第 10卷 第 35期 2010年 12月
1671—1815(2010)35-8752-08

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 10 No. 35 Dec. 2010
© 2010 Sci. Tech. Engng.

综 述

计算机技术

增量聚类算法综述

李桃迎 陈 燕 秦胜君 李 楠
(大连海事大学交通运输管理学院,大连 116026)

摘 要 给出了增量聚类的概念。分析了增量聚类方法可以用于解决数据的变化和大量存储空间的需求问题。增量聚类算法选择恰当时,可以保证数据在变化时有效地提高聚类的精度和效率。从传统聚类、生物智能聚类和数据流聚类三个角度归纳了增量聚类问题,分析了增量聚类问题的研究进展:包括发展的过程及特点,阐述了研究增量聚类问题的关键技术,最后给出了未来的发展趋势。

关键词 聚类分析 增量聚类 生物智能 数据流
中图法分类号 TP301.6; 文献标志码 A

聚类就是将数据点划分成组同时满足组内数据点之间的相似性尽可能大,不同组的数据点之间的相似性尽可能小[1],聚类在数据挖掘中起着非常重要的作用,并广泛地应用于模式识别、计算机可视化、模糊控制等领域。

随着信息技术的发展,特别是 Web的出现,数据和环境无时无刻不在发生变化,需要更多的空间存储数据,如何解决大量数据的存储问题成为当前一个迫在眉睫的问题。增量聚类由于有限的空间需要而被提出来,即不需要将所有数据存储到内存中[2]。目前有关增量聚类的研究主要是将增量数据看成是时间序列数据或按特定顺序的数据,主要可以分成两类:一类是每次将所有数据进行迭代,即从第一个数据到最后一个数据进行迭代运算,其优点是精度高,不足之处是不能利用前一次聚类的结果,浪费资源;另一类是利用上一次聚类的结果,每次将一个数据点划分到已有簇中,即新增的数据点被划入中心离它最近的簇中并将中心移向新增的数据点,也就是说新增的数据点不会影响原有划分,其优点是不需要每次对所有数据进行重新聚类,不足之处是泛化能力弱,监测不出孤立点。因此,如何设计增量聚类算法以提高聚类效率,成为当前聚类分析的一个重要挑战。

1 增量聚类方法综述

目前存在各种各样的聚类方法[3],传统的聚类方法主要被划分成五类:基于层次的、基于划分的、基于密度的、基于网格的和基于模型的聚类。基于层次的聚类和基于划分的聚类是实际生活中应用最为广泛的两类。前者可以进一步划分为自底向上和自顶向下两种[1],例如 CLIQUE[3]、ENCLUS和 MAFIA[4]属于自底向上算法,PROCLUS[5]和 OR-

2010年 10月 8日收到 国家十一五科技支撑计划项目(2009BAG13A03)、国家自然科学基金资助项目(70940008、70801007)、教育部博士点基金项目(200801510001)、教育部科学技术研究重点项目(209030)、中央高校基本科研业务费专项资金(2009QN085)资助

第一作者简介:李桃迎(1983—),安徽人,博士生,E-mail: ytaoli@126.com。研究方向:数据挖掘。

http://www.kxjsshu.com

CLUS[6]属于自顶向下的算法。但是,传统的层次聚类算法由于计算量过大不适用于大数据集,例如BIRCH[2]和CURE[7]。传统的基于划分的算法包括k-means、k-modes等等,其中k-means是现存聚类算法中最经典的聚类算法[8,9]。

增量聚类是维持或改变k个簇的结构的问题。比如,一个特定序列中的新的数据点可能被划分到已有k个簇的一个簇中,也可能被划分到新的簇中,此时会需要将另外两个簇变成一个[10]。自从Hartigan在文献[11]中提出的算法被实现[12],增量聚类就吸引了众人的关注。D. Fisher[13]提出的COBWEB算法是一种涉及到增量形式数据点的增量聚类算法。文献[14,15]中给出了与数据库的动态方面相关的增量聚类的详细阐述,文献[16—18]中列出了其广泛应用的领域。对增量聚类产生兴趣的动力是主存空间有限,有些信息不需要存储起来,例如数据点之间的距离,同时增量聚类算法可以根据数据点集的大小和属性数进行扩展[19]。文献[10,17]中也对于求解增量聚类问题的算法进行了研究。

现在很多聚类算法都是对单一数据类型的数据进行聚类,但是现实数据中非常多的数据都是混合数据类型的数据,既包含数值属性数据,还是分类属性数据,简单地丢弃其中一种数据类型,或者将其中一种数据类型转换成另一种,都会影响聚类的精度。因此,混合属性数据增量聚类的研究具有非常重要的意义。

2 基于传统聚类方法及其变形的增量聚类算法

现在对于增量聚类方法的增量处理主要集中在三个方面,一类是基于传统聚类方法及其各种变形的增量聚类算法,一类是基于生物智能的增量聚类算法,另一类是针对数据流的聚类算法。

2.1 方法概述

有的传统聚类方法同样适用于增量模式的聚类,如BRICH和COBWEB算法。也有些是在传统聚类算法的基础进行了变形,来满足增量聚类的需要。文献[20]中首次提出了增量聚类的概念,也就是增量的DBSCAN,它是基于DBSCAN的基础上提出的。由于DBSCAN算法是基于密度的特性,插入或删除一个新的数据点只影响当前聚类中近邻该点的簇,这种方法的优点是它的聚类结果和非增量聚类的结果相似,但是它的不足是只能一个一个的划分数据点,从而导致聚类的效率很低。文献[21]中提出了基于网格的增量聚类,其类似于增量的DBSCAN。Huang和Zou与Xu和Xie[22,23]采用批量处理的基于密度的增量聚类,克服了一个一个处理数据的缺点,以批量的形式处理数据,但是用这种聚类方法由于计算量过大而不能用于大数据集。文献[24]中描述了一种高效的基于密度的增量聚类算法,利用划分和抽样技术来处理大数据集,在划分高维数据时会产生抽样误差。

Chen等[25]依据物理学中的重力理论提出了一种增量的层次聚类,即GRIN算法。该算法分为两个阶段,首先,它把到达的增量数据缓冲在一个数据池中,从池中选出一些样本数据对其建立树状图(dendrogram),删除包含数量过少的簇,去除噪音数据等过程建立暂时的树状图。GRIN的第二个阶段就是处理数据池中的其他数据,即确定待处理的数据是否应插入第一阶段得到的图中的叶节点。如果该数据属于多过两个叶节点,就用重力学原理确定它最终属于哪个叶节点。虽然GRIN具有较好的聚类质量及线性的时间复杂度,对数据输入顺序和参数值的设定不敏感,但是,GRIN实质上并不是真正意义上的增量聚类算法,而是批处理的方法。Widyantoro等[26]提出了凝聚的增量层次聚类算法——IHC,该方法的目的是构建一个拥有两个性质的概念层次:同质性(homogeneity)和单调性(monotonicity)。同质簇即为簇内对象有相似密度,而在层次聚类的簇中,单调性是指一个簇的密度总是高于其父辈簇。Charikar等[27]基于信息检索的需求,提出了基于层次凝聚的增量聚类算法,即当以增量数据提交给算法时,要么分配给已知簇,要么形成一新簇同时合并两个已知簇。这些在传统聚

类算法基础上提出的增量算法基本上都是采用直接对学习结果重新组织反映数据的动态变化,耗时耗空间,可扩展性差,没能实现数据特征的提取,也没有起到数据压缩的作用。

目前相关文献中,只有 Hsu和 Huang[28]针对混合属性数据的增量聚类进行研究,借用概念层次树来求解混合型数据的相似度,但是要求用户必须清楚地了解属性的值,对任意属性(包括分类属性)的取值范围给出大小关系并设置有效地差值,不恰当的设置会导致误差非常大,由于增量的数据存在未知的因素,所以此方法不适用于最终的求解。

2.2 关键技术

现在有关增量聚类的研究主要集中在三个方面:聚类的初始化问题、聚类过程中簇的调整和聚类的有效性。很多聚类算法对初始化方法的选择非常敏感,其选择的优劣直接会影响聚类的结果,所以初始化问题是一个值得关注的问题;随着数据的逐渐增加,原有簇的结构及归类方法应该如何调整,以此提高聚类的效率和精度,是一个不可忽视的问题;如果聚类算法和参数选择不当,会导致聚类结果与实际不符,所以聚类结果的有效性问题成了一个对聚类算法至关重要的方面。

2.2.1 簇的初始化

聚类的初始化方法主要划分为三类:随机抽样法、距离优化法和密度评估法[29]。其中,随机抽样法是出现最早且最为常用的算法,它是通过从样本集中随机抽取 k个样本作为初始 k个簇的中心,进而根据目标函数的最小化问题来对簇进行调整。这种算法相对比较简单,易于实现,但是它的目标函数易于陷入局部极小点。距离优化法考虑到类内距离尽可能小的同时,考虑类间距尽可能的大,下一个待选簇的中心和现有簇的中心之间的距离要尽可能的大,这种方法也很简单,但是对数据集的多次扫描和距离的计算导致计算量过大,同时无法识别孤立点。密度评估方法认为数据集的数据服从高斯混合分布的假设,将输入域的密集区域本身作为包含自然的聚类划分,通过辨识数据的密集区域来得到初始聚类中心,但是当数据量很大时,计算量过大[30]。

现有很多对聚类初始化问题的各种改进方法[31—37]。张霞等[31]在数据对象的模糊粒度空间上给定归一化距离函数,对距离小于给定值的数据对象进行初始聚类,其个簇的中心作为初始的聚类中心。田生文等[32]面向复杂网络的特征,选择其对象的度、聚集度和聚集系数高的 k个点作为 k-means的初始簇的中心。而杨圣云等[33]则用数据集的多个子集和山函数来对 k-means进行初始化。钱线等[34]首先估计出 k个簇的特征中心的位置来对 k-means进行初始化。申晓勇等[35]针对基于目标函数的直觉模糊聚类方法容易陷入局部点的问题,将样本密度函数在较高局部密度的区域中选取 c个样本,遍历剩余样本进行粗归类,将每类各维数据的平均值选取为初始聚类中心。马秀丽等[36]利用待聚类数据所包含的空间邻近信息和特征相似性,引入 K调和平均算法来克服原始谱聚类算法对初始化的敏感性。盛莉等[37]采用网格和密度的方法对模糊 C均值算法进行初始化。

2.2.2 增量过程中簇的调整

对增量聚类进行聚类时,由于新数据的到达,可能需要对已有簇进行调整,即合并、拆分或产生新簇,基本上都是在增量聚类算法的聚类过程中讲述如何对簇进行调整。Edwin Lughofer[38]通过扩展矢量量化方法对增量数据进行聚类,过程中设定一个阈值,当新增数据与现有中心点的最小距离小于这个阈值时,就将其归入现有簇中,否则增添一个新的簇,并将该数据点当作新簇的中心。吴琪等[39]处理新增的数据点时,如其在原有聚类结果某一类的范围内,那就将其归入该内,否则当作孤立点,增加新的一类。对于新增数据点的同时,删除已有的部分数据点的情况,如果经过删除的簇分布比较均匀,仍然作为一类,否则需要进行类分裂处理。刘建晔等[24]处理增加数据点时,如果新增数据点所处网格原来是密集的,则该网格不需继续运算,其肯定是密集的。否则,计算该网格的密度,当为密集单元时,需要查看相邻单元格,如果相邻单元格中有一个或多个是密集且属于某一类,则该网格也属

于此类;如果没有相邻单元格属于某一类,则该网格属于新增的一个类;如果该网格与两个个或两个以上类相邻,则合并这几个类。否则,如果该网格非密集单元,则该网格不属于任何一类。删除数据点时,判断删除的数据点所处的网格,如果原来不是密集的,则无需计算,肯定不是密集的。否则计算该网格的密度,如果是密集的,则聚类不变,否则判断相邻单元格,如果只有一个单元格是某类,则在此类中删除该单元格;如果没有单元格属于某类,则删除该单元格,也就是该单元格原来所处的那个类;如果有两个或两个以上相邻单元格属于同一类,则删除该单元格,同时检查这些类的连通性,若这些单元格相邻,则直接删除该单元格,如果不相邻,则需要将该类划分成若干类。

2.2.3 聚类的有效性度量

聚类有效性函数的定义方法目前主要分为3类[40]:基于数据集模糊划分的方法;基于数据几何结构的方法和基于数据统计信息的方法。

聚类有效性度量多采用多个指标进行研究,根据对数据集的分离性与紧致性的定义方式的差异,目前已有不少基于几何结构的聚类有效性函数。其中,Xie和Beni将模糊聚类的目标函数与聚类分离度相结合提出的紧致度和分离度有效性函数具有一定的代表性[41],具体如式(1)。

$$V_{xie}(U,V,c)=\frac{\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\|v_i-v_j\|^2}{\min\limits_{i\neq j}\|v_i-v_j\|^2} \tag{1}$$

该方法通过在类内紧凑度和类间分离度之间找一个平衡点使目标函数达到最小。但是当 $c->n$ 时,该函数接近于0,难于发现最优聚类数 c。随后出现了很多对该算法改进的研究[42,43]。

彭勇等[44]根据模糊不确定性理论及聚类问题的特性,基于数据集的紧密度与分离度特征,综合考虑了数据成员的隶属度及数据集的几何结构,提出了一个旨在寻求最优聚类数的有效性函数。安中华等[45]将多元分析中的F-统计量和变化的显著性水平引入模糊聚类分析中进行有效性分析,其F-统计量的形式如式(2)所示。

$$F_l=\frac{\frac{1}{l-1}\sum_{j=1}^{l}r_{jl}\sum_{t=1}^{m}(\bar{x}_t^{(jl)}-\bar{x}_t)^2}{\frac{1}{n-1}\sum_{j=1}^{l}\sum_{k=1}^{r_{jl}}\sum_{t=1}^{m}(\bar{x}_{kt}^{(jl)}-\bar{x}_t^{(jl)})^2} \tag{2}$$

其有效性函数定义为:$f(M_l)=1-\min\{\alpha\mid P(F_l\geq F_\alpha(l-1,n-l))=\alpha\}$,其中 $F_\alpha(l-1,n-l)$ 是显著性水平为 α、自由度为 $l-1$,$n-l$ 的 F 检验的临界值。

3 基于生物智能的增量聚类算法

研究者从生物的智能行为中受到启发,建立了各种模型,包括人工神经网络(Artificial Neural Network-ANN)、人工免疫系统(Artificial Immune System-AIS)、遗传算法(Genetic Algorithm-GA)、蚁群算法(Ant Colony Algorithm)、粒子群优化(Particle Swarm Optimization-PSO)等。基于神经网络的聚类方法,已经归结为传统聚类方法中的基于模型的聚类方法中,研究的已经相对比较成熟,这里主要讲述其他的用于增量聚类的生物智能方法。

基于遗传算法的聚类自提出经常和传统聚类方法相结合,来克服传统聚类方法对初始值的选择比较敏感,避免陷入局部极小点或者来优化目标函数[46],遗传算法和 k-means、EM[47]等的混合算法,利用遗传算法优化 EM 的参数和初始点,取得了较好的效果。将遗传算法引入到传统聚类算法中,成果丰硕,在一定程度上克服了传统聚类算法的缺点。

粒子群算法在聚类分析中的应用与遗传算法类似,基本上都是与传统聚类方法结合,来克服传统聚类算法的缺陷,如刘靖明[48]提出的 PSO 与 k-means的混合算法,文献[49]提出了一种基于 PSO 和 k-harmonic means[50]的混合算法,于是 Yang等将 PSO 算法引入其中克服此缺点。Kao[51]等人也提出了三种算法结合的混合算法来聚类数据,收敛速度好于 PSO、NM-Psoriasis和 K-PSO。在增量聚类方面,粒子群算法也取得一定的成果。Bo Liu等[52]和 Chen Zhuo等[53]在原有粒子群聚类算法的基础上,对带新增数据点的 Agent进行判断,如果其信息素大于某一特定值,则将其移向子空间区的一个空

置区,否则将其移向没有被别的 Agent占用的随机选择的区域。如果 Agent不带新增数据点,则判断其是否信息素的是否小于特定值,如果小于就将其移向最近的新增数据点,否则将其移向没有被别的 Agent占用的随机选择的区域。随后在对信息素进行调整。

基于蚁群的聚类方法 1991年由 Deneubour提出[54], E. Lumer和 B. Faieta将该模型应用到了数据分析领域[55]。基于蚁群算法的聚类方法从原理上可以分为四种; 1) 运用蚂蚁觅食的原理,利用信息素来实现聚类[56]; 2) 利用蚂蚁自我聚集行为聚类; 3) 基于蚂蚁堆的形成原理实现数据聚类; 4) 运用蚁巢分类模型, 利用蚂蚁化学识别系统进行聚类的[57]。蚁群算法在增量聚类方面存在一些应用,文献[58, 59]使用基于蚁群聚类的增量式 Web用户聚类算法,其思想是在原有算法的基础上,判断新用户是否已在聚类算法中,如果在,则由新用户替换原有用户信息,表达用户兴趣的变迁,类标识不变。如果新用户不在已聚类的用户中,则调用基于方向相似性的蚁群聚类算法进行聚类。为了避免簇过于庞大,每次增量聚类后监测簇的误差大于给定值,解体簇并释放其对象,再进行蚁群聚类,直到没有新的簇解题和新数据到达位置。

基于人工免疫系统 (AIS)的聚类算法中最著名的是 Timmis等提出的资源受限人工免疫系统 (Resource limited AIS, RLAIS)[60]和 de Castro等构造的进化人工免疫网络 (aiNet)[61],他们都是抽取了免疫网络隐喻实现了对静态数值数据的聚类,显示出 AIS在数据分析方面的潜能。在基于 AIS的增量聚类方面,也取得了一定的成果。文献[62]中提出的 AIS框架及其框架上的一个自组织的增量聚类算法,利用 Logistic混沌序列生成初始抗体种群,利用其多样性识别新增的不属于任何已知簇的数据,该过程模拟了初次免疫应答。同时,初次免疫应答形成的记忆抗体可用于二次免疫应答, 即识别新增的属于已知簇的数据。为了减少数据冗余,算法用中心点和代表点表示已知簇并动态更新其识别区域,这样算法不但能动态、自组织地形成聚类,而且实现了数据特征的提取。

4 面向数据流的增量聚类算法

由于数据流的连续不断数据到达特点,很多人都将对数据流的聚类问题看作增量聚类的问题。因为数据流连续不断的特点,对算法的处理的效率要求较高,需要针对新数据的不断流入,动态地调整和更新聚类的结果,以此真实反应数据流的聚类形态[63]。由于内存的限制,只能考虑对数据流进行单遍扫描或有限次的扫描[64]。相对来说用户对最近一段时间的数据更感兴趣,而不是对所有数据都有同样的兴趣,鉴于这一思想,多种数据倾斜技术被应用于数据流[65—68]。现有文献中很多都将常规的聚类方法应用于数据流的聚类中,例如基于划分的数据流聚类[66, 69—73],基于层次的数据流聚类[65, 66, 74],基于网格的数据流聚类[66, 68, 75],基于密度的数据流聚类[66, 76],基于模型的数据流聚类[63, 66],基于回归的数据流聚类[66, 77]等。此外,还有针对特殊数据流进行聚类的算法,如面向 XML数据流的聚类算法[78, 79], Web流数据聚类[80],基于 Web Service的多数据流聚类[81],具有增量挖掘功能的 Web点击流聚类算法[82]。

5 结束语

本文对现有增量聚类算法进行了概述,列举了现有传统增量聚类算法及增量聚类算法的关键技术问题,对基于生物智能的增量聚类方法和数据流聚类进行分析。聚类方法多种多样,无论哪一种方法都不是放之四海而皆准的。于是,研究者们就考虑将这些聚类算法相结合,取长补短。考虑分阶段、分层次的增量聚类,也就是将聚类的对象进行加细,或者可以将同一时刻不同的增量聚类的结果进行融合来考虑,可以大大提高聚类的精度。尤其是在基于生物智能的聚类算法方面,这种结合尤其显著,将会是聚类算法进一步发展的方向之一。

另外,一些面向特定领域的增量聚类方法将是

http://www.cnki.net

未来一段时间的研究方向,比如数据流增量聚类问题,即将数据流的连续实时到达的特定用时间窗的方式出来,进而用增量聚类的方法来实现。此外就是复杂网络聚类、web聚类或者文本聚类的增量问题。

参 考 文 献

1 Jing L, Ng M K, Huang J Z. An entropy weighting *k*-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Transactions on Knowledge and Data Engineering, 2007; 19(8): 1026—1041

2 Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. ACM SIGMOD International Conference on Management of Data. Montreal: ACM Press, 1996: 103—114

3 Agrawal R, Gehrke J, Gunopulos D, *et al*. Automatic subspace clustering of high dimensional data for data mining applications. ACM SIGMOD International Conference on Management of Data. Seattle: ACM Press, 1998: 94—105

4 Cheng C H, Fu A W, Zhang Y. Entropy-based subspace clustering for mining numerical data. The Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego: ACM Press, 1999: 84—93

5 Aggarwal C, Procopiuc C, Wolf J L, *et al*. Fast algorithms for projected clustering. The fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego: ACM Press, 1999: 61—72

6 Aggarwal C C, Yu P S. Finding generalized projected clusters in high dimensional spaces. ACM SIGMOD international conference on management of data. Dallas: ACM Press, 2000: 70—81

7 Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for clustering large databases. ACM SIGMOD International Conference on Management of Data. Seattle: ACM Press, 1998: 73—84

8 Hsu C C, Huang Y. Incremental clustering of mixed data based on distance hierarchy. Expert Systems with Applications, 2008; 35(3): 1177—1185

9 Jain A, Dubes R. Algorithms for clustering data. Englewood Cliffs, N J: Prentice Hall College Div, 1988

10 Somlo G L, Howe A E. Incremental clustering for profile maintenance in information gathering web agents. The Fifth International Conference on Autonomous Agents, New York: ACM Press, 2001: 262—269

11 Hartigan J A. Clustering algorithms. New York: John Wiley & Sons, Inc, 1975

12 Carpenter G, Grossberg S. Art3: hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. Neural Networks, 1990; 3(2): 129—152

13 Fisher D. Knowledge acquisition *via* incremental conceptual clustering. Machine Learning, 1987; 2: 139—172

14 Can F. Incremental clustering for dynamic information processing. ACM Transaction for Information Systems, 1993; 11: 143—164

15 Can F, Fox E A, Snavely C D, *et al*. Incremental clustering for very large document databases: initial MARIAN experience. Information Systems, 1995; 84: 101—114

16 Lin J, Vlachos M, Keogh E J, *et al*. Iterative incremental clustering of time series. Lecture Notes in Computer Science, 2004: 106—122

17 Charikar M, Chekuri C, Feder T, *et al*. Incremental clustering and dynamic information retrieval. The Twenty-ninth Annual ACM Symposium on Theory of Computing. El Paso: ACM Press, 1997: 626—635

18 Ester M, Kriegel H P, Sander J, *et al*. Incremental clustering for mining in a data warehousing environment. The 24rd International Conference on Very Large Data Bases. NY: Morgan Kaufmann, 1998: 323—333

19 Simovici D, Singla N, Kuperberg M. Metric incremental clustering of nominal data. The 4th IEEE International Conference on Data Mining. Brighton: IEEE Computer Society Press, 2004: 523—526

20 Ester M, Kriegel H P, Sander J, *et al*. Incremental clustering for mining in a data warehousing environment. The 24rd International Conference on Very Large Data Bases. NY: Morgan Kaufmann, 1998: 323—333

21 陈 宁,陈 安,周龙骧. 基于密度的增量式网格聚类算法. 软件学报, 2002; 13(1): 1—7

22 黄永平,邹力鹍. 数据仓库中基于密度的批量增量聚类算法. 计算机工程与应用, 2004; 29: 206—208

23 徐新华,谢永红. 增量聚类综述及增量 DBSCAN 聚类算法研究. 华北航天工业学院学报, 2006; 16(2): 15—17

24 刘建畔,李 芳. 一种基于密度的高性能增量聚类算法. 计算机工程. 2006; 32(21): 76—78

25 Chen C, Hwang S, Oyang Y. An incremental hierarchical data clustering algorithm based on gravity theory. In: Proceedings of the Sixth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin: Springer, 2002; 2 336: 237—250

26 Widyantoro D H, Ioerger T R, Yen J. An incremental approach to building a cluster hierarchy. In: Proceedings of the 2002 IEEE International Conference on Data Mining. New York: IEEE, 2002: 705—708

27 Charikar M, Chekuri C, Feder T, *et al*. Incremental clustering and

http://www.ixueshu.com

dynamic information retrieval. In: Proceedings of the Twenty-ninth Annual ACM Symposium on the Theory of Computing New York: ACM, 1997: 626—635

28 Hsu C C, Huang Y. Incremental clustering of mixed data based on distance hierarchy. Expert Systems with Applications, 2008; 35 (3): 1177—1185

29 He J, Lan M, Tan C L, *et al*. Initialization of cluster refinement algorithms: a review and comparative study. In: Proceedings of International Joint Conference on Neural Networks. Buda Pest, Hungary, 2004: 297—302

30 李新良. 数据挖掘中聚类初始化方法的优化研究, 计算技术与自动化, 2008; 27(2): 130—133

31 张 霞,王素珍,尹怡欣,等. 基于模糊力度计算的 K-means文本聚类算法研究. 计算机科学, 2010; 37(2): 209—211

32 田生文,王伊蕾,李阿丽. 一种应用复杂网络特征的 K-means初始化方法. 计算机工程与应用, 2010; 46(6): 127—129

33 杨圣云,袁德辉,赖国明. 一种新的聚类初始化方法. 计算机应用与软件, 2007; 24(8): 50—52

34 钱 线,黄萱菁,吴立德. 初始化 K-means的谱方法. 自动化学报, 2007; 33(4): 342—346

35 申晓勇,雷英杰,蔡 茹,等. 一种基于密度函数的直觉模糊聚类初始化方法. 计算机科学, 2009; 36(5): 197—199

36 马秀丽,焦李成. 联合模型初始化独立谱聚类算法. 西安电子科技大学学报 (自然科学版), 2007; 34(5): 768—772

37 盛 莉,邹开其,邓冠男. 基于网格和密度的模糊 c均值聚类初始化方法. 计算机应用与软件, 2008; 25(3): 22—24

38 Lughofer E. Extensions of vector quantization for incremental clustering. Pattern Recognition, 2008; 41: 995— 1011

39 吴 琪,左万利. 一种基于距离的增量聚类算法. 湖南工程学院学报, 2005; 15(3): 41—44

40 高新波. 模糊聚类分析及其应用. 西安: 西安电子科技大学出版社, 2004

41 唐明会,杨 燕. 模糊聚类有效性的研究进展. 计算机工程与科学, 2009; 31(9): 122—124

42 Rhee H. A validity measure for fuzzy clustering and its use in selecting optimal number of clusters. In: Proc of the 5th IEEE Int'l Conf on Fuzzy System, 1996: 1020—1025

43 Kwon S H. Cluster validity index for fuzzy clustering. Electronics Letters, 1998; 34(22): 2176—2177

44 彭 勇,吴友情. 一种新的聚类有效性函数. 计算机工程与应用, 2010; 46(6): 124—126

45 安中华,安 琼. 模糊聚类的有效性研究. 湖北大学学报, 2006; 28(3): 222—226

46 李 洁,高新波,焦李成. 一种基于 GA的混合属性特征大数据集聚类算法. 电子与信息学报, 2004; 26(8): 1203—1209

47 Lauritzen S L. The EM algorithm for graphical association models with missing data. Computational Statistics and Data Analysis, 1995; 19(2): 191—201

48 刘靖明,韩丽川,候立文. 一种新的聚类算法——粒子群聚类算法. 计算机工程与应用, 2005; 41(20): 183—185

49 Yang F Q, Sun T L, Zhang C H. An efficient hybrid data clustering method based on K-harmonic means and partical swarm optimization. Expert Systems with Applications, 2009; 36(6): 9847—9852

50 Hammerly G, Elkan C. Alternatives to the *k*-means algorithm that find better clusterings. In: Proceedings of the 11th International Conference on Information and Knowledge Management. New York: ACM Press, 2002: 600—607

51 Kao Y T, Zahara E, Kao IW. A hybirdized approach to data clustering. Expert Systems with Applications, 2008; 34 (3): 1754—1762

52 Liu Bo, Pan Jiuhui, McKay R I (Bob). Incremental clustering based on swarm intelligence. In: Proceedings of Simulated Evolution and Learning-6th International Conference, 2006: 189—196

53 Chen Zhuo, Meng Qingchun. An incremental clustering algorithm based on swarm intelligence theory. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics, 2004; 3: 1768—1772

54 Deneubourg J L, Goss S, Franks N, *et al*. The dynamics of collective sorting: Robot- like ants and ant- like robots. In: Meyer J A, Wilson S, eds. Proceedings of the First international Conference on Simulation of Adaptive haviour, From Animals to Animals J, MIT Press, Cambridge MA, 1991: 356—365

55 Bonabeau E, Dorigo M, Theraulaz G. Swarm intelligence-from natural to artificial system. New York, NY: Oxford University Press, 1999

56 杨新斌,孙京诰,黄 道. 一种进化聚类学习新方法. 计算机工程与应用, 2003; 39(15): 60—62

57 张建华,江 贺,张宪超. 蚁群聚类算法综述, 计算机工程与应用, 2006; 16: 171—174

58 沈 洁,林 颖,陈志敏,等. 基于增量式蚁群聚类的用户访问模式挖掘. 计算机应用, 2005; 25(7): 1654—1657

59 张 斌,苏一丹,曹 波. 基于蚁群聚类模型的增量式 Web用户聚类. 微计算机信息, 2008; 24(5—3): 231—233

60 Timmis J, Neal M. A resource limited artificial immune system for data analysis. Knowledge-Based Systems, 2001; 14 (34): 121—130

61 Castro L N, Von Zuben F J. An evolutionary immune network for data clustering. In: Proceedings of the 6th Brazilian Symposium on Neural Networks. Rio de Janeiro, Brazil: IEEE, 2000: 84—89

62 李向华,王钲旋,吕天阳,等. 基于混沌和免疫应答的增量聚类

http://www.ixueshu.com

新算法,自动化学报,2010; 32(2): 208—214
63 张晓龙,曾 伟. 实时数据流聚类的研究新进展. 计算机工程与设计, 2009; 30(9): 2177—2181
64 金澈清,钱卫宁,周傲英. 流数据分析与管理综述. 软件学报, 2004; 15(8): 1172—1181
65 Aggarwal C C, Han J W, Wang J Y, *et al*. A framework for projected clustering of high dimensional data streams. Proceedings of the 30th VLDB Conference. Toronto: VLDB Endowment, 2004: 852—863
66 Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. Kaufmann M: Elsevier Inc, 2006: 467—589
67 Aggarwal C C, Han J W, Wang J Y, *et al*. A framework for clustering evolving data streams. Proceedings of the 29th VLDB Conference. Berlin: VLDB Endowment, 2003: 81—92
68 Chen Y X, Tu L. Density-based clustering for real-time stream data. Proceedings of the 13th ACM SIGKDD international conference on Knowledge Discovery and Data Mining. California: ACM, 2007: 133—142
69 Guha S, Mishra N, Motwani R, *et al*. Clustering data streams. The 41st Annual Symp. on Foundations of Computer Science, FOCS 2000. Redondo Beach: IEEE Computer Society, 2000: 359—366
70 Guha S, Meyerson A, Mishra N, *et al*. Clustering data streams: theory and practice. IEEE Transactions on Knowledge and Data Engineering, 2003; 15(3): 515—528
71 Babcock B, Datar M, Motwani R, *et al*. Maintaining variance and *k*-medians over data streams windows. Proceedings of the 22nd Symposium on Principles of Database Systems, 2003: 234—243
72 Charikar M, O'Callaghan L, Panigrahy R. Better streaming algorithms for clustering problems. In: Proc of 35th ACM Symposium on Theory of Computing, 2003: 30—39
73 O'Callaghan L, Mishra N, Meyerson A, *et al*. Streaming-data algorithms for high quality clustering. Proceedings of IEEE International Conference on Data Engineering, 2002, 685—694
74 Udommanetanakit K, Rakthanmanon T, Waiyamai K. E-stream: evolution-based technique for stream clustering. Springer-Verlag Berlin Heidelberg, 2007: 605—615
75 Bhatnagar V, Kaur S. Exclusive and complete clustering of streams. Springer-Verlag Berlin Heidelberg, 2007: 629—638
76 Cao F, Ester M, Qian W, *et al*. Density-based clustering over an evolving data stream with noise. Proceedings of the SIAM Conference on Data Ming, 2006: 328—339
77 Motoyoshi M, Miura T, Shioya I. Clustering stream data by regression analysis. The Australasian Workshop on Data Mining and Web Intelligence (DMWI2004), Dunedin, New Zealand, 2004: 115—120
78 姚文集,高明霞,毛国君,等. 基于滑动窗口的 XML 数据流聚类算法. 计算机工程, 2010; 36(13): 87—89
79 常建龙,曹 锋,周傲英. 基于滑动窗口的进化数据流聚类. 软件学报, 2007; 18(4): 905—918
80 彭 源. Web流数据聚类挖掘技术研究. 电脑知识与技术, 2010; 6(4): 935—936
81 邹凌君,高开周. 基于 Web Service的多数据流聚类研究. 广西轻工业, 2009; 11: 85—87
82 李晓明,夏秀峰,张 斌. 一种具有增量挖掘功能的 Web点击流聚类算法. 沈阳大学学报, 2010; 22(3): 8—10

http://www.ixueshu.com

Survey of Incremental Clustering Algorithms

LI Tao-ying, CHEN Yan, QIN Sheng-jun, LI Nan
(Transportation Management College, Dalian Maritime University, Dalian 116026, P. R. China)

[Abstract] The concepts of incremental clustering is interpreted. Incremental clustering could be used to solve the demands of data changing and large storage space, correct incremental clustering could promoted the accuracy and efficiency of clustering when data was changing. Methods of incremental clustering based on traditional clustering, that based on biological intelligence and stream clustering are studied and their development process and characteristics are described, and their key technologies are analyzed. Finally, the direction of further work of incremental clustering is given.

[Key words] clustering analysis incremental clustering biological intelligence data stream

word版下载：http://www.ixueshu.com
免费论文查重：http://www.paperyy.com
3亿免费文献下载：http://www.ixueshu.com
超值论文自动降重：http://www.paperyy.com/reduce_repetition
PPT免费模版下载：http://ppt.ixueshu.com