文章编号: 1001-0920(2018)03-0471-08 DOI: 10.13195/j.kzyjc.2017.0049

基于局部分布的贝叶斯自适应共振理论增量聚类算法

王 玲†, 孟建瑶

(1. 北京科技大学 自动化学院，北京 100083；2. 北京科技大学
工业过程知识自动化教育部重点实验室，北京 100083)

摘 要: 针对传统的贝叶斯增量聚类算法需要人为设置参数,且对分布不均衡数据聚类效果不佳的问题,提出一种基于局部分布的贝叶斯自适应共振理论增量聚类算法. 首先,利用数据快照读取数据;然后,在无需设置参数的情况下,考虑类簇的局部分布情况,自适应地确定新数据的所属类别,并更新获胜类簇;最后,确定相邻快照中类簇的演化关系. 不同数据集的仿真结果表明,所提出的算法在准确性和自适应性方面均有显著提高.

关键词: 增量聚类算法；贝叶斯；自适应共振理论；不均衡数据

中图分类号: TP273 **文献标志码:** A

Incremental clustering algorithm of Bayesian adaptive resonance theory based on local distribution

WANG Ling†, MENG Jian-yao

(1. School of Automation，University of Science and Technology Beijing，Beijing 100083，China；2. Key Laboratory of Knowledge Automation for Industrial Processes，Ministry of Education，University of Science and Technology Beijing，Beijing 100083，China)

Abstract: Traditional incremental clustering algorithm needs to be set parameters and cannot deal with the imbalance data.To solve the problem, the incremental clustering algorithm of bayesian adaptive resonance theory based on local distribution is proposed. Firstly, the new data are collected by data snapshots. Then, in current data snapshot, the new data are clustered into the winning cluster adaptively according to the local distribution of the clusters without predefined parameters. Then, the evolving relationships between the clusters in two neighboring data snapshots are determined. Finally, the simulation result shows that the proposed algorithm can improve the accuracy and the adaptability.

Keywords: incremental clustering algorithm；Bayesian；adaptive resonance theory；imbalance data

0 引 言

聚类分析是数据挖掘领域中重要的技术之一,用于将数据对象分成多个类簇,同时保证同一个类簇内的对象相似性最大,类簇之间的对象具有较高的差异性[1-2]. 数据分布随着时间的推移不断变化[3],因此人们在已有典型聚类算法的基础上,提出了各种各样的增量聚类算法来发现动态数据中的演化结构. 例如,基于密度的增量聚类算法[4]、基于距离的增量聚类算法[5]、基于网格的增量聚类算法[6]、基于模型的增量聚类算法[7]等. 其中,基于模型的增量聚类算法以概率为基础,具有很好的灵活性,得到了广泛的应用[8-9]. 文献[8]提出了一种基于模型的增量聚类算法,以自组织映射为模型基础,对动态数据进行聚类,但此算法需要预设速度因子来控制模型的初始化;文献[9]结合了EM算法和概率PCA混合模型,以实现增量聚类算法,然而此算法需要不断迭代获取聚类参数,算法效率低. 上述两种聚类算法的聚类结果都缺乏可解释性,对此,文献[10]提出了一种基于贝叶斯的增量聚类算法,利用贝叶斯理论计算模型中的概率,使得算法具有良好的可解释性. 增量高斯混合模型(IGMM)算法[11]是一种典型的基于贝叶斯的增量聚类算法,可以增量地根据新数据确定任意形状的类簇,但是初始类簇个数的设置较为困难. 文献[12]

收稿日期: 2017-01-13；**修回日期:** 2017-06-09.
基金项目: 国家自然科学基金项目(61572073)；北京科技大学研究生教育发展基金项目(230201506400060).
责任编委: 侯忠生.
作者简介: 王玲(1974–), 女, 副教授, 博士, 从事数据挖掘、机器学习的研究；孟建瑶(1992–), 女, 硕士生, 从事数据挖掘的研究.
†通讯作者. E-mail: lingwang@ustb.edu.cn

提出了一种改进的基于GMM的增量聚类算法,但是仍然没有彻底解决由于迭代寻优导致算法效率低的问题;文献[13]采用了贝叶斯聚类算法结合增量的分散-聚合策略对新数据进行聚类,然而该算法对学习参数敏感,且由于分散-聚合策略不完善,导致冗余类簇的产生;文献[14]提出了一种基于无参数贝叶斯方法的增量聚类算法,该算法根据数据自适应地确定类簇个数,然而该算法不能对分布复杂的数据进行准确聚类;文献[15]利用两种变量学习方法对贝叶斯增量聚类中的参数进行学习,但是初始化参数直接影响最终的聚类效果.

为了避免不断迭代寻优引起聚类效率下降以及人为设置聚类个数的困难,文献[16]引入了贝叶斯自适应共振理论(BART)增量聚类算法. BART算法继续沿用GMM算法中的类簇结构,根据当前输入数据不断调整获胜类簇的均值向量和协方差矩阵,能够发现任意形状的类簇. 此算法主要包括3个阶段:类簇选择、匹配测试和更新学习. BART算法无需设置初始值,结合贝叶斯理论,实现了增量聚类,提高了算法效率[17-18]. 然而,BART算法需要人为确定警戒参数来控制是否建立新的类簇,警戒参数的大小直接影响聚类结果的好坏. 文献[17]将BART算法与卡尔曼滤波算法相结合,对新类簇的建立和获胜类簇的参数更新进行改进,虽然优化了获胜或者新建类簇的参数,但是增加了时间复杂度,且没有解决人为设置参数的问题. 文献[19]对警戒参数的设置进行了调整,根据各个类簇的协方差矩阵调整警戒参数,在一定程度上提高了算法的自适应性,但仍需人为设置警戒系数.

鉴于此,本文在BART算法的基础上,考虑类簇的局部分布情况,对匹配测试阶段进行调整,给出一种基于局部分布的贝叶斯自适应共振理论增量聚类算法(ILocal-BART). ILocal-BART算法在保留BART算法优点的同时,对BART 算法中匹配测试阶段进行调整,同时考虑获胜类簇的选择函数和加入新输入导致类簇局部分布的不平衡程度,使得两者达到均衡,选择的类簇更合理,提高了聚类结果的有效性; 根据类簇内的数据分布自动确定匹配测试阶段的阈值,提高了算法的自适应性. 为了发现相邻快照中类簇的演化关系,通过计算类簇间的演化关系定量地分析了类簇间的动态变化.

1 BART算法

BART算法是一种无监督自适应的增量聚类算法,每个类簇用多元高斯混合分布的密度函数表示,多元高斯混合分布的密度函数包括均值向量、协方差矩阵和先验概率3个参数,能够发现任意形状的类簇. 这3个参数反映了类簇中心、类簇形状以及与其他类簇对比的优势,根据新数据不断地对类簇的3个参数进行调整. 该算法主要包括3个部分:类簇选择,匹配测试和学习更新.

1) 类簇选择:为了将新数据划分到合适的类簇中,选择新数据相对所有已存在类簇的后验概率最大值,最大后验概率对应的类簇作为获胜类簇.

2) 匹配测试:如果获胜类簇的样本点容量小于警戒阈值,则执行学习更新步骤;否则,寻找下一个获胜类簇. 如果所有类簇的样本点容量都大于警戒阈值,则建立新的类簇,新类簇的均值向量为输入本身,协方差矩阵是一个极小值.

3) 学习更新:本阶段根据输入本身调整获胜类簇的均值向量和协方差矩阵.

2 基于局部分布的贝叶斯自适应共振理论增量聚类算法

2.1 数据切片处理

数据快照是保留某一时刻数据影像的技术,保留的影像称为快照[20]. 数据快照将动态数据分割成若干快照,每个快照的大小由在线数据的速度和数据快照应用程序的读取速度有关. 随着信息不断地采集输入,数据快照的任务是以有效的方式管理输入数据,存储信息. 基本上,聚类过程的快照性能是通过利用历史记录来减少需要的硬盘空间,从而提高处理能力. 在增量聚类过程中,这些快照将被收集,并用于构造特定时间点的聚类历史. 相邻快照记录将被用来进行演化过程中类簇之间的对等比较. 一旦某个类簇失效或者被替代,其相关的快照记录将改变状态,并用作相似性的识别和融合的参考. 如图1所示,$\boldsymbol{S}[1],\boldsymbol{S}[2],\cdots,\boldsymbol{S}[q]$是不同时刻的快照,假设第1个快照$\boldsymbol{S}[1]$是$t$时刻获取的快照,第2个快照$\boldsymbol{S}[2]$是$t+r$时刻获取的快照,第$q$个快照$\boldsymbol{S}[q]$是$t+(q-1)r$时刻获取的快照,$r$是相邻快照之间的时间间隔,快照内类簇的位置或个数都随着快照内数据的变化而变化.

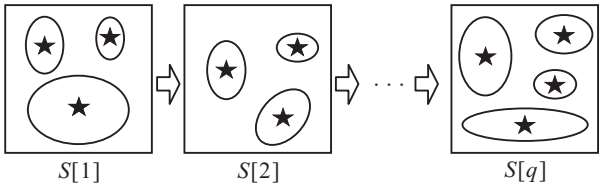


图1 快照的聚类过程

2.2 算法提出

通过对BART算法的描述可知,该算法只能对数据进行逐条处理,且在对新数据进行聚类时,根据各

个类簇的先验概率和新数据相对于各个类簇的后验概率,共同决定新数据的所属类簇. 针对数据分布不均衡的问题,BART算法中含有大量数据的类簇更容易在类簇选择阶段中获胜,最后的聚类结果不能准确地体现数据的分布,忽视了数据局部分布对聚类的影响. 同时,警戒阈值设置不精确将直接影响新数据所属类簇的选择,无法保证聚类结果的正确性. 如图2所示,点代表类簇中已有的数据,三角形代表新数据,类簇C_1中包含的数据较多,类簇C_2中包含的数据较少. 可以看出,类簇C_1的先验概率比类簇C_2的先验概率大,新数据相对于类簇C_1与类簇C_2的后验概率相差较小. 根据BART算法中的类簇选择原理,最终的获胜类簇是类簇C_1. 然而,新数据加入类簇C_1中会导致类簇C_1的局部分布混乱. 如果警戒阈值设置较大,则类簇C_1中包含的数据会越来越多,类簇C_1的局部分布会越来越混乱.

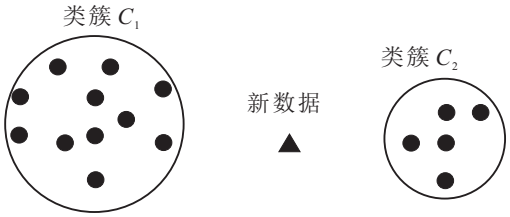


图2　新数据聚类

为了在无需设置任何参数的情况下确保对分布不均衡数据聚类的准确性,本文首先利用数据快照在线获取数据,然后在BART算法的基础上加入局部分布的概念,提出一种基于局部分布的贝叶斯自适应共振理论增量聚类算法. ILocal-BART 算法根据先验概率和后验概率获取获胜类簇之后,再测试新数据的加入是否导致获胜类簇J的局部分布不平衡程度加大. 类簇的局部分布不平衡程度用协方差行列式$|\Sigma_k|(k=1,2,\cdots,|C|)$表示,值越大,类簇的不平衡程度越大. 本文中使用的协方差矩阵Σ_k是对角线矩阵,形如

$$\Sigma_k=\begin{bmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma_L^2\end{bmatrix}, \tag{1}$$

其中$\sigma_j^2(1\leqslant j\leqslant L)$是属于第$k$个类簇的所有数据的第$j$个维度的方差,$L$是数据集的维度.

在第q个快照中的第i条数据$\boldsymbol{x}_i[q]$加入获胜类簇J之后,根据类簇协方差行列式的变化,判断数据$\boldsymbol{x}_i[q]$的加入是否会导致获胜类簇J的局部分布不平衡程度变大,计算公式为

$$d(i)=N_k[q]\times\log(|\Sigma_J[q]|)+\log(|\Sigma(\boldsymbol{x}_i[q])|)-(N_k[q]+1)\times\log(|\Sigma_{\rm com}(i)|). \tag{2}$$

其中

$$\Sigma_{\rm com}(i)=\frac{1}{N_J[q]+1}\times\Big\{N_J[q]\times\Sigma_J[q]+\Sigma(\boldsymbol{x}_i[q])+\frac{N_J[q]\times(\mu_J[q]-\boldsymbol{x}_i[q])\times(\mu_J[q]-\boldsymbol{x}_i[q])^{\rm T}}{N_J[q]+1}\Big\}, \tag{3}$$

$$\Sigma(\boldsymbol{x}_i[q])=\frac{\rm mean(diag(\mathit{D}))}{10^5}\times I. \tag{4}$$

$\Sigma_{\rm com}(i)$是数据$\boldsymbol{x}_i[q]$合并到获胜类簇J后的合并协方差矩阵,$\Sigma(\boldsymbol{x}_i[q])$是假设数据$\boldsymbol{x}_i[q]$单独为一个类簇时的协方差矩阵,$\mu_J[q]$是第$q$个快照中获胜类簇$J$的均值向量,$\Sigma_J[q]$是第$q$个快照中获胜类簇$J$的协方差矩阵,$N_J[q]$是第$q$个快照中获胜类簇$J$中数据的个数,diag(·)是向量对角化函数,mean(·)是求均值函数,I是单位矩阵.

为了平衡获胜类簇的选择函数与类簇的局部分布,既不能直接选取选择函数值最大的作为最终的获胜类簇,也不能直接选取协方差行列式变化量$d(i)$最小的作为最终的获胜类簇. 因此,将协方差行列式变化量$d(i)$与合并阈值$d_\theta(i)$对比,确定对于数据$\boldsymbol{x}_i[q]$更加合适的类簇,合并阈值$d_\theta(i)$的计算公式为

$$d_\theta(i)=\log(|\Sigma(\boldsymbol{x}_i[q])|)-\frac{1}{(N_J[q]-1)}\times\sum_{i=1}^{N_J[q]-1}\log(|\Sigma(\boldsymbol{x}_{C_J}^1[q]-\boldsymbol{x}_{C_J}^m[q]|), \tag{5}$$

$$\Sigma(\boldsymbol{x}_{C_J}^1[q]-\boldsymbol{x}_{C_J}^m[q])=\frac{(\boldsymbol{x}_{C_J}^1[q]-\boldsymbol{x}_{C_J}^m[q])^2}{4}. \tag{6}$$

其中:$\boldsymbol{x}_{C_J}^1[q]$是第$q$个快照中获胜类簇$J$的第1条数据,$\boldsymbol{x}_{C_J}^m[q](m=2,3,\cdots,N_J[q])$是第$q$个快照中获胜类簇$J$中第$m$条数据,$N_J[q]$是第$q$个快照中获胜类簇$J$中数据的个数.

2.3　基于局部分布的贝叶斯自适应共振理论增量聚类算法步骤

为了自适应地得到数据对象的聚类结果,本文算法在数据快照获取数据的基础上,能够在无需定义任何参数的条件下保证聚类结果的有效性,并且得到相邻快照中类簇的演化规律. 首先利用数据快照在线获取数据,根据选择函数为快照中的每条数据选取获胜类别;然后利用历史快照已有的聚类类别,对当前数据快照中的每条数据进行匹配测试,如果匹配成功,则获胜类别根据现有数据进行更新,否则建立新的聚类;最后分析相邻快照中聚类之间的演化关系.

ILocal-BART算法的详细执行步骤如下.

Step 1　利用数据快照获取数据.

Step 2 初始化. 令$N=1,q=1$,将第q个快照中的第1条数据$\boldsymbol{x}_1[q]$作为该快照中第1个类簇$C_1[q]$的初始化均值向量$\mu_1[q]$,初始化协方差矩阵$\Sigma_1[q]$为1个与输入同维度的对角线矩阵,初始化公式为

$$\mu_1[q]=\boldsymbol{x}_1[q], \tag{7}$$

$$\Sigma_1[q]=\frac{\mathrm{mean}(\mathrm{diag}(D))}{10^5}\times I, \tag{8}$$

其中已输入数据集为$D=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]$.

Step 3 对每个快照内的数据进行增量聚类. 若$q\neq 1$, $i=1$,则首先继承第$q-1$个快照的聚类结果,然后执行Step 3.1;否则,直接执行Step 3.1.

Step 3.1 选择函数的计算. 将第q个快照中第i条数据$\boldsymbol{x}_i[q]$与已有的所有类簇进行匹配,在与第q个快照中第k个类簇$C_k[q]$进行匹配时,选择函数的计算公式如下所示:

$$M_k[q]=P(C_k[q]|\boldsymbol{x}_i[q])=\frac{P(\boldsymbol{x}_i[q]|C_k[q])P(C_k[q])}{\sum\limits_{l=1}^{|C_k[q]|}P(\boldsymbol{x}_i[q]|C_k[q])P(C_k[q])}. \tag{9}$$

其中:$|C[q]|$是第q个快照中已有的类簇个数;$P(C_k[q])$是第q个快照中类簇C_k的先验概率,$k=1,2,\cdots,|C[q]|$;$P(\boldsymbol{x}_i[q]|C_k[q])$是第$q$个快照中的第$i$条数据$\boldsymbol{x}_i[q]$选择第$q$个快照中的类簇$C_k[q]$为最合适类簇的概率,计算公式为

$$P(C_k[q])=N_k[q]/N, \tag{10}$$

$$P(\boldsymbol{x}_i[q]|C_k[q])=\frac{1}{(2\pi)^{L/2}|\Sigma_k[q]|^{1/2}}\times\exp\{-0.5(\boldsymbol{x}_i[q]-\mu_k[q])^{\mathrm{T}}\Sigma_k[q]^{-1}(\boldsymbol{x}_i[q]-\mu_k[q])\}. \tag{11}$$

$N_k[q]$是第q个快照中类簇$C_k[q]$中数据的个数,L是输入的维度,$\mu_k[q]$是第q个快照中类簇$C_k[q]$的均值向量,$\Sigma_k[q]$是第q个快照中类簇$C_k[q]$的协方差矩阵,N是已处理数据的个数.

Step 3.2 选择最大的选择函数值. 获胜类簇J有最大的选择函数值

$$J=\arg\min_{J}\min_{k=1}^{|C[q]|}(M_k[q]), \tag{12}$$

其中$|C[q]|$是第q个快照中已有的类簇个数.

Step 3.3 匹配测试. 计算数据$\boldsymbol{x}_i[q]$加入获胜类簇J后,协方差行列式的变化量,计算公式如式(2)所示. 为了平衡获胜类簇的选择函数和类簇的局部分布,引入合并阈值$d_\theta(i)$,计算公式如式(5)和(6)所示. 若$d(i)>d_\theta(i)$,则执行Step 3.4;否则,从剩下的类簇中寻找选择函数最大的类簇,继续执行匹配测试. 如果匹配成功,则执行Step 3.4;如果所有类簇都匹配不成功,则执行Step 3.5.

Step 3.4 更新已选择类簇J的均值向量和协方差矩阵:

$$\mu_{J,\mathrm{new}}[q]=\frac{1}{N_J[q]+1}\times\{N_J[q]\times\mu_{J,\mathrm{old}}[q]+\boldsymbol{x}_i[q]\}, \tag{13}$$

$$\Sigma_{J,\mathrm{new}}[q]=\frac{N_J[q]}{N_J[q]+1}\times\Sigma_{J,\mathrm{old}}[q]+\frac{1}{N_J[q]+1}\times(\boldsymbol{x}_i[q]-\mu_{J,\mathrm{new}}[q])\times(\boldsymbol{x}_i[q]-\mu_{J,\mathrm{new}}[q])^{\mathrm{T}}. \tag{14}$$

其中:$\mu_{J,\mathrm{old}}[q]$是更新之前第q个快照中获胜类簇J的均值向量,$\Sigma_{J,\mathrm{old}}[q]$是更新之前第$q$个快照中获胜类簇$J$的协方差矩阵,$\mu_{J,\mathrm{new}}[q]$是更新之后第$q$个快照中获胜类簇$J$的均值向量,$\Sigma_{J,\mathrm{new}}[q]$是更新之后第$q$个快照中获胜类簇$J$的协方差矩.

Step 3.5 增加新的类簇,新类簇的均值向量$\mu_{|C[q]|+1}[q]$和协方差矩阵$\Sigma_{|C[q]|+1}[q]$为

$$\mu_{|C[q]|+1}[q]=\boldsymbol{x}_i[q], \tag{15}$$

$$\Sigma_{|C[q]|+1}[q]=\frac{\mathrm{cov}(D)\times(N-1)}{N}+\frac{\mathrm{mean}(\mathrm{diag}(D))}{10^5}\times I, \tag{16}$$

其中$\mathrm{cov}(\cdot)$是协方差函数.

2.4 算法复杂度分析

ILocal-BART算法主要由类簇选择、匹配测试和学习更新3个阶段组成. 对于每条新数据,首先,在类簇选择阶段计算已有的$|C|$个选择函数M_k所需的计算量是$O(1)$. 然后,在匹配测试阶段计算获胜类簇与新数据之间的距离,同时更新阈值,经过$k(1\leqslant k\leqslant|C|)$次尝试之后,所需计算量是$O(k)$. 最后,在学习更新阶段,若在上一步匹配测试中存在匹配成功的类簇,则更新获胜类簇的均值向量和协方差矩阵,所需的计算量是$O(1)$;如果所有的类簇都没有匹配成功,则建立新的类簇,构建新类簇的均值向量和协方差矩阵所需的计算量是$O(1)$. 综上所述,假设数据快照中包含N条数据样本,该算法的复杂度是$O(N)+O(Nk)+O(N)$, $1\leqslant k\leqslant|C|$. 在最坏的情况下,该算法的复杂度是$O(N)+O(N|C|)+O(N)$. 因此,该算法的时间复杂度与样本个数密切相关.

2.5 演化分析

ILocal-BART算法对各个快照内的数据进行聚类,提高了聚类算法读取数据的效率,通过计算相邻快照中类簇间的演化关系,得到快照内部的演化规律. 如图3所示,$S[1],S[2],\cdots,S[q]$是不同时刻的快照,五角星代表不同类簇的中心. 在快照$S[1]$中,利用ILocal-BART算法对当前快照内的数据聚类,类簇

的个数根据数据的分布自动得到. 在快照$S[2]$中,根据当前快照中的数据,利用ILocal-BART算法对快照$S[1]$内的聚类结果进行更新或者建立新的类簇. 显然,随着快照内数据的不断变化,利用ILocal-BART算法迭代地得到的类簇的位置和个数也在不断发生变化. 每一个快照内的聚类结果直接影响它的下一个快照内的聚类结果,通过计算相邻快照中类簇之间的演化关系,定量地表示类簇间的动态变化关系,其中,演化关系越大,下一个快照中的类簇继承上一个快照中类簇的信息越多;反之,继承的信息越少. 根据第$q-1$个快照$S[q-1]$中第k个类簇与第q个快照$S[q]$中的部分数据,得到快照$S[q]$中的第p个类簇,类簇之间的演化关系$W_{kp}^{[q-1,q]}$可以根据下式计算:

$$w_{kp}^{[q-1,q]}=1-(\mu_p[q]-\mu_k[q-1])^{\mathrm{T}}\Sigma_k[q-1]^{-1}(\mu_p[q]-\mu_k[q-1])\Big/\Big[\sum_{j=1}^{|C[q-1]|}(\mu_p[q]-\mu_j[q-1])^{\mathrm{T}}\Sigma_j[q-1]^{-1}(\mu_p[q]-\mu_j[q-1])\Big]. \tag{17}$$

其中:$|C(q-1)|$是第$q-1$个快照中类簇的个数,$|C(q)|$是第q个快照中类簇的个数,$\mu_k[q-1]$是第$q-1$个快照中的第$k(1\leqslant k\leqslant|C[q-1]|)$个类簇的均值向量,$\mu_p[q]$是第$q$个快照的第$p(1\leqslant p\leqslant|C[q]|)$个类簇的均值向量,$\mu_j[q-1]$是第$q-1$个快照中的第$j(j=1,2,\cdots,k,\cdots,|C[q-1]|)$个类簇的均值向量,$\Sigma_j[q-1]$是第$q-1$个快照中的第$j$个类簇的协方差矩阵.

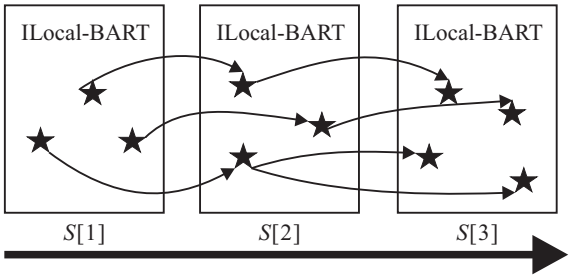


图 3　ILocal-BART 算法的类簇演化过程

3　仿真分析

3.1　实验安排和测试环境

将ILocal-BART算法和BART算法与IGMM对比,选取1个人工合成不均衡数据集、UCI数据库中的5个不均衡数据集[21]和2个不均衡数据流[22]作为实验数据进行分析,以测试ILocal-BART算法. 测试结果表明,在无需设置任何参数的情况下,所提出算法执行效率和自适应性能均得到提高. 各个数据集的不均衡分布如图4所示.

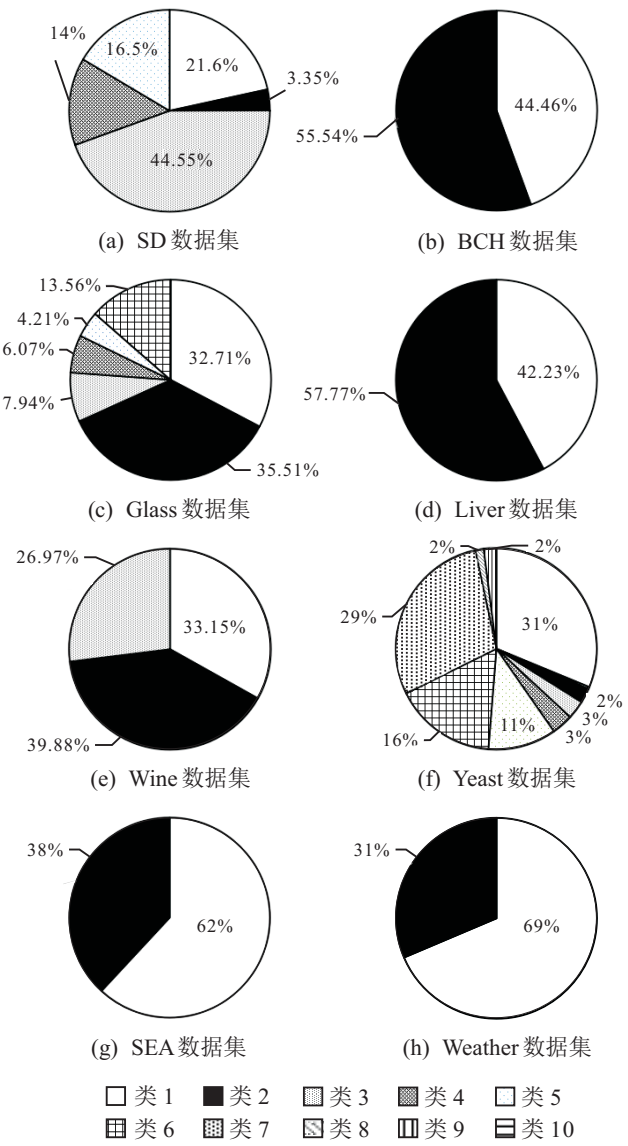


图 4　数据集不均衡分布

对于所有数据,各个维度均进行归一化处理,使数据集的所有数据值均在[0,1]区间. 所有实验均在Inter(R) CUP 2.30 GHz工作台上运行,利用Matlab 2014 a软件进行仿真. 数据集的详细信息如表1所示.

表 1　实验数据

数据集	样本数	聚类数	维度
SD	2 000	5	2
BCH	1 372	2	4
Glass	214	2	9
Liver	341	2	6
Wine	178	3	11
Yeast	1 484	10	8
SEA	50 000	2	3
Weather	18 159	2	8

3.2　评价指标

本文采用以下3种聚类评价指标对算法的聚类效果进行评价:聚类准确率(Acc)、Rand指数(RI)和平均平方距离和(MSSQ).

1) Acc用于统计在所有的数据中经过聚类算法正确划分数据个数占整个数据集的比例,定义公式为

$$\mathrm{Acc}=\sum_{k=1}^{|C|}a_k\Big/N. \tag{18}$$

其中: a_k是第k个类簇中算法的聚类结果与实际数据集分类情况相一致的数据个数, $|C|$是类簇个数, N是数据集中包含的数据个数.

2) RI用于度量聚类划分结果与数据集实际划分的一致性,定义公式为

$$\mathrm{RI}=\frac{a_d+a_s}{N(N-1)/2}. \tag{19}$$

其中: a_d表示实际属于不同的类簇,聚类划分结果也属于不同类簇的数据对数; a_s表示实际属于同一个类簇,聚类划分结果也属于同一个类簇的数据对数; N表示数据集中包含的数据个数.

3) MSSQ用于衡量在未知数据集实际分类情况下的聚类质量,定义公式为

$$\mathrm{MSSQ}=\sum_{k=1}^{|C|}\sum_{i=1}^{N_k}\|\boldsymbol{x}_k^m-\mu_k\|^2\Big/N. \tag{20}$$

其中: $|C|$是类簇个数, N是数据集中包含的数据个数, N_k是第k个类簇内的数据个数, $\boldsymbol{x}_k^m$是第k个类簇内的第m个数据, μ_k是第k个类簇的均值向量.

3.3 聚类效果测试

为了更加充分地验证ILocal-BART算法的有效性,对比BART算法和IGMM算法在多个数据集下的聚类效果. 表2给出了3种不同聚类算法对各个数据集的聚类个数、Acc和RI的对比结果.

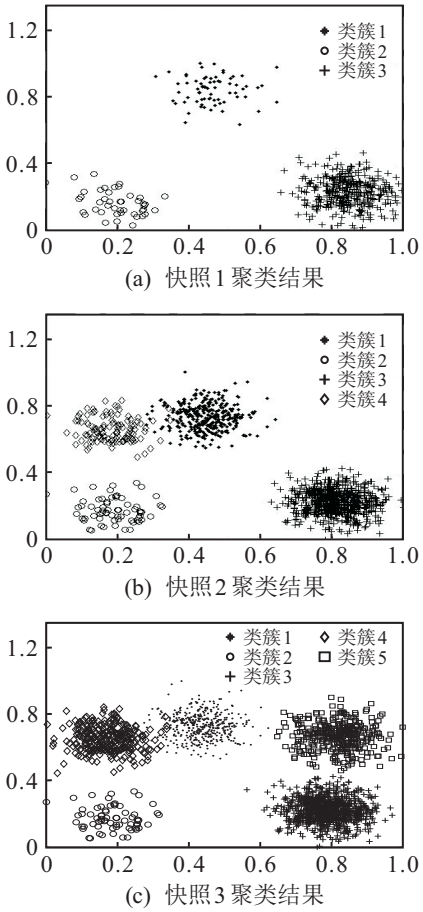
表 2 聚类算法评价指标对比

数据集	IGMM			BART($\rho=0.1$)			BART($\rho=0.5$)			BART($\rho=1$)			ILBART		
	$\lvert C\rvert$	Acc	RI	$\lvert C\rvert$	Acc	RI	$\lvert C\rvert$	Acc	RI	$\lvert C\rvert$	Acc	RI	$\lvert C\rvert$	Acc	RI
SD	5	0.842	0.876	5	0.962	0.958	5	0.852	0.849	5	0.648	0.594	5	0.995	0.974
BCH	5	0.564	0.662	4	0.823	0.835	4	0.743	0.782	2	0.453	0.396	3	0.905	0.882
Glass	4	0.539	0.633	5	0.683	0.654	3	0.847	0.898	1	0.483	0.492	2	0.986	0.952
Liver	5	0.546	0.552	3	0.886	0.854	3	0.942	0.935	1	0.456	0.543	2	0.982	0.964
Wine	4	0.895	0.885	5	0.645	0.694	4	0.795	0826	3	0.924	0.923	3	0.983	0.954
Yeast	8	0.794	0.789	12	0.745	0.723	8	0.903	0.932	5	0.483	0.492	9	0.935	0.941
SEA	2	0.673	0.658	4	0.784	0.821	2	0.879	0.899	2	0.682	0.703	2	0.959	0.962
Weather	2	0.596	0.632	3	0.793	0.792	2	0.896	0.852	2	0.532	0.549	2	0.966	0.974

由表2可以得出,ILocal-BART算法的聚类个数与数据集的实际划分个数相等. 除此之外,在无需设置任何参数的情况下,ILocal-BART算法对所有数据集的Acc和RI高于其他2种算法,并且均大于0.9,基本获得了与原数据集分布一致的聚类划分.

为了对算法的聚类过程进行更加直观地分析,图5给出了人工合成数据集SD在不同数据快照中的类簇分布情况,对各个类簇中的样本用不同的符号标记,可以清楚地看到在第1个快照中有3个类簇,第2个快照和第3个快照都是在上一个快照的聚类结果的基础上进行更新,并增加了一个类簇.

图6是人工合成数据集SD在不同数据快照中,类簇中心的动态变化过程,各快照中样本点的个数分别是494、720和786. 在快照1中,3个类簇的中心分别是(0.469 2, 0.841 5), (0.200 2, 0.163 4)和(0.845 7, 0.229 0); 在快照2中,前3个类簇的中心调整为(0.447 4, 0.731 9), (0.191 1, 0.175 2)和(0.811 5, 0.2232),增加了类簇4,类簇4的中心是(0.185 2, 0.655 1); 在快照3中,前4个类簇的中心调整为(0.434 3, 0.732 5), (0.186 3, 0.174 7), (0.788 1, 0.221 3)和(0.172 3, 0.6529)增加了类簇5,类簇5的中心是(0.819 1, 0.669 1). 从图6可以看出,相邻快照中同一个类簇之间的演化关系



(a) 快照1聚类结果

(b) 快照2聚类结果

(c) 快照3聚类结果

图 5 数据集SD的增量聚类过程

很大,类簇之间演化关系的最大值是0.997 2,最小值是0.883 3,类簇间的联系很紧密.

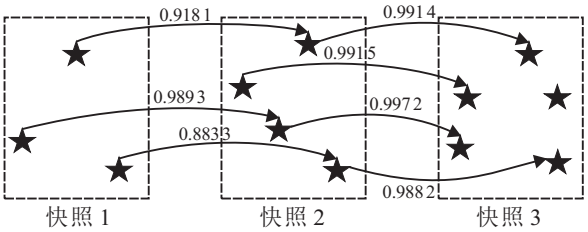


图 6　不同快照中聚类的动态变化

图7是IGMM算法、BART算法和ILocal-BART算法随不同快照所得聚类质量的变化情况. 以SEA不均衡数据流为例,对比IGMM算法、BART算法和ILocal-BART算法的评价指标MSSQ.

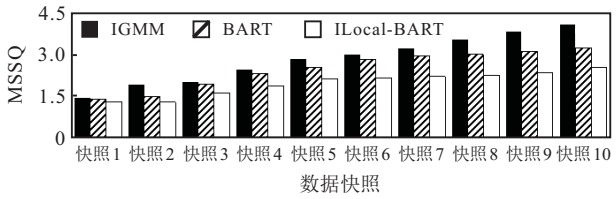


图 7　聚类过程质量比较

由图7可以看出,在同一快照中,ILocal-BART算法所得MSSQ均小于IGMM算法和BART算法,ILocal-BART算法的聚类效果优于IGMM算法和BART算法. 其中,在快照1中,3种算法的MSSQ几乎一样,分别是1.47、1.38和1.29. 在快照10中,3种算法的MSSQ分别是4.13、3.26和2.56,IGMM算法的MSSQ大于ILocal-BART算法.

3.4　快照对聚类结果的影响测试

为了分析数据快照个数和快照大小对ILocal-BART算法聚类结果的影响,本文利用人工合成不均衡数据集SD和不均衡数据流SEA形成不同的快照,对其进行聚类结果的比较. 图8和图9分别给出了SD不均衡数据集和SEA不均衡数据流的Acc和RI随着数据快照变化而变化情况. 由于人工合成不均衡数据集SD是一个静态数据集,并且数据量较少,将SD划分成5个快照,并且每个快照中的数据量相等;不均衡数据流SEA是一个动态数据集,快照的个数及每个快照中的数据量取决于数据流的频率,SEA被划分成10个快照,每个快照中的数据量分别是4 346、5 498、5 462、3 469、5 684、6 490、2 484、6 025、6 437和4 105.

从图8和图9的聚类评价指标的变化曲线可以看出,无论是每个快照中的数据量相等还是不断变化,Acc和RI几乎不变. 这是由于ILocal-BART算法是对数据进行逐条聚类,无论每个数据快照包含多少样本,聚类结果几乎不发生变化. 综上所述,快照的个数和大小对聚类结果几乎没有影响.

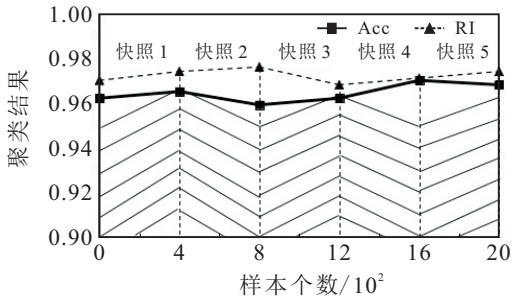


图 8　SD不均衡数据集在不同快照中的聚类结果

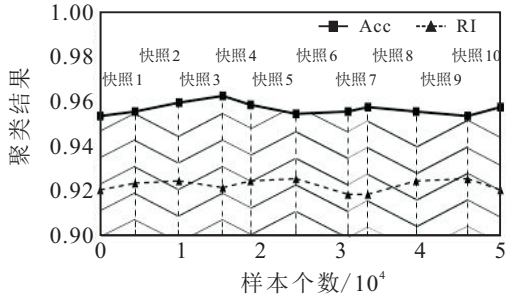


图 9　SEA不均衡数据流在不同快照中的聚类结果

3.5　聚类效率测试

采用不同数据集对3种不同算法的运行时间进行对比,对比结果如图10所示.

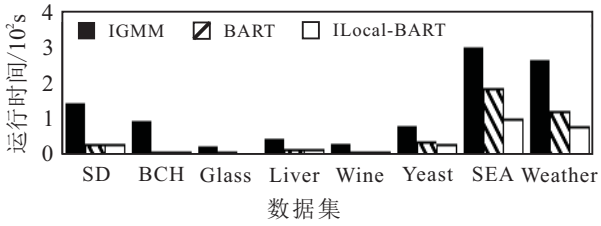


图 10　算法执行时间对比

由图10可以看出,针对同一个数据集,IGMM算法的运行时间最长,BART算法的运行时间几乎与ILocal-BART算法的运行时间相同. 这是由于IGMM算法要不断执行寻优操作,增加了时间消耗;BART算法和ILocal-BART算法采用相同的算法结构,虽然ILocal-BART算法对BART算法作出改进,但是没有增加时间复杂度,其中ILocal-BART算法对各个数据集的运行时间分别为34.634 s、6.508 s、7.222 s、3.528 s、15.522 s、29.756 s、103.232 s和80.136 s. 对比Yeast和Weather这两个数据集,它们的维度均为8,但Weather样本个数远多于Yeast,因此ILocal-BART对Weather数据集的运行时间远多于Yeast.

综上所述,ILocal-BART算法无需人为设置参数,能够自动确定匹配测试阶段的阈值,具有良好的自适应性;考虑类簇内数据的局部分布,提高了算法的聚类准确性;能够发现相邻快照中类簇的演化关系.

4 结 论

ILocal-BART算法首先利用数据快照读取数据,然后自适应地确定警戒测试阶段的阈值,并根据获胜类簇的局部分布确定新数据的所属类别,进一步提高了算法的聚类正确率. 同时,为了确保算法能够发现任意形状的聚类,根据新数据不断调整获胜类别的均值向量和协方差矩阵. 最后,发现类簇的演化关系. 实验结果表明,ILocal-BART算法在无需设置任何参数的情况下,能够对数据实现增量聚类,且聚类质量较原始增量聚类算法在准确性和自适应性方面均有显著提高,大大缩短了算法执行时间.

参考文献(References)

[1] 肖维. 用于高斯混合模型参数估计的EM算法及其初始化研究[J]. 电子测试, 2011(6): 26-30.
(Wei W. EM algorithm and its initialization research for parameter estimation of Gaussian mixture models[J]. Electronic Test, 2011(6): 26-30.)

[2] 何明, 冯博琴, 马兆丰, 等. 一种基于高斯混合模型的无监督粗糙聚类方法[J]. 哈尔滨工业大学学报, 2006, 38(2): 256-259.
(He M, Fend B Q, Ma Z F, et al. An unsupervised rough clustering method based on the gaussian mixture model[J]. J of Harbin Institute of Technology, 2006, 38(2): 256-259.)

[3] Khan I, Huang K Z, Ivanov K. Incremental density-based ensemble clustering over evolving data streams[J]. Neurocomputing, 2016, 191: 34-43.

[4] Huang D C, Xiao Chang L I. Incremental relative density-based clustering algorithm for mixture data sets[J]. Contorl and Decision, 2013, 28(6): 815-822.

[5] Patra B K, Ville O, Launonen R, et al. Distance based incremental clustering for mining clusters of arbitrary shapes[C]. Pattern Recognition and Machine Intelligence. Berlin: Springer, 2013: 229-236.

[6] Dong W, Chen L, He H, et al. Adjustable probability density grid-based clustering for uncertain data streams[J]. Int J of Advancements in Computing Technology, 2011, 3(8): 163-169.

[7] Wang T, Yu X, Alahakoon D, et al. An enhancing dynamic self-organizing map for data clustering[C]. The 10th IEEE Int Conf on Control and Automation(ICCA). Hangzhou: IEEE, 2013: 1324-1329.

[8] Vasighi M, Amini H. A directed batch growing approach to enhance the topology preservation of self-organizing map[J]. Applied Soft Computing, 2017, 55: 424-435.

[9] Bellas A, Bouveyron C, Cottrell M, et al. Model-based clustering of high-dimensional data streams with online mixture of probabilistic PCA[J]. Advances in Data Analysis and Classification, 2013, 7(3): 281-300.

[10] Gomes R, Welling M, Perona P. Incremental learning of nonparametric Bayesian mixture models[C]. 2008 IEEE Conf on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008: 1-8.

[11] Engel P M, Heinen M R. Incremental learning of multivariate Gaussian mixture models[C]. Advances in Artificial Intelligence-Sbia. São Bernardo Do Campo: DBLP, 2010: 82-91.

[12] Han K J, Narayanan S S. Agglomerative hierarchical speaker clustering using incremental Gaussian mixture cluster modeling[C]. 2008 Conf of the Int Speech Communication Association. Brisbane: DBLP, 2008: 20-23.

[13] Lughofer E, Sayed-Mouchaweh M. Autonomous data stream clustering implementing split-and-merge concepts Towards a plug-and-play approach[J]. Information Sciences, 2015, 304: 54-79.

[14] Lee H, Kwak K, Jo S. An incremental nonparametric Bayesian clustering-based traversable region detection method[J]. Autonomous Robots, 2017, 41(4): 795-810.

[15] Fan W, Sallay H, Bouguila N, et al. A hierarchical Dirichlet process mixture of generalized Dirichlet distributions for feature selection[J]. Computers and Electrical Engineering, 2015, 43: 48-65.

[16] Vigdor B, Lerner B. The Bayesian ARTMAP[J]. IEEE Trans on Neural Networks, 2007, 18(6): 1628-1644.

[17] Oentaryo R J, Meng J E, San L, et al. Bayesian ART-based fuzzy inference system: A new approach to prognosis of machining processes[C]. Prognostics and Health Management. Montreal: IEEE, 2011: 1-10.

[18] Islam M N, Chu K L, Seera M. Incremental clustering-based facial feature tracking using Bayesian ART[J]. Neural Processing Letters, 2017, 45(3): 887-911.

[19] Oentaryo R J, Er M J, Linn S, et al. Online probabilistic learning for fuzzy inference system[J]. Expert Systems with Applications, 2014, 41(41): 5082−5096.

[20] 沈豪, 张延园. 高效数据快照方法设计[J]. 微处理机, 2008, 29(4): 141-143.
(Shen H, Zhang Y Y. Design of high availability snapshot solution[J]. Microprocessors, 2008, 29(4): 141-143.)

[21] Comas D S, Meschino G J, Nowe A, et al. Discovering knowledge from data clustering using automatically-defined interval type-2 fuzzy predicates[J]. Expert Systems with Applications, 2016, 68: 136-150.

[22] Ghazikhani A, Monsefi R, Yazdi H S. Online neural network model for non-stationary and imbalanced data stream classification[J]. Int J of Machine Learning and Cybernetics, 2014, 5(1): 51-62.

（责任编辑：齐 霁）