于它具有较强的灵活性、鲁棒性、不受领域知识所约束等特点. 作为一个新的研究方向,ABC 算法引起了许多学者的关注,现在已经取得了一些研究成果,并在人工智能、神经网络和数据挖掘等领域有着广泛应用.

蜂群算法来源于人类对蜜蜂采蜜过程的认知. 在蜜蜂采蜜过程中,不同的蜜蜂有着不同的分工,并且它们相互之间通过共享信息,发现最优蜜源. ABC 算法求解实现采蜜的群体智能行为的过程需要 4 部分:食物源(蜜源)、采蜜蜂、跟随蜂和侦察蜂. 食物源在 ABC 算法中对应候选解所处的位置,候选解质量的优劣与食物源的收益度值成正比例,其表示为收益率函数,决定收益率函数值的多少有诸多因素;采蜜蜂(引领蜂)发现食物源并记录下与食物源收益率有关的信息然后将这些信息以一定的概率与其他蜜蜂分享,与数量与食物源成一一对应的关系(引领蜂的个数 = 食物源的个数),即每个食物源都有一只引领蜂与之对应. 跟随蜂是等待在蜂巢中通过观察引领蜂分享的信息决定跟随哪个引领蜂并到相应的蜜源进行采蜜. 跟随蜂一般按照轮盘赌的方法选择引领蜂. 食物源收益率的多少与食物源被选中的概率成正比例关系;侦察蜂负责一旦解出现停滞现象时在蜂巢周围随机搜索发掘新的食物源,增强 ABC 算法跳出局部最优解的能力,当新食物源出现的时候,侦察蜂记忆下新食物源的相关信息并采蜜,同时进行角色的改变,由侦察蜂变为采蜜蜂.

食物源的花蜜量对应于相应解的质量,用适应度 fit_i 表示,用适应度来表示收益率,fit_i 根据下式计算:

$$\mathrm{fit}_i = \begin{cases} \dfrac{1}{1+f_i} & f_i \geqslant 0 \\ 1+|f_i| & f_i < 0 \end{cases} \tag{1}$$

其中 f_i 是目标函数,跟随蜂按照概率值 P_i 选择食物源,P_i 根据下式计算:

$$P_i = \frac{fit_i}{\sum_{j=1}^{FN} fit_i} \tag{2}$$

引领蜂和跟随蜂依据表达式(公式 3 -3)进行食物源位置更新:

$$v_{ij} = x_{ij} + \varphi_{ij}(x_{ij} - x_{kj}) \tag{3}$$

其中 $j \in \{1,2,\cdots,FN\}$, $k \in \{1,2,\cdots,M\}$,其中 FN 为食物源的个数,M 为食物源的维度空间,x_{ij} 为当前的食物源,x_{kj} 为随机选择是新食物源,用来控制领域搜索的大小,便于向最优解靠近. j,k 都是随机选取的,$k \neq i$,而 φ_{ij} 是 $[-1,1]$ 之间的随机数,控制着领域的搜索范围. 越接近最优解,$(x_{ij} - x_{kj})$ 之间的距离缩小,搜索的空间也缩小.

对于采蜜蜂陷入局部最优,经过 Limit 次迭代依然没有变化,而跟随蜂所得适应度又不是当前全局最优时,放弃该食物源,侦察蜂则通过表达式(公式 4)产生一个新的解来替代原. Limit 是 ABC 算法中一个重要的控制参数,可根据经验或者统计知识进行选取.

$$x_i^j = x_{\min}^j + rand(0,1)(x_{\max}^j - x_{\min}^j) \tag{4}$$

其中 $j \in \{1,2,\cdots,D\}$,$x_{\max}^j$ 表示变量的最大值,$x_{\min}^j$ 表示变量的最小值.

该算法的主要过程如下:

(1)采蜜蜂在其所附食物源的邻域内进行局部搜索,一旦发现附近有好的食物源时,更新所依附的食物源;

(2)跟随蜂根据由采蜜蜂所提供的食物源相关信息,按某种概率依附到特定食物源,然后每只跟随蜂在所附食物源的邻域内进行局部搜索新食物源. 当发现好的食物源时. 通知相应的采蜜蜂更新其食

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

物源的位置. 在此过程中,好的食物源(适应度高的蜜源) 会吸引较多的跟随蜂. 随着对邻域多次进行搜索,这些食物源会得到更频繁的进化;

(3) 侦察蜂当解的进化陷入局部最优时进行处理,采蜜蜂放弃原来依附的食物源成为侦察蜂,然后随机寻找一个可行解作为食物源位置重新开始新的搜索.

蜂群算法描述如下:

(1) 初始化: 设定食物源个数(FN) 、最大循环次数 MCN 和更新解的判断参数 Limit;

(2) 产生初始解集,对于初始解集中的每个解 x_{ij},根据公式(1) 计算解 x_{ij} 的适应度;

(3) 采蜜蜂根据公式(3) 做邻域搜索产生新解 v_{ij}, 并根据公式(1) 计算其适应度;

(4) 比较 v_{ij} 的适应度大于 x_{ij} 的适应度,则 x_{ij} 被 v_{ij} 替换; 否则,x_{ij} 不变; //对初始解进行优化

(5) 计算 x_{ij} 的适应度,并根据公式(2) 计算概率值 p_{ij};

(6) 跟随蜂根据 p_{ij} 选择食物源,并做邻域搜索产生新解 v_{ij}, 计算其适应度

(7) 比较 v_{ij} 的适应度大于 x_{ij} 的适应度,则 x_{ij} 被 v_{ij} 替换; 否则,x_{ij} 不变; //领域搜索新最优解

(8) 经过 Limit 次循环后,判断是否有要丢掉的解,若存在,则侦察蜂根据公式(4) 产生一个新解代替 x_{ij};

(9) 若当前迭代达到最大循环次数,则停止迭代,返回最优解; 否则转到(3) 继续求解.

4　基于蜂群算法的数据流聚类算法

4.1　基于蜂群算法的数据流聚类算法的相关知识

聚类算法就是根据事物之间的某些潜在属性,把事物聚集成类,使得不同类中的事物尽可能的不相似,而同一类中的事物尽可能相似. 出于以上考虑,将目标函数定义为与两个指标有关的函数分别表示簇内的聚合程度和离散程度. 表示聚合程度部分越小表明簇越紧凑,越相似,表示离散程度部分越大表明簇之间越分散,越不相似. 对于在线阶段每个微簇的目标函数可定义如下.

$$f_i = \sum_{k=1}^{q} \sum_{i \subset z} \left[\varphi(x) \cdot \frac{n}{|\overline{CF2_i^x} - \overline{CF2_z^x}| + |-(\overline{CF1_i^x} - \overline{CF1_z^x})^2|} + (|\overline{CF1_i^x} - \overline{CF1_k^x}|^2 - |\overline{CF2_i^x} - \overline{CF2_k^x}|) \right] \quad (5)$$

其中 $z \neq k$; $z,k \in (1,2,\cdots,q)$, q 表示微簇的个数,M_z 是微簇 M_i 聚类的中心,M_k 是另一个聚类的中心,$\varphi(x)$ 是一个修正函数用于定义前半部分的权重. 该函数的前半部分为用于衡量簇内紧密程度的函数,后半部分为用于衡量簇间离散情况的函数.

ABCClustream 算法使用在线—离线两阶段框架来进行数据流的聚类分析,在此过程中涉及到微簇的概念,而表示微簇概念的一般方式是微簇聚类特征,由于表示微簇聚类的特征相关知识在 2.3 中已有介绍,而对于本文所提的 ABCClustream 算法,为了反映微簇的活跃程度,对微簇进行更新,沿用原始的微簇聚类结构是不合理的,因此本文对原始微簇聚类特征结构进行了修改,介绍如下:

ABCCluStream 算法的微簇聚类特征使用 $(\overline{CF2^x},\overline{CF1^x},CF2^t,CF1^t,n_i,t_1,t_2)$ 来定义,它是一个 7 元组. 其中,$\overline{CF2^x}$,$\overline{CF1^x}$,$CF2^t$,$CF1^t$ 和 n_i 的定义和 CluStream 算法中的定义一致, t_1 上一次更新的时间,t_2 表示该微簇的最后更新时间.

为了插入新簇,当存储空间不足时,定义一个用于删除微簇的标量,为了反映数据流的特点,需要删除的微簇应该是最近不太活跃的微簇,而且里面包含较小的数据元素,该标量可用以下函数来衡量,函数定义如下:

$$f_t = (t_2 - t_1)^2 + \lambda(x)(CF2^t - (CF1^t)^2) \quad (6)$$

其中 $\lambda(x)$ 为修正函数,控制 f_t 的大小在一个合理的范围,该函数主要用于在空间不足删除微簇和时间片到孤立点检测时参考.

4.2　基于蜂群算法的数据流聚类算法思想

ABCClustream 算法借鉴经典算法 CluStream 的思想,利用蜂群算法中蜜蜂觅食基础上来对数据流进行分析. 算法由在线和离线两个部分构成,在线部分快速接受输入的数据流进行微聚类,并且基于修改的 micro-cluster 结构定时存储数据流的概要信息,随着新数据点的到达该结构实时进行更新; 离线部分 macro-cluster 通过对在线部分保存的 micro-cluster 结构再处理,以便得到用户感兴趣的不同时间范围内数据流聚类结果. ABCClustream 算法采用金字塔时间型的时间窗口[3] 分级保存数据的摘要信息,在线部分用先用 K-means 算法将其进行聚类,对新的数据点到达,微聚类要对数据流的变化做出反应. 算法采用判断采用距离确定归属点微,然后对微簇进行更新、添加、合并和删除维护和离群点检测、用于发现任意形状的数据点,产生高质量的微

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

簇. 当内存空间不足时删除一个已有微聚类,考虑刚产生的包含少量样本的微聚类有可能成为新模式,所以删除这种微聚类存在风险. 为了避免误删除新生聚类,本为采用了一种基于时间定义函数公式(6)的策略来删除某个微簇. 对于可能的孤立点,并不立即将其抛弃,而是先作为新微簇插入,由于它有可能会成为有效数据. 然后根据时间片周期检测做判断删除,防止离群点删除影响了聚类效果. ABCClustream 算法增加了一个离群点删除检测机制,基于自定义函数公式(6)选择低于某个阈值的离群点进行删除. 由于微聚类一旦被合并将无法恢复,所以仅当空间不足尽可能选择合适聚类可删除,当没有合适的删除时才考虑合并操作. 离线聚类阶段,根据用户指定要观察的时间窗口长度 h,并指定期望获得的聚类数量 K. 因为微聚类反映的是其创建以来数据流的统计信息,所以需要利用微聚类的可减特性获得指定时段内微簇的统计信息. 然后使用该信息根据用户的要求进行宏聚类.

4.3 ABCclustream

其 ABC 算法中,食物源的位置对应问题的一个可能的聚类(聚类中心),食物源的收益度对应聚类问题的相似度,寻找食物和采蜜的速度对应聚类的求解速度,最大收益度对应聚类问题的最佳结果.

ABCclustream 在线算法描述

Online-ABCCluStream ( )

(1) 初始化,得到 k 个微聚类

(2) 当新的数据点达到,判断该点和已有微簇的关系

(3) if(该点和已有的第 k_i 个微簇非常相近)

(4) 更新微簇 k_i 的结构,准备接收下一个数据点

(5) else

(6) if (微聚类没有到达上限)

(7) 创建一个新微聚类

(8) else

(9) if (满足微聚类删除条件)

(10) 根据条件选定某个微簇删除

(11) else

(12) 根据插入的节点和其他簇之间的关系选择微簇合并

(14) When(检测时间片 σ 到)

(15) if (判断微簇是否为孤立点)

(16) 则抛弃该点,接收下一点

ABCClustream 离线算法描述:

Offline-ABCClustream Algorithm ( )

(1) 从磁盘上取出 horizon 时间范围内的微簇

(2) 根据用户指定聚类数目 k,随机选取 k 个微簇并计算每个簇的收益度

(3) 根据概率 p 随机的在先前聚类的邻域类,选择一个微簇,并计算它的收益度

(4) 判断该领域类的每个微簇和原有簇之间的收益度

(5) if(收益度高)

(6) 则用该微簇替换原有微簇

(7) when(limit)

(8) 对于未访问的微簇利用以上方法更新候选解

(9) 把结果进行输出

5 算法的实验结果仿真

本实验将从三个方面对 ABCCluStream 算法的性能进行比较,它们分别是聚类质量、算法的有效性和可扩展性.

5.1 聚类质量的验证

平方距离和(SSQ)只适用于评价球状簇,对于任意形状簇,该方法的评价效果不佳,所以本实验选取聚类纯度[12]作为衡量算法性能的标准. 聚类纯度定义为聚类结果中数据点正确聚类所占的比例,比例越大则质量越好. 实验数据采用 KDD-CUP99 网络入侵检测数据集,实验结果如图 2 所示.

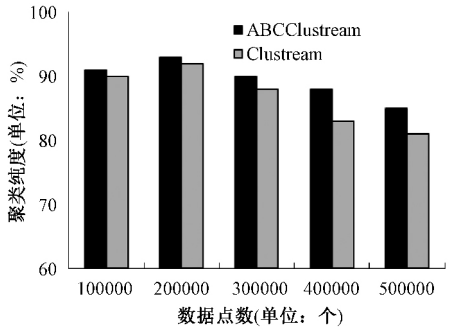


图 2 ABCCluStream 算法的聚类纯度比较

从图 2 中可以看出,两种算法的聚类纯度都不低于 80%,但是 ABCCluStream 算法的聚类纯度更高一些,这是因为该算法在离线阶段采用了蜂群算法,避免了全局距离参数对聚类结果的影响,因此数据点能够自然地分配到簇中.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

5.2　有效性的验证

有效性是衡量数据流算法的重要指标,本实验采用数据流处理速率来测量 ABCCluStream 算法的有效性,实验数据采用 KDD-CUP99 网络入侵检测数据集,实验结果如图 3 所示.

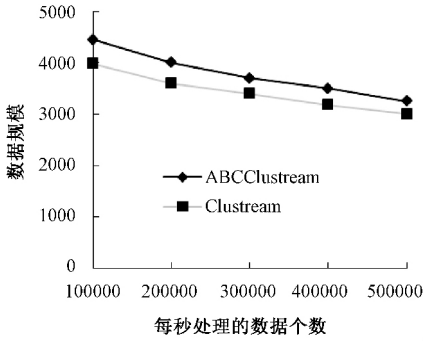


图 3　ABCCluStream 算法的有效性比较

可以看出,这两种算法的处理速率都可以达到了每秒上千个数据点,都属于高效的数据流聚类算法. 但是,实验结果表明 ABCCluStream 算法的处理速度高于 CluStream 算法,这是因为该算法在迭代执行的过程中使用适应度函数不需要大量计算距离,更新微簇时对于聚类特征的计算也较简单,因此大大节省了处理时间.

5.3　可扩展性的验证

为了验证该算法的扩展性,本实验采用数据流处理速率来测 ABCCluStream 算法的有效性,实验数据采用 KDD-CUP99 网络入侵检测数据集,实验结果如图 4 所示.

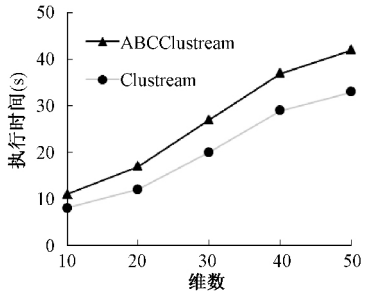


图 4　ABCCluStream 算法在维度上的可扩展性(流速 v = 100)

从该图中可以看到,随着数据流维度和簇数目不断地增长,该算法的处理时间基本也都呈线性增长,说明本文所提算法比较稳定,能够适应不同规模的数据集.

6　结束语

本文研究现状数据流聚类的相关算法和蜂群算法的基本知识与概念. 引进蜂群算法的适应度和收益率等概念,并且在宏聚类阶段使用了蜂群算法,对微簇的结构进行重新定义,改进了在线阶段微簇更新策略,改进了算法对发现任意形状数据的能力,通过对离群点进行检测删除,增强算法适应性,然后,根据数据流快速、海量等特点设计出高效的聚类算法 ABCCluStream. 通过真实数据集的验证表明该算法聚类纯度高、可扩展性好,ABCCluStream 是一种高效的数据流聚类算法.

参　考　文　献

[1] 范明,孟小峰. 数据挖掘概念与技术(第三版) [M]. 北京: 机械工业出版社,2012: 288 – 319.

[2] O'Callaghan L, Mishra N, Meyerson A, et al. Streaming-data algorithms for high-quality clustering [C]. //Proceedings of the 18th International Conference on Data Engineering. San Jose: IEEE Computer Society. 2002: 685 – 694.

[3] Henzinger R M, Raghavan P, Rajagopalan S. Computing on data stream [C]. //Proceedings of the SRC Technical Note. California: SRC Press, 1998: 107 – 118.

[4] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases [J]. ACM SIGMOD Record,1996,25(2) : 103 – 114.

[5] Aggarwal C C, Han J, Wang J, et al. A framework for projected clustering of high dimensional data streams [C]. //Proceedings of the 30th International Conference on Very Large Data Bases. Toronto: VLDB Endowment,2004: 852 – 863.

[6] Aggarwal C C, Han J, Wang J, et al. A framework for clustering evolving data streams [C]. //Proceedings of the 29th International Conference on Very Large Data Bases. Berlin: VLDB Endowment,2003: 81 – 92.

[7] 王辉,钱锋. 群体智能优化算法 [J]. 化工自动化及仪表,2007,34(5) : 7 – 13.

[8] Colorni A, Dorigo M, Maniezzo V. Distributed Optimization by Ant Colonies [C]. Proceedings of the First European Conference on Artificial Life, 1991: 134 – 142.

(下转第 35 页)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Causes and Countermeasures of University Physics Learning Difficulties in Electrodynamics Learning ——Based on two Normal Universities in South H Province in the Middle of China

Zhang Tingyu[1]　Hu Yangyang[2]

(1. Department of Physics, Capital Normal University, Beijing　100048;
2. College of Education, Capital Normal University, Beijing　100048)

Abstract

In this article we analyzed the status quo of reformation of electrodynamics teaching in normal schools, basing on the specific aspects of electrodynamics course, the status quo and real experiences of teaching reformation. Then we proposed a couple of suggestions concerning the reformation of electrodynamics course contents, teaching methods and concepts.

Key words: electrodynamics, learning difficulties, physics teaching, normal education.

(上接第 29 页)

[9] Kennedy J, Eberhart R C. Particle swarm optimization [C]. Proceeding Soft the IEEE International Conference on Neural Networks, Perth, Australia. Piscataway, NJ: IEEEPress, 1995: 1942 – 1948.

[10] 李晓磊,邵之江,钱积新. 一种基于动物自治体的寻优模式: 鱼群算法 [J]. 系统工程理论与实践,2002,22(11): 32 – 38.

[11] Karaboga. An idea based on honey bee swarm for numerical optimization. Technical report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department, 2005.

[12] Aggarwal C C, Han J, Wang J, et al. A framework for projected clustering of high dimensional data streams [C]. Proceedings of the 30th International Conference on Very Large Data Bases. Toroto, Canada, 2004: 852 – 863.

Clustering of Data Streams Based on Artificial Bee Colony Algorithm

Cheng Junfeng

(Longnan Teachers College, LongnanGansu　742500)

Abstract

Clustering is one of the very important technology in data mining field, which can find the hidden knowledge and value in the data stream, it has an important significance in the data flow analysis. In this paper, aiming at the shortcomings of the existing data stream clustering algorithm, according to the characteristics of a data flow, propose an high efficient and extensible calle ABCCluStream The algorithm use the classic CluStream algorithm of online-offline two-stage framework, based on the theory of the artificial bee colony defines related parameters and micro-cluster clustering feature structure in the process of clustering. Experiments show that this clustering of the ABCCluStream algorithm is designed with good scalability and high purity of clustering, and produce high quality of the clustering results, suitable for analysis and research on the clustering of dynamic large-scale data stream.

Key words: data stream, clustering, artificial bee colony, fitness, yields.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net