软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(12):3884−3900 [doi: 10.13328/j.cnki.jos.006118]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于动态赋权近邻传播的数据增量采样方法*

陈晓琪[1,2], 谢振平[1,2], 刘 渊[1,2], 詹千熠[1,2]

[1](江南大学 人工智能与计算机学院,江苏 无锡 214122)

[2](江苏省媒体设计与软件技术重点实验室(江南大学),江苏 无锡 214122)

通讯作者: 谢振平, E-mail: xiezp@jiangnan.edu.cn

摘 要: 数据采样是快速提取大规模数据集中有用信息的重要手段,为更好地应对越来越大规模的数据高效处理要求,借助近邻传播算法的优异性能,通过引入分层增量处理和样本点动态赋权策略,实现了一种能够非常有效地平衡处理效率和采样质量的新方法.其中的分层增量处理策略考虑将原始的大规模数据集进行分批处理后再综合;而样本点动态赋权则考虑在近邻传播过程中对样本点进行合理的动态赋权,以获得采样的数据空间上更好的全局一致性.实验中,分别使用人工数据集、UCI 标准数据集和图像数据集进行性能分析,结果表明:新方法与现有相关方法在采样划分质量上可达到同等水平,而计算效率则可实现大幅提升.进一步将新方法应用于深度学习的数据增强任务中,相应的实验结果表明:在原始数据增强方法上结合进高效增量采样处理后,在保持总训练数据集规模的情况下,所获得的模型性能可实现显著的提升.

关键词: 数据采样;近邻传播;动态赋权;增量采样;数据增强

中图法分类号: TP311

中文引用格式: 陈晓琪,谢振平,刘渊,詹千熠.基于动态赋权近邻传播的数据增量采样方法.软件学报,2021,32(12):3884−3900. http://www.jos.org.cn/1000-9825/6118.htm

英文引用格式: Chen XQ, Xie ZP, Liu Y, Zhan QY. Incremental data sampling method using affinity propagation with dynamic weighting. Ruan Jian Xue Bao/Journal of Software, 2021,32(12):3884−3900 (in Chinese). http://www.jos.org.cn/1000-9825/6118.htm

Incremental Data Sampling Method Using Affinity Propagation with Dynamic Weighting

CHEN Xiao-Qi[1,2], XIE Zhen-Ping[1,2], LIU Yuan[1,2], ZHAN Qian-Yi[1,2]

[1](School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China)

[2](Jiangsu Key Laboratory of Media Design and Software Technology (Jiangnan University), Wuxi 214122, China)

Abstract: Data sampling is an important manner to efficiently extract useful information from original huge datasets. In order to fit with the requirements of efficiently dealing with more and more large-scale data, a novel incremental data sampling method originated from affinity propagation method is proposed, in which two integrated algorithm strategies including hierarchical incremental processing and the dynamic weighting of data samples are introduced. The proposed method mainly can balance the computational efficiency and sampling quality very well. For hierarchical incremental processing strategy, it firstly samples data items in batches and then composites samples by hierarchical way. For dynamic weighting of data samples strategy, it dynamically re-weights the preference to retain better global consistency of possible samples on data space in the incremental sampling procedure. In the experiments, artificial datasets, UCI datasets, and image datasets are used to analyze the sampling performance. The experimental results with several compared algorithms indicate that, the proposed method can gain similar sampling quality but with notably higher computational efficiency especially for more large-scale datasets. This study further applies the new method to data augmentation task in deep learning, and the corresponding experimental results show that the proposed method performs excellently. Concretely, if basic training dataset are processed by sampling

* 基金项目: 国家自然科学基金(61872166); 江苏省“六大人才高峰”项目(XYDXX-161); 江苏省科技计划(BE2018056)

Foundation item: National Natural Science Foundation of China (61872166); Six Talent Peaks Project of Jiangsu Province (XYDXX-161); Science and Technology Planning Project of Jiangsu Province (BE2018056)

收稿时间: 2019-08-01; 修改时间: 2019-11-23, 2020-06-15; 采用时间: 2020-07-03

enhancement with the proposed new method, the trained model performance using similar number of training samples can be significantly improved compared to traditional data enhancement strategies.

Key words: data sampling; affinity propagation; dynamic weighting; incremental sampling; data augmentation

现代信息技术的不断发展和进步,使各个领域累计了大量的数据,广泛涉及图像、文本、音频以及其他各类非结构化数据等.面对海量规模增长的数据内容,如何更有效地平衡信息获取的效率和有效性,正成为大数据处理的新重要问题.数据采样(data sampling)技术是一种能够在一定程度上解决上述问题的手段之一,该技术考虑从数据集中选取有代表性的样本作为整个数据集合的代表,在减小数据规模的同时,最大可能地保留数据集的有用信息,从而精简地表达数据集合包涵的重点知识.

当前,与数据集代表点采样相关的研究更多的是围绕图像数据.文献[1]提出了多字典不变稀疏编码方法,在不依赖人工标记、种子图像或其他先验知识的情况下,采集互联网上的代表性商标图像,为商标识别、商标侵权检测、品牌保护等领域自动提供原型、代表性图像或弱标签训练图像.文献[2]使用图像位置信息和图像相关标记,基于上下文和内容挖掘与位置和标记视觉相关联的图像,结合 *k*-means 聚类方法从视觉相关图像中选择代表性图像,为世界著名标志性建筑提供浏览摘要.文献[3]基于情感特征生成图像摘要,通过概率情感模型提取输入图像情感特征,在情感空间中对图像做聚类处理,结合覆盖性、情感一致性和显著性对簇排序并选择代表性图像.文献[4]提出一种区别于大多数利用视觉特征系统的基于语义知识的图像集合摘要方法,其通过图像间的语义相似度构建语义相关性网络,依据网络中心性选择代表性图像.文献[5]将图像自动摘要问题看作稀疏表示的字典学习问题,将图像摘要任务看作是字典学习任务,去实现大规模图像数据集的代表性图像选择.

现有的代表点采样方法大多基于聚类[6−9]方法,通常可抽象为 3 个步骤:(1) 用一组特征向量来表示数据集中的样本点,数值类型数据应用简单处理可转化为特征向量,图像数据则根据自身特点结合特征提取方法表示为一组高维特征向量;(2) 在特征向量空间中,采用聚类方法将样本点划分为若干簇;(3) 利用样本点代表性排序方法,从簇中选择具有代表性的样本.代表点采样方法可以从数据表示、聚类方法和代表点选择方法这 3 个方面进行研究.针对样本点聚类和代表点排序,常用的方法是使用 AP(affinity propagation)算法[6,7,10,11]进行代表点的选择.与传统聚类算法相比,AP 算法不需要预先设定聚类个数,对初始值不敏感,并且其生成的聚类中心是数据集中真实存在的样本点,不仅具有很好的代表性,且可以直接将聚类中心当作簇代表点.尽管 AP 算法在处理许多问题上具有优势,但它一个较大的不足是算法复杂度较高,在 Frey 等人所做的代表性灰度图像选择实验中[12],运行 1 次 AP 算法所消耗的时间与运行 100 次 *k*-center 算法所消耗的时间基本相同.与 *K*-means 算法相比,AP 算法的时间复杂度正比于数据规模的平方,而 *k*-means 算法的时间消耗则是线性增长.当数据规模较大时,AP 算法的计算时间将很长,且对存储空间要求也呈平方规模增长,这极大地限制了 AP 算法应用于大规模数据处理的实用性.

针对大规模数据的代表点采样问题,本文提出了一种基于动态赋权近邻传播的数据增量采样算法 ISAP (incremental data sampling using affinity propagation with dynamic weighting).主要包含两个策略.

(1) 近邻传播偏向参数的动态赋权.在 AP 算法的迭代过程中,利用样本点单次迭代聚类结果计算样本点自身轮廓系数,对 AP 算法的重要参数——偏向参数做动态调整,使最终采样结果能够包含更多的潜在性样本;

(2) 引入分层增量处理,将数据集划均匀划分为规模适中的子集,在各子集上分别执行全局偏向参数动态赋权的 AP 算法,获得局部最优代表点集合;然后在局部最优代表点集合上执行局部偏向参数动态赋权的 AP 算法,推选出整个数据集的最终代表点.

相比于现有的主要增量采样算法[13,14],它们的主要目的在于实现数据密度上的均匀采样.而 ISAP 算法目的在于实现数据空间上基于聚类划分的代表性样本获取.

为检验 ISAP 算法的性能及其在图像数据应用问题的价值,我们分别在人工合成数据集、UCI 标准数据集和图像数据集上进行了代表性样本采样任务,对比一些经典的和较新的方法,结果表明:我们的方法与现有相关方法在采样划分质量上处于同一水平,而计算效率则获得了大幅提升.进一步将新方法应用于深度学习的数据

增强任务中,相应的实验结果表明:在原始数据增强基础上,结合进高效增量采样处理后,训练数据多样性增加,在不改变总训练数据集规模的情况下,新方法介入所获得的模型质量可实现显著的提升.

本文的主要贡献:

(1) 提出了一种基于动态赋权近邻传播的数据增量采样算法,引入分层增量处理和样本点动态赋权策略,良好地实现了数据采样质量和效率的平衡;

(2) 在人工合成数据集、UCI 标准数据集和图像数据集上进行代表性样本采样任务,本文方法在提高采样效率的同时,保证了代表性样本的显著性和覆盖性;

(3) 将本文方法应用于深度学习的数据增强任务中,实验结果表明:在保持训练数据集规模不变的情况下,本文方法介入所获得的模型质量有显著提升.

本文第 1 节将介绍方法框架.第 2 节、第 3 节介绍标准 AP 算法及基于动态赋权近邻传播的数据增量采样算法 ISAP.第 4 节对算法进行相关分析.第 5 节通过实验评估算法的效果.第 6 节对本文进行总结,对未来工作进行展望.

1 基于动态赋权近邻传播的数据增量采样方法 ISAP 框架

本文提出的基于动态赋权近邻传播的数据增量采样算法框架如图 1 所示,主要包含以下内容.

(1) 动态赋权的 AP 算法.本文提出一种用于代表点采样的动态赋权 AP 算法,在 AP 算法的迭代过程中,利用样本点单次迭代聚类结果计算样本点自身轮廓系数(silhouette coefficient),对样本点的偏向参数做动态调整,不断赋予新的权值直至方法收敛,使最终采样结果能够包含更多的潜在性样本;

(2) 分层增量采样.基于上述偏向参数动态赋权算法,结合整体样本点的全局初始偏向参数和局部代表点之间的相对于整体数据集的局部初始偏向参数,提出了分层增量的代表点采样算法 ISAP.算法架构如图 1 所示:首先,将样本集合划分为规模上大小均匀的子集,在每个子集上执行全局偏向参数动态赋权的 AP 算法,得到每个子集产生的局部代表点,所有子集产生的局部代表点组成局部代表点集,本文将该过程称为增量局部推选层;然后,在局部代表点集上利用动态赋权 AP 算法对局部代表点进行合并,产生数据集整体代表点,本文将该过程称为合并推选层;最后,将数据集中的非代表点划分给与其相似度最大的代表点,完成全局簇划分.

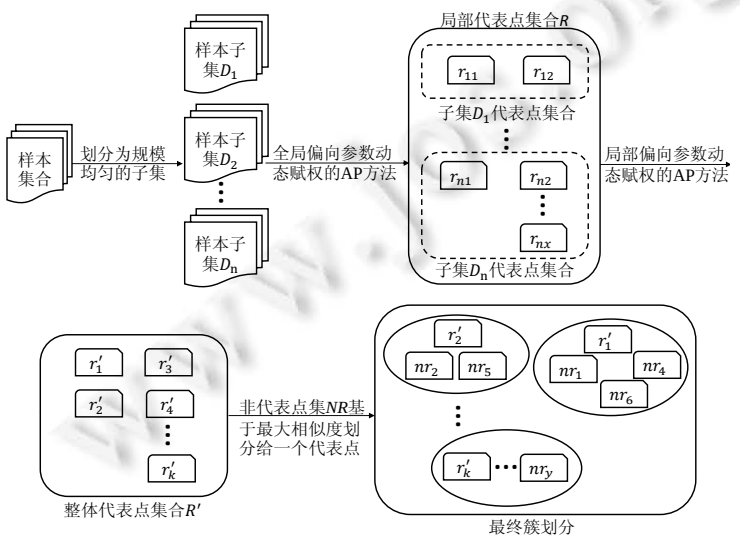


Fig.1 Flow chart of proposed incremental data sampling method

图 1 本文数据增量采样算法流程图

2 AP 算法中偏向参数的动态赋权

AP 算法[12]是一种 exemplar-based 聚类方法,它将所有数据样本看作是网络中的节点,通过节点间的双向消息传递,收敛得到最优的簇代表点集合.AP 算法与其他聚类方法的最大不同点是:其生成的聚类中心是数据集中真实存在的样本,具有很好的代表性,可以直接作为簇的代表点.

在基于聚类方法的代表点采样问题中,聚类质量与代表点采样质量紧密相关,聚类质量的提升,在一定程度上能够解决采样结果的代表性问题.AP 算法改善聚类质量的一个方向是偏向参数,偏向参数直接影响最终聚类数目,决定聚类质量的好坏.偏向参数一般根据经验设置为统一常数[12,15,16],在 AP 算法的迭代过程中恒定不变,即所有样本点成为簇代表点的可能性从始至终是一样的.但是在实际情况中,样本点之间的差异使得它们成为簇代表点的可能性不尽相同;此外,在迭代过程中因为信息交换的影响,样本点成为簇代表点的概率是会变动的.因此,给所有样本点设置同样且恒定的偏向参数的做法并不恰当,容易导致不好的聚类结果.为便于后续讨论,表 1 中汇总给出了本小节与新方法有关的符号信息.

Table 1 Summary of relevant symbols on chapter 2

表 1 第 2 节新方法相关符号汇总

符号	意义
$P=\{p_k\}_{1\times N}$	输入 AP 算法的偏向参数
$S=\{s(i,j)\}_{N\times N}$	输入 AP 算法的相似度矩阵
θ	相似度截断值
$R=\{r(i,j)\}_{N\times N}$	吸引度
$A=\{a(i,j)\}_{N\times N}$	归属度
$Silhouette(i)$	一个样本点的轮廓系数
Δp_k	样本点的偏向参数变动量
$O=\{o_k\}_{1\times N}$	聚类中心标识

2.1 初始偏向参数定义

在实际问题中,根据局部密度、语义性等信息可以知道,样本点成为簇代表点的可能性是有区别的,所以将偏向参数 $P=\{p_k\}_{1\times N}$ 设置为统一常数的做法并不恰当.

为了充分考虑数据本身所蕴含的信息,减少信息迭代中不必要的计算,文献[17,18]依据数据集中样本点的分布情况,为每个样本赋予不同的初始偏向参数.本文方法仅考虑样本点在局部范围内与其他样本之间的平均相似度,不考虑比例系数的影响,采用如下初始偏向参数计算公式:

$$p_k=\frac{\sum_{i\neq k}(s(i,k)\cdot I(cutoff))}{\sum_{i\neq k}I(cutoff)} \tag{1}$$

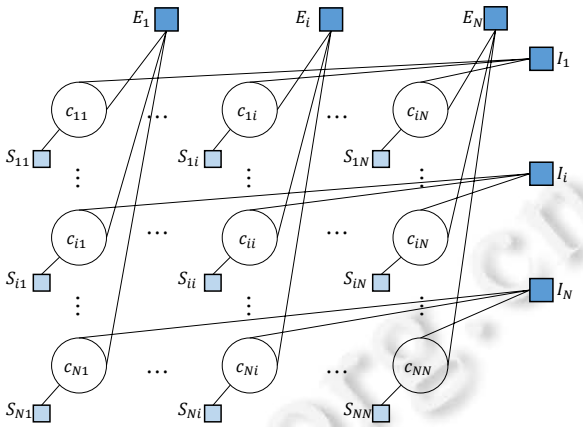
其中,

- $I(\cdot)$为指示函数;
- 参数 $cutoff$ 表示不等式 $s(i,k)\geqslant\theta$的结果:如果 $s(i,k)\geqslant\theta$成立,则 $I(cutoff)=1$;否则为 0.

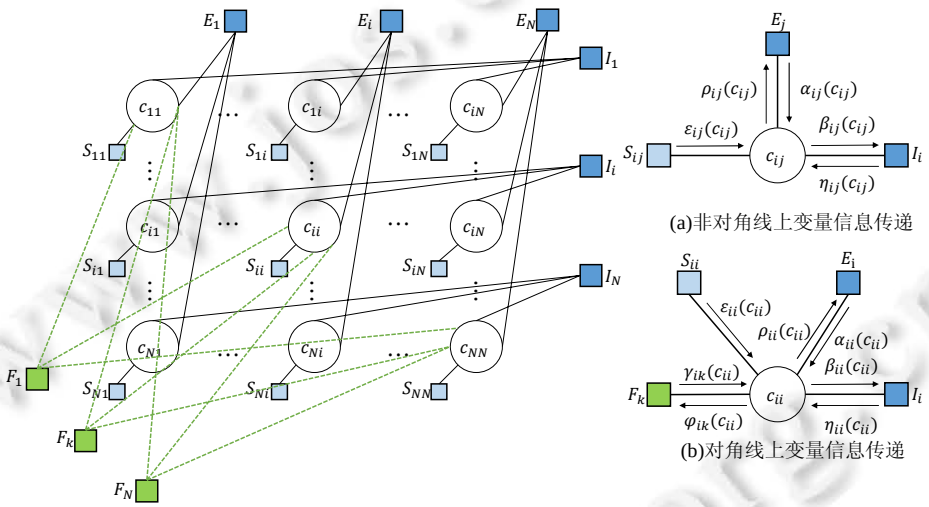
参数θ定义为相似度截断值,其大小的设置与数据集所有样本点间的相似度紧密程度相关,其在某种程度上也反映了原始距离度量 $s(\cdot,\cdot)$的上限合理尺度.

2.2 偏向参数动态赋权

标准 AP 算法的输入是相似度矩阵 $S=\{s(i,j)\}_{N\times N}$以及包含在 S 中的偏向参数 $P=\{p_k\}_{1\times N}(s(k,k)\leftarrow p_k)$,$p_k$ 表示样本点 k 被选为聚类中心的可能性.对于 N 个样本点的聚类问题,AP 算法旨在找到一组聚类中心,将每一个非中心点分配给唯一的聚类中心,以便最大化非中心点和其聚类中心之间的相似性以及各聚类中心的偏向参数的总和.一种典型的 AP 算法二元变量因子图[12]如图 2(1)所示.



(1) 标准 AP 算法二元变量因子图



(2) 加入新约束的 AP 算法因子图及其信息传递

Fig.2 Factor graph of the AP method

图 2 AP 算法二元变量因子图及其改进后的因子图

在图 2(1)中,$\{c_{ij}\}_{N\times N}$ 中的 c_{ij} 属于{0,1},c_{ij}=1 表示点 j 是点 i 的聚类中心.标准 AP 算法需要满足两个聚类约束条件:一个是聚类中心的唯一性,即一个点只能属于一个类(图 2(1)中约束 *I*);另一个是聚类中心的存在性,即当一个点是另一个点的聚类中心时,该点必然是自己的聚类中心(图 2(1)中约束 *E*).基于这两个约束,AP 算法通过在因子图上的信息传递更新使得全局函数 $S(C)$最大,其具体定义见公式(2).

$$I_i(c_{i:})=\begin{cases}0, & \sum_j h_{ij}=1\\ -\infty, & \text{otherwise}\end{cases}$$

$$E_j(c_{:j})=\begin{cases}0, & c_{jj}=\max_i h_{ij}\\ -\infty, & \text{otherwise}\end{cases} \tag{2}$$

$$\text{Max}: S(C)=\sum_{i,j}S_{ij}c_{ij}+\sum_j E_j(c_{:j})+\sum_i I_i(c_{i:})$$

针对这一 NP 难问题,依据 max-sum 算法[12,19]可推导出 AP 算法的迭代更新公式.在 AP 算法中,有两种重要的消息在样本点间传递,分别是吸引度(responsibility)和归属度(availability),在算法中表现为吸引度矩阵 $R=\{r(i,j)\}_{N\times N}$ 和归属度矩阵 $A=\{a(i,j)\}_{N\times N}$,其更新公式如下:

$$\begin{cases} r(i,j)=s(i,j)-\max\limits_{k\neq j}(a(i,k)+s(i,k)) \\ a(i,j)=\min\left(0,r(j,j)+\sum\limits_{k\neq i,j}\max(0,r(k,j))\right), j\neq i \\ a(i,i)=\sum\limits_{k\neq i}\max(0,r(k,i)) \end{cases} \tag{3}$$

吸引度 $r(i,j)$表示样本点 j 作为样本点 i 的聚类中心的合适程度,归属度 $a(i,j)$表示样本点 i 选择样本点 j 作为聚类中心的合适程度.对于任意的样本点 k,如果其对于自身的归属度 $a(k,k)$和吸引度 $r(k,k)$之和大于 0,则样本点 k 是一个聚类中心.

从标准 AP 算法的迭代过程和聚类中心选取方法可以看出,偏向参数 $P=\{p_k\}_{1\times N}(s(k,k)\leftarrow p_k)$的大小直接影响最终聚类数目.在大多数基于 AP 过程的算法中,各样本的 p_k 被设为同一常数并且在迭代过程中保持不变.一旦 p_k 被确定,聚类结果也就被确定下来,过程中没有其他的方法来对结果进行修正.为了降低初始 p_k 对聚类结果的影响,可以引入额外的信息在聚类过程中动态改变偏向参数.本文考虑引入节点的轮廓系数对每一次迭代产生的中间结果进行评价,从而依据评价结果动态改变每个点的 p_k.即:当 AP 算法产生至少一个聚类中心时,认为点与点之间产生了相互作用.依照这一思想,AP 算法的因子图模型结构将相应地发生变化,在原本的聚类约束条件下,将新增有关于聚类中心的约束,改变后的因子图模型及其信息传递如图 2(2)所示.新的因子图新增了聚类约束条件 F,该约束表示当产生至少一个聚类中心时,各节点之间存在相互作用,新增约束公式及改变后的全局函数 $S(C)$如下:

$$\begin{aligned} &F_k(c_{11},\ldots,c_{NN})=\begin{cases}0, & \sum_i c_{ii}>0\\ -\infty, & \text{otherwise}\end{cases} \\ &Max:S(C)=\sum_{i,j}S_{ij}c_{ij}+\sum_j E_j(c_{:j})+\sum_i I_i(c_{i:})+\sum_k F_k(c_{11},\ldots,c_{NN}) \end{aligned} \tag{4}$$

依据 max-sum 算法,可以从新的全局函数及信息传递过程中[17,20]推导出如下公式:

$$\begin{cases} \rho_{ij}=s_{ij}-\max\limits_{k\neq j}(s_{ik}+\alpha_{ik}) \\ \alpha_{ij}=\min\left(0,\rho_{jj}+\sum\limits_{k\neq i,j}\max(0,\rho_{ij})\right) \\ \alpha_{ii}=\sum\limits_{k\neq i}\max(0,\rho_{ki}) \\ \varphi_{ik}=-\max\limits_{i'\neq i}(s_{ii'}+\alpha_{ii'})+s_{ii}+\alpha_{ii}-\gamma_{ik} \\ \gamma_{ik}=-\max\limits_{i'\neq i}(0,\varphi_{i'k}) \end{cases} \tag{5}$$

与公式(2)比较可以发现,新增信息量 γ_{ik} 和 φ_{ik} 不影响原本吸引度和归属度的更新过程.因此,引入额外信息改变 p_k 不会改变 AP 算法的核心公式.

轮廓系数是确定聚类质量的一种常用指标,通常用来评价整体聚类质量的好坏,但也可以用来评估某一个样本点簇归属的合适程度.假设数据集被划分为 x 个子集$\{C_1,C_2,\ldots,C_x\}$,$a(i)$是子集 C_k 中的样本点 i 到同簇其他样本点间的平均相似度,$b(i,C_j)(i\neq j)$是样本点 i 到子集 C_j 中所有样本点的平均相似度.$b(i,:)=\min\{b(i,C_j)\}$.一个样本点的簇归属合适程度计算方法如下:

$$Silhouette(i)=\frac{b(i,:)-a(i)}{\max\{a(i),b(i,:)\}} \tag{6}$$

Silhouette(i)的取值范围为[−1,1]:越接近 1,则样本点 i 的簇归属越合理;接近−1,说明样本点 i 应属于其他簇;为 0,则说明样本点 i 在两个簇的边界上.

在 AP 算法迭代产生聚类中心的情况下,引入对样本点的轮廓系数评价.参考 *Silhouette* 指标的结果可以得知样本点是否被正确地划分,以及被选作聚类中心的样本点是否是正确的聚类中心,从而调整每个样本的偏向参数$\{p_k\}$,动态改变样本点成为聚类中心的可能性.显然,聚类中心样本与非聚类中心样本在调整上存在差异,因

此给出两条偏向参数更新规则.

规则 1. 对于非聚类中心样本:*Silhouette* 指标大于 0,说明该样本被划分到正确的簇,其成为聚类中心的可能性应该降低,p_k 应适当减小;*Silhouette* 指标小于 0,则该点没有被划分到正确的簇,其应该被划分到其他簇或成为一个新的聚类中心,p_k 应维持不变或适当增加.

规则 2. 对于聚类中心样本:*Silhouette* 指标大于 0,说明该样本是正确的聚类中心,p_k 应维持不变或适当增加;*Silhouette* 指标小于 0,说明该点不是正确的聚类中心,p_k 应适当减小.

基于上述两条规则,为了将 *Silhouette* 转化为合适的Δp,定义如下转换函数:

$$\Delta p = o_k \cdot \delta \cdot \frac{1-\mathrm{e}^{-Silhouette(k)}}{1+\mathrm{e}^{-Silhouette(k)}} \tag{7}$$

$O=\{o_k\}_{1\times N}$ 存放聚类中心标志,如果 o_k=1,表明样本点 i 是聚类中心;否则,o_k=−1.参数δ调整Δp 的变动幅度.引入基于 *Silhouette* 计算的Δp,当算法产生至少一个聚类中心后,偏向参数动态赋权的 AP 算法更新公式如下:

$$\begin{cases} r(i,k)=s(i,k)-\max\limits_{k'\neq k}(a(i,k')+s(i,k')) \\ a(k,k)=\sum\limits_{k'\neq k}\max(0,r(k',k)) \\ a(i,k)=\min\left(0,r(k,k)+\sum\limits_{k'\neq i,k}\max(0,r(k',k))\right) \\ \Delta p=o_k\cdot\delta\cdot\dfrac{1-\mathrm{e}^{-Silhouette(k)}}{1+\mathrm{e}^{-Silhouette(k)}} \\ s(k,k)=s(k,k)+\Delta p_k \end{cases} \tag{8}$$

偏向参数动态调整的 AP 算法在复杂度上与标准 AP 算法没有差别,但是因为考虑到了样本点之间的相互作用关系,使得偏向参数能够即时反映样本点当前迭代时刻的状态.对于基于聚类划分的采样问题来说,期望得到的代表点能够更大程度地覆盖原始数据的空间区域,包含更多数据密度较小区域的特异性样本,而引入偏向参数的动态调整能够在一定程度上消除原始数据密度分布给初始偏向参数带来的影响,从而能够找到更多低数据密度空间区域中的代表性样本.引入轮廓系数动态改变偏向参数后,算法过程如下.

算法 1. *adjustPreferenceAP*(S,P,λ,δ,*maxiter*,*conviter*).

输入:数据集相似度矩阵 $S=\{s_{ij}\}_{N\times N}$,偏向参数 $P=\{p_{11},...,p_{NN}\}$,阻尼系数λ,变动幅度δ,

最大迭代次数 *maxiter*,最大收敛状态比较次数 *conviter*;

输出:聚类中心 *classcenter*.

init $R\leftarrow\{0\}_{N\times N}$, $A\leftarrow\{0\}_{N\times N}$

init $E\leftarrow\{0\}_{N\times conviter}$

for *curriter*←1 to *maxiter* **do**

  init $R_{old}\leftarrow R$, $A_{old}\leftarrow A$

  update R and A according to Eq.(8) line 1~ line 3

  $R\leftarrow(1-\lambda)\times R+\lambda\times R_{old}$

  $A\leftarrow(1-\lambda)\times A+\lambda\times A_{old}$

  init classcenter←{False}$_{N\times 1}$, *centernum*←0

  for *k*←1 to *N* **do**

    if $R[k][k]+A[k][k]>0$

      classcenter[*k*]←True

      centernum←*centernum*+1

    end if

  end for

  E[:,*curriter*%*conviter*]←*classcenter*

if *centernum*>0
init $C\leftarrow\{0\}_{1\times N}$
for $k\leftarrow$1 to N **do**
$C[k]\leftarrow\text{argmax}_j(R[k][j]+A[k][j])$
end for
for $k\leftarrow$1 to N **do**
calculate Δp according to Eq.(8) line 4
update $S[k][k]\leftarrow S[k][k]+\Delta p$
end for
end if
if *curriter*$\geqslant$*conviter*
conver$\leftarrow$*judge convergence* by E (True or False)
if (*conver* and *centernum*) or (*curriter*=*maxiter*)
return *classcenter*
end if
end if
end for

3 分层增量采样

AP 算法中,聚类中心为真实样本点这一特性,使得其非常适用于代表点的采样.但是标准 AP 算法的复杂度较高,不适用于大规模数据的代表点采样.本文结合分层及增量处理的采样策略,基于上述偏向参数的动态赋权 AP 算法,实现对数据集的高效采样,以期实现处理效率和采样质量的更好兼顾.

本文提出的分层增量采样方法框架如图 1 所示,是一层增量局部推选加一层合并推选的“1+1”模式的采样.算法的输入为已划分好的子集序列,为了保证算法的计算效率,各子集的规模要相同或相近,可以采用顺序划分或随机采样划分的方法.

将数据集划分为规模适中的子集$\{D^1,D^2,\ldots,D^n\}$,分层增量采样的第 1 层(增量局部推选)依次处理样本子集D^i,选出基于全局偏向参数信息和局部相似度信息的子集代表点 $R^i=\{r_{i1},r_{i2},\ldots,r_{ix}\}$,构成子集代表点集 R.第 2 层(合并推选)完全基于 R 的全局信息,即数据集的局部信息,从 R 中挑选出数据集的整体代表点 $R'=\{r'_1,r'_2,\ldots,r'_k\}$.

在增量局部推选层中,为了将更多的潜在代表点挑选出来,需要考虑数据集的全局相似度信息.因此,将样本点基于整个数据集中的全局初始偏向参考度作为 AP 算法的输入.依据公式(1)计算数据集的全局偏向参数 $PG=\{pg_{kk}\}$,每个子集的全局偏向参数集合为 PG^i.

合并推选层采样是对增量局部推选层采样结果的合并推选.在第 1 层中已经基于全局偏向参数获得了所有潜在的整体代表点组成局部代表点集,在第 2 层采样中,只需基于局部代表点集的全部相似度信息,即整体数据集的局部节点之间的相似度信息进行潜在代表点的采样选择,其中,仍采用公式(1)初始化其中局部代表点集相对于整体数据集的局部偏向参数 $PN=\{pn_{kk}\}$.

结合截断值参数θ的引入,本文方法可考虑对原始的所有数据点间的相似度矩阵进行约简,将小于相似度截断值的边进行删除,进而达到约简相似度矩阵的目的,加速 AP 算法的计算效率.综前所述,可归纳给出如下的综合算法过程.

算法 2. *ISAP*($\{D^1,D^2,\ldots,D^n\}$,θ,λ,δ,*maxiter*,*conviter*).

输入:子集序列 $D^1,D^2,\ldots,D^n$,相似度截断值θ,阻尼系数λ,变动幅度δ,
最大迭代次数 *maxiter*,最大收敛状态比较次数 *conviter*;

输出:代表点集 R',全局划分 *Cluster*.

calculate global similarity matrix $SG_{N\times N}$ base $D^1\cup\ldots\cup D^n$
calculate global preference $PG_{1\times N}$ base θ and SG according to Eq.(1)
for $i\leftarrow$1 to n **do**
  $M\leftarrow|D^i|$
  calculate similarity matrix $S_{M\times M}$ base $D^i_{1\times M}$
  $S[S<\theta]\leftarrow 0$
  init $PG^i\leftarrow\varnothing$
  for $j\leftarrow$1 to M **do**
    add $PG[D^i[j]]$ to PG^i
  end for
  $R^i\leftarrow adjustPreferenceAP(S,PG^i,\lambda,\delta,maxiter,conviter)$
end for
init $R\leftarrow\varnothing$
for $i\leftarrow$1 to n **do**
  for $j\leftarrow$1 to $|R^i|$ **do**
    add $R^i[j]$ to R
  end for
end for
$K\leftarrow|R|$
calculate lobal similarity matrix $SL_{K\times K}$ base R
$SL[SL<\theta]\leftarrow 0$
calculate lobal preference $PL_{1\times K}$ base θ and SL according to Eq.(1)
$R'\leftarrow adjustPreferenceAP(S,PL,\lambda,\delta,maxiter,conviter)$
init $Cluster\leftarrow\{\{\varnothing\}\}_{1\times|R'|}$, $NR\leftarrow D\backslash R'$
for $r\leftarrow$1 to $|R'|$ **do**
  add $R'[r]$ to $Cluster[r]$
end for
for $i\leftarrow$1 to $|NR|$ **do**
  $c\leftarrow\text{argmax}_{r\in 1,\ldots,|R'|}(SG[NR[i]][R'[r]])$
  add $NR[i]$ to $Cluster[c]$
end for
return R', $Cluster$

选出数据集的整体代表点后,如果需要实现单簇多代表点的选择,则可依据最大相似度将非代表点分配给一个代表点,完成基于最大相似度分配的全局簇划分.然后依据最终的簇划分,在每个簇中选取一组点放入代表点集.

4 算法分析

假设数据集被划分为 K 个规模为 M 的子集($K<<M$),对各子集执行 AP 算法的时间复杂度为 $O(M^2\log M)$,时间总和为 $t_1=O(KM^2\log M)$.假设各子集推选出的平均局部代表点数与子集规模 M 的比例为ρ($0<\rho<<1$),即各子集推选出的平均局部代表点数为$\rho\cdot M$,则局部代表点总数为$\rho\cdot K\cdot M$.在局部代表点集合上再次执行 AP 算法的时间为 $t_2=O(\rho^2K^2M^2\log\rho KM)$,因此分层增量采样方法耗费的总时间为 t_1+t_2.需要明确的是:子集规模 M 要远大于子集个数 K,子集推选出的平均局部代表点数与子集规模的比例ρ是一个大于 0 且远小于 1 的数.因此,子集规模是对

算法时间消耗影响最大的因素.在数据集规模不变的情况下,可以推算(t_1+t_2)正比于Mlog$M$,而标准 AP 算法处理相同规模数据的时间复杂度为 O(K^2M^2log(KM)),显然,分层增量采样方法的效率更高.

在各子集上执行 AP 算法的空间复杂度为 O(M^2),每一次增量推选过程中消耗的存储空间不变,因此增量执行 K 次规模为 M 的 AP 算法的空间消耗为 s_1=O(M^2).依照上述设定,在局部代表点数为ρ·K·M 的情况下,在局部代表点集上执行 AP 算法的空间消耗为 s_2=O($\rho^2K^2M^2$),因此分层采样方法的空间复杂度为 O($\rho^2K^2M^2$).而标准 AP 算法的空间复杂度为 O(K^2M^2).ρ是一个大于 0 且远小于 1 的数,因此分层增量采样方法在空间上的消耗同样优于标准 AP 算法.分层增量采样方法在时间效率和存储空间方面均可以得到提升.

相比于标准的 AP 算法,ISAP 中引入了大规模数据的分层处理策略,在第 1 层处理中对原始数据进行一次局部约简处理后,仅选取部分代表性数据参与第 2 层的再处理,从理论上看,可能会造成代表性数据对原始数据信息携带不足的问题,进而导致最终采样得到的代表性样本在全局上的偏离问题.虽然如此,结合我们以前的相关研究[21,22]发现:在增量学习过程中,如果阶段性选择的代表点集能够构成对目标问题解的一个整体有效的表达,那么在增量过程中逐步筛选掉非代表性样本,仅基于保留下来的数据参与后续处理的做法,理论上在一定条件下可以达到与全局方法的一致.

具体到 ISAP 算法,考虑在第 1 层处理后保留的局部代表性样本集是在整体上对原始数据空间区域进行代表性表达,而非数据密度上的约简采样.这样,AP 算法自身的特性就显得非常有优势,不仅可以有效地约简冗余数据,减少计算量,而且能够一定程度上解决原始数据空间中局部数据密度可能存在较大程度不平衡的问题,以使得最终得到的采样结果更优.

5 实验分析

本文提出的代表点采样方法基于 AP 聚类,为检验算法有效性,同样与基于 AP 的其他聚类方法进行比较,几种比较方法如下.

(1) 标准 AP 算法的代表点采样:使用文献[12]提出的标准 AP 算法来做代表点选取;

(2) 分层近邻传播算法 HAP 的代表点采样:使用文献[23]提出的一种基于底层推举和上层推举的层次 AP 算法来做代表点选取.该方法同样采用分层的策略,底层基于标准 AP 算法,上层基于自适应 AP 算法;

(3) 近邻赋值的增量式 AP 聚类算法 IAPNA 的代表点采样:使用文献[16]提出的一种增量式 AP 算法来做代表点选取.该方法通过近邻赋值构建新增数据与已有数据之间的消息传递关系,增量式地扩充吸引度和归属度矩阵,直至方法收敛得到结果.

相关算法实验中,首先需要计算样本点 i 和 j 之间的相似度 $s(i,j)$,$s(i,j)$的值越大,表示两个样本越相近.实验中,对两个样本点间的相似度 $s(i,j)$采用归一化计算定义:

$$s(i,j)=\frac{\max(D)-d(i,j)}{\max(D)-\min(D)} \tag{9}$$

D={$d(i,j)$}是样本点之间的欧式距离矩阵,max 和 min 操作分别表示取矩阵元素中的最大值和最小值.上式计算后,两个样本点完全相同时的相似度为 1,完全不同时的相似度为 0.

实验中,参考各自算法特点和相关文献,AP 算法的偏向参数取相似度矩阵中位值;对于 IAPNA 方法参数 pc,根据算法输出簇数和采样质量选取对比选取;对于 ISAP 算法中的截断参数θ和变动幅度参数δ,一方面也可类似地根据算法输出簇数和采样质量选取对比选取,其中的采样质量可以是使用有监督的交叉验证选取;或者使用无监督的质量指标(如代表点间平均欧式距离)作为参照.

此外,截断参数θ一定程度上反映了数据点间距离度量合理性的界限范围,而变动幅度参数δ是从微观上对偏向参数进行调整,解释样本点对于轮廓系数的学习能力.由此,取θ选择范围可取[10^{-2},max(D)],δ选择范围可设定为[10^{-3},10^{-1}],其中,max(D)同公式(9)中的含义.

5.1 评价指标

实验从聚类质量、采样质量和方法效率这 3 个方面对代表点采样方法进行评价.标准化互信息(normalized

mutual information,简称 NMI)是广泛用于聚类质量评估的聚类评价指标,用于度量方法结果与标准结果之间的相似度,其定义如下:

$$NMI=\frac{-2\cdot\sum_{i=1}^{C_X}\sum_{j=1}^{C_Y}C_{ij}\cdot\log\left(\frac{C_{ij}\cdot N}{C_i\cdot C_j}\right)}{\sum_{i=1}^{C_X}C_i\cdot\log\left(\frac{C_i}{N}\right)+\sum_{j=1}^{C_Y}C_j\cdot\log\left(\frac{C_j}{N}\right)} \tag{10}$$

其中,N 是数据样本总数;C_X 和 C_Y 分别是数据集的真实划分和算法聚类结果包含的簇数;C_{ij} 表示同时属于真实划分簇数 i 和聚类结果簇数 j 的样本数量.NMI 的取值范围为[0,1],其值越大,表示聚类结果与真实结果越匹配.

针对采样质量,本文引入代表点间平均欧式距离(average euclidean distance of representative points,简称 AEDRP)和距离方差(variance of representative points,简称 VRP)以及新设计综合覆盖率(comprehensive coverage rate,简称 CCR)来评价选取代表点的显著性和覆盖性.代表点间平均欧式距离越大,距离方差越小,说明选取的代表点的显著性越好.综合覆盖率考虑反映在合理代表点个数下,代表点集对原始数据集的覆盖程度.值越大,说明选取的代表点的覆盖性越好.其定义如下:

$$CCR=\mathrm{e}^{-\frac{N_R}{N_{NR}}}\cdot\frac{\sum_{i,j}I(dist(:,j)_{\min}\leqslant d_{mean})}{N_{NR}},i\in R,j\in NR \tag{11}$$

其中,N_R 是算法划分的类数即代表点数目,N_{NR} 是非代表点数,R 和 NR 分别为代表点集和非代表点集;$dist(i,j)$是代表点 i 和非代表点 j 间的欧式距离,$dist(:,j)_{\min}$ 是非代表点 j 与所有代表点之间的最小距离;d_{mean} 是整个数据集中数据点间的平均欧式距离;$I(\cdot)$为指示函数.CCR 值综合考虑代表点占原始数据大小的比例以及代表点对数据集在一定范围内的覆盖程度,通常情况下,选取更多数量的代表点能够明显提升代表点对数据集的覆盖程度.因此,在不考虑选取代表点数量的情况下去比较代表点对数据集的覆盖程度是不合理的,所以在评价时引入与代表点数目相关的系数,可以更为合理地评价代表点对数据集的覆盖程度.当然,在采样问题中,覆盖度是更被看重的问题,因此在 CCR 评价指标中,代表点数目的影响权重较小,代表点对数据集覆盖程度的影响权重较大.

5.2 数值数据实验分析

在 3 组 UCI 标准数据集和 4 组人工合成数据集上进行代表性样本采样实验,数据集详情见表 2.

Table 2 Descriptions of numerical experimental data sets

表 2 数值型实验数据集描述

	数据集	大小	类数	是否平衡
1	Iris	150	3	是
2	wine	178	3	否
3	yeast	1 484	10	否
4	Set 1	200	4	是
5	Set 2	450	3	是
6	Set 3	800	6	否
7	Set 4	1 200	8	是

真实数据集 iris,wine,yeast 来源于 UCI,人工合成数据集 Set 1~Set 4 是随机生成的二维数据集,均符合正态分布,其样本分布情况如图 3 所示.对每个数据集,依据公式(9)计算任意数据集内样本点间的相似度.4 种算法的 AP 阻尼系数λ设为 0.9,HAP 算法和 ISAP 算法的分区子集大小设为数据规模的 0.2[23],如果子集规模超过 500,则设为 500.采样选取每个最终簇的一组数据(算法输出的代表点)作为代表点.

在实验过程中,调整 IAPNA 算法参数 pc、ISAP 算法截断参数θ和变动幅度参数δ进行多次实验.参数 pc 在真实数据集上的数值来源于文献[16],参数θ,δ以及人工数据集上的 pc 综合算法输出的簇数和采样质量选取.最终实验参数设置见表 3.

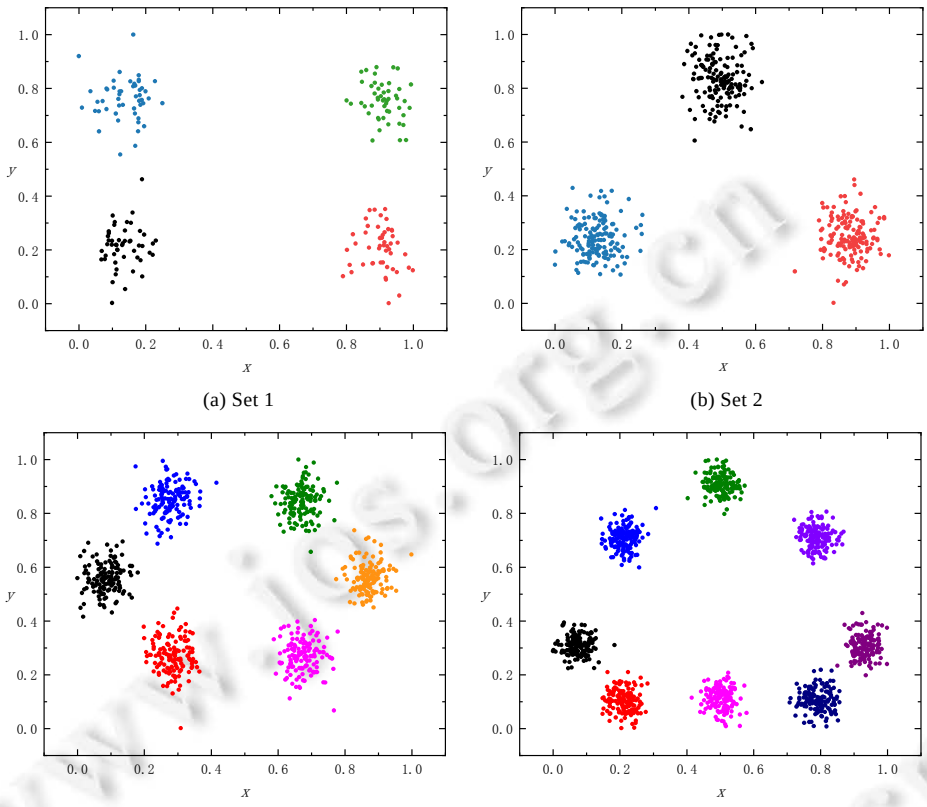


Fig.3 Four synthetic data sets

图 3 人工合成数据集情况

Table 3 Parameter setting for numerical experimental data sets

表 3 数值型数据实验算法参数设置情况

数据集		Iris	wine	yeast	Set 1	Set 2	Set 3	Set 4
IAPNA	*pc*	0.015	0.015	0.011	0.015	0.017	0.015	0.013
ISAP	θ	0.05	0.23	0.45	0.05	0.13	0.25	0.42
	δ	0.01	0.1	0.01	0.01	0.005	0.001	0.009

相应的实验结果见表 4、表 5.从表中结果可以看到:AP 算法在小规模数据上耗时最短,但随着数据规模增加其效率大幅下降,并且在所有数据集上的聚类质量和采样质量都不太理想;IAPNA 算法是增量式输入数据的全局 AP 算法,其聚类质量和采样质量最优;但随着数据规模扩大,其时间消耗剧增,不适用于大规模数据的代表点采样;HAP 算法与 ISAP 算法都是分层采样代表点的方法,两种方法的聚类质量和采样质量均优于 AP 算法,接近 IAPNA 算法,但两种算法耗费的时间远低于 IAPNA 算法.但是 HAP 算法在合并推选层上采用的是自适应 AP 聚类算法,需要执行多次标准 AP 算法得到最优的结果;随着数据规模的扩大,参与合并推选层采样的数据量也随之增大.因此,HAP 算法的时间消耗增加幅度比 ISAP 算法要大.ISAP 算法在聚类质量和采样质量与 IAPNA,HAP 算法处于相同水平,但计算消耗的时间显著较短.

上述实验结果也从实际应用角度表明:引入改进 AP 算法过程和分层处理的 ISAP 算法不仅获得了比标准 AP 算法更好的采样效果,且可以具有更好的计算效率.这也在一定程度上佐证了前文第 4 节中关于 ISAP 算法性能的理论分析结果.

4 种算法在 yeast 数据集上的效果都不理想.wine,yeast 和 Set 3 数据集内各簇的规模不尽相同,但是 wine 和

Set 3 各簇之间规模相似,而 yeast 的簇规模差别较大,是典型的不平衡数据.从实验结果看,几种对比算法在不平衡数据集的效果均相对较差.

对于代表点综合覆盖率性能指标,其与代表点对数据集的覆盖度和簇数(代表点数)有关,一般簇数越大,选取的代表点越多,其代表点对数据集的覆盖度越大.从 UCI 标准数据集上的结果来看:相比于标准 AP 算法,ISAP 算法能在产生更少代表点的同时获得较大的代表点覆盖度,因此其综合覆盖度较高.此外,从人工合成数据集上的结果来看:ISAP 算法在与 HAP,IAPNA 算法产生相同簇数的情况下,ISAP 算法产生的代表点间的平均距离较大,方差较小,说明代表点间的相似性低,挑选结果平滑,显著性较好.

Table 4 Performance results obtained by several compared sampling methods on three UCI datasets

表 4 不同方法在 3 个 UCI 数据集上的性能对比结果

评价指标	Iris				wine				yeast			
	AP	HAP	IAPNA	ISAP	AP	HAP	IAPNA	ISAP	AP	HAP	IAPNA	ISAP
NMI	0.373	0.576	0.576	0.767	0.306	0.448	0.872	0.743	0.124	0.189	0.106	0.212
类数	13	2	2	3	19	2	3	3	120	5	3	14
AEDRP	0.684	0.820	0.965	0.855	0.964	1.071	1.049	0.976	0.513	0.508	0.332	0.550
VRP	0.120	0.000	0.000	0.026	0.063	0.000	0.029	0.030	0.070	0.117	0.0026	0.101
CCR	0.611	0.647	0.647	0.640	0.619	0.595	0.590	0.595	0.916	0.653	0.387	0.991
耗时(s)	0.026	0.072	0.135	0.095	0.031	0.093	0.220	0.100	6.327	3.006	21.434	1.984

Table 5 Performance results obtained by several compared sampling methods on four synthetic datasets

表 5 不同方法在 4 个人工数据集上的性能对比结果

评价指标	Set 1				Set 2			
	AP	HAP	IAPNA	ISAP	AP	HAP	IAPNA	ISAP
NMI	0.68	1.00	1.00	1.00	0.40	1.00	1.00	1.00
类数	8	4	4	4	15	3	3	3
AEDRP	0.65	0.74	0.75	0.75	0.55	0.73	0.73	0.76
VRP	0.09	0.03	0.03	0.03	0.08	0.07	0.08	0.06
CCR	0.96	0.99	0.99	0.99	0.97	0.98	0.98	0.98
耗时(s)	0.04	0.11	0.20	0.11	0.13	0.14	1.14	0.18
评价指标	Set 3				Set 4			
	AP	HAP	IAPNA	ISAP	AP	HAP	IAPNA	ISAP
NMI	0.53	1.00	1.00	1.00	0.61	1.00	1.00	1.00
类数	28	6	6	6	29	8	8	8
AEDRP	0.48	0.54	0.54	0.54	0.55	0.61	0.61	0.61
VRP	0.05	0.02	0.02	0.02	0.07	0.04	0.04	0.04
CCR	0.96	0.99	0.99	0.99	0.98	0.99	0.99	0.99
耗时(s)	0.82	0.56	4.25	0.46	1.22	0.64	10.7	0.56

5.3 代表性图像选择

在本实验中,实验图像集来自搜索得到的车标图像以及 ILSVRC2014 图像集.车标图像集 Carlogo 共有 270 张图像,包含 18 个类别的车标,每类包含 15 幅图像.分别取 ILSVARC2014 验证集的前 50 类、前 100 类和前 150 类构成图像数据集 ILSVARC50,ILSVARC100,ILSVARC150,分别包含 2 500、5 000 和 7 500 张图像.

实验过程中,调整 IAPNA 算法参数 pc、ISAP 算法截断参数θ和变动幅度参数δ多次执行算法,综合算法输出的簇数和采样质量选取,最终实验参数设置见表 6.代表性图像选择实验仅评估算法的采样质量.其中,Carlogo 图像集采用 SIFT 特征匹配度作为图像间相似度,SIFT 相似度经过公式(9)转换后得出 Carlogo 数据集上的 *AEDRP* 指标.ILSVARC 图像数据集则使用卷积神经网络(convolutional neural networks,简称 CNN)提取图像的特征向量,依据公式(9)计算图像间的特征相似度.

Table 6 Parameter setting in representational image selection experiment

表 6 代表性图像选择实验参数设置情况

数据集		CarLogo	ILSVARC50	ILSVARC100	ILSVARC150
IAPNA	pc	0.003	0.000 5	0.000 1	–
ISAP	θ	0.27	0.3	0.2	0.1
	δ	0.03	0.01	0.008	0.05

实验结果见表 7.从表中结果可以看到:综合考虑代表性图像的平均距离、距离方差、综合覆盖度以及时间效率,ISAP 算法具有较为明显的优势.标准 AP 算法在各方面都不占优势;IAPNA 方法在数据集规模较大时时间耗费过大;而 HAP 算法得到的代表图像之间的平均距离较大,但是代表图像间距离的方差明显超过另外 3 种方法,其代表图像间的距离分布不平滑.

Table 7 Performance results obtained by serval compared methods on representational image selection problem

表 7 对比方法在代表性图像选择实验上的性能结果

评价指标	CarLogo				ILSVARC50			
	AP	HAP	IAPNA	ISAP	AP	HAP	IAPNA	ISAP
类数	41	12	19	18	167	57	36	46
AEDRP	0.79	0.77	0.79	0.81	108	108	91	91
VRP	0.02	0.03	0.02	0.01	458	937	161	146
CCR	0.84	0.94	0.93	0.94	0.93	0.97	0.98	0.97
耗时(s)	0.05	0.79	0.49	0.27	10.3	15.2	121	6.20
评价指标	ILSVARC100				ILSVARC150			
	AP	HAP	IAPNA	ISAP	AP	HAP	IAPNA	ISAP
类数	304	90	107	104	449	180	–	178
AEDRP	107	119	100	96	109	122	–	100
VRP	477	840	252	171	541	953	–	187
CCR	0.94	0.98	0.97	0.98	0.94	0.97	–	0.97
耗时(s)	51.3	191	1 303	26.5	155	691	–	58.4

ISAP 算法从数据集 CarLogo 中选择的代表性图像如图 4 所示.

Fig.4 Representational images selected by ISAP on CarLogo data set

图 4 ISAP 在 CarLogo 数据集上挑选的代表性图像

可以看到:通过本文方法得到的代表性图像很好地覆盖了数据集,能够作为数据集的代表.

5.4 数据增强应用

深度学习是数据驱动的方法,用规模更大、质量更好的数据集去训练神经网络一般都能够得到泛化性能更好的模型.但在实际情况中,数据的采集面临多重困难,人工采集的样本在多样性和规模上均不能满足实际训练的需求.数据增强即数据扩增,是一种有效扩充数据规模,解决训练样本不足问题的方法[24−26].数据增强能够扩充数据规模,增加数据噪声,使用增强后的数据集训练神经网络能够提高模型的泛化能力和鲁棒性.在图像识别领域,数据增强可以很好地提升训练模型的识别率.但简单的数据增强策略容易产生许多极其相似的图像序列.

考虑检验 ISAP 算法在数据增强任务上的价值,实验数据来源于加州理工大学开源数据集 leaves,包含 3 种类型的叶片,共 186 张图像.利用仿射变换、高斯噪声、区域衰减、高斯模糊等数据增强手段,将 leaves 数据集的规模扩充 10 倍至 1 860 张图像,命名为 leavesDa10.在 leavesDa10 的基础上,再次利用上述数据增强手段将数据集规模扩充 5 倍至 9 300 张图像,命名为 leavesDa50.

在 leavesDa50 上执行参数不同的 ISAP 算法,采样选取每个最终簇的 10 幅图像(ISAP 算法输出的代表点及

每个簇中离代表点最近的另外 9 幅图像,规模不足 10 的簇则选取其全部图像)作为代表图像,并考虑形成约 1 860 张增强样本图像为目标,最终生成增强数据集 leavesDaIsap0~leavesDaIsap4.leaves 叶片类型与部分增强图像如图 5 所示.

Fig.5 leaves blade type and partial enhanced images

图 5 leaves 叶片类型与部分增强图像

在各数据集上,从每类叶片中取 3/4 的图像用于卷积神经网络 CNN 训练,其余 1/4 的图像用于模型测试,使用参数不同的 ISAP 算法产生的约简数据集训练得到的模型平均识别率(10 次实验结果的平均)如表 8 所示.其中,采样指标 AEDRP 值依据算法直接输出的代表点进行计算.

Table 8 Performance results obtained by different ISAP-augmented deep learning on leaves recognition task

表 8 不同参数 ISAP 算法数据增强下 leaves 叶片识别问题的深度学习性能结果比较

	数据集名称	ISAP 参数	大小	AEDRP	平均识别率(%)
1	leavesDaIsap0	θ=0.05, δ=0.005	1 973	64.985	95.8
2	leavesDaIsap1	θ=0.01, δ=0.005	1 975	64.536	95.2
3	leavesDaIsap2	θ=0.2, δ=0.005	1 993	64.638	94.8
4	leavesDaIsap3	θ=0.05, δ=0.001	2 233	64.536	95.1
5	leavesDaIsap4	θ=0.05, δ=0.01	2 102	63.224	95.4

考虑以采样比例和 AEDRP 指标为依据,针对表 8 中的实验结果,将 leavesDaIsap0 数据集上训练的结果与其他方法进行比较.

考虑在 leavesDa50 上执行 HAP 算法,用上述选取代表性图像方法构成增强数据集 leavesDaHap;为了更好地与 HAP 算法进行对比分析,调整 ISAP 参数为 θ=0.3,δ=0.05,可形成规模为 4 584 的数据集 leavesDaIsap5,相应的实验结果见表 9,其中的平均识别率为 10 次实验的平均结果.

Table 9 Performance results obtained by different augmented datasets on leaves recognition task

表 9 不同增强数据集下 leaves 叶片识别问题的深度学习性能结果比较

	数据集名称	大小	平均识别率(%)
1	leaves	186	80.5
2	leavesDa10	1 860	86.0
3	leavesDa50	9 300	97.2
4	leavesDaHap	4 940	96.3
5	leavesDaIsap5	4 584	96.1
6	leavesDaIsap0	1 973	95.8

从表 9 中实验结果可以看到:leavesDa10 与 leavesDaIsap0 的规模相近,在经过 ISAP 算法采样约简的数据集 leavesDaIsap0 上训练学得的模型识别率远好于在使用简单数据增强手段的 leavesDa10 上训练学得的模型识别率.leavesDa50 是基本数据增强扩充 50 倍后的数据集,其规模大致是 leavesDaIsap0 的 5 倍.用 leavesDaIsap0 训练得到的模型的识别率与用 leavesDa50 训练得到的模型的识别率相差已不大.在经过 HAP 算法采样约简的数据集 leavesDaHap 上训练得到的模型识别率接近在 leavesDa50 上学得的模型,其数据规模只有 leavesDa50 的 1/2 左右.但是相比于 ISAP 算法的采样,HAP 算法的采样并不占优势.因为 HAP 算法无法有效获得可控数量的

代表点集,ISAP 算法在调整参数的情况下可以控制输出代表点的规模.而在最终样本规模相近的情况下,ISAP 算法数据增强策略相对于 HAP 算法数据增强策略获得的模型识别率也较为接近.而在实际使用时,由于 ISAP 算法计算效率显著较高,显然具有更高的实用价值.

综上表明,数据增强手段结合进高效增量采样处理后,在不改变总训练数据集规模的情况下,ISAP 算法介入所获得的模型质量可实现显著的提升.且 ISAP 算法能够控制约简后数据集的规模,有效地在减小数据规模的同时,保证数据集的多样性;在保持数据集规模不变的情况下有效提升数据质量,增加样本多样性.此外,因为 ISAP 高效的处理速度,可以快速地处理更大规模的数据增强数据集,更好地满足现实应用需求.

6 结论与展望

本文针对数据集代表点采样的一般性问题,提出了一种基于动态赋权近邻传播的数据增量采样算法 ISAP. 算法通过引入分层增量处理和样本点动态赋权策略,结合偏向参数动态赋权的 AP 算法,有效地实现了处理效率和采样质量的兼顾,更好地满足大规模数据集上的高效代表点选择.设计实验分别使用人工数据集、UCI 标准数据集和图像数据集进行性能分析,与其他方法相比,ISAP 算法在获得了采样划分质量与其他方法处于同一水平的同时,计算效率获得了大幅提升.进一步将 ISAP 算法应用于深度学习的数据增量任务中,相应实验结果表明:基本数据增强策略结合进高效处理的 ISAP 算法后,在不改变总训练数据集规模的情况下增加了样本的多样性,在保留样本多样性的同时约简了训练数据集的规模,新数据增强后所获得的模型质量可实现显著的提升.

在下一阶段,我们将从以下几个方面进行尝试.

(1) 本文中使用的数据规模与实际情况可能会面临的数据规模相比,规模还不够大.当数据规模扩大到一层增量局部推选加一层合并推选的“1+1”模式的采样无法处理时,研究如何将该方法扩充至 n 层增量局部推选加一层最终合并推选的“n+1”模式的采样;

(2) 类似于标准的 AP 算法,本文方法还不能很好地适应于类规模差别较大的数据集,在不平衡数据集上的采样效果不太理想.对算法过程作何改进,能够使其适用于不平衡数据集,是一个值得思考的问题;

(3) 作为一种同步约简的增量式采样算法,关于其中理论性能的分析研究还不够深入,这也将在我们的后续研究中进一步展开.

References:

[1] Liu XB, Zhang B. Automatic collecting representative logo images from the Internet. Tsinghua Science and Technology, 2013, 18(6):606−617. [doi: 10.1109/TST.2013.6678906]

[2] Kennedy L, Naaman M. Generating diverse and representative image search results for landmarks In: Proc. of the 17th Int'l Conf. on World Wide Web. 2008. 297−306. [doi: 10.1145/1367497.1367539]

[3] Kim EY, Ko E. Generating summaries for photographic images based on human affects. In: Proc. of the IEEE 14th Int'l Conf. on Cognitive Informatics & Cognitive Computing (ICCI*CC). 2015. 360−367. [doi: 10.1109/ICCI-CC.2015.7259411]

[4] Samani ZR, Moghaddam ME. A knowledge-based semantic approach for image collection summarization. Multimedia Tools and Applications, 2017,76(9):11917−11939. [doi: 10.1007/S11042-016-3840-1]

[5] Yang CL, Shen JL, Peng JY, *et al*. Image collection summarization via dictionary learning for sparse representation. Pattern Recognition, 2013,46(3):948−961. [doi: 10.1016/J.PATCOG.2012.07.011]

[6] Zhao Y, Hong RC, Jiang JG. Visual summarization of image collections by fast RANSAC. Neurocomputing, 2016,172(172):48−52. [doi: 10.1016/J.NEUCOM.2014.09.095]

[7] Qi MB, Zhu JJ, Ji P, *et al*. Representative image selection from image dataset. Acta Automatica Sinica, 2014,40(4):706−712 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2014.00706]

[8] Xue Y, Qian XM. Visual summarization of landmarks via viewpoint modeling. In: Proc. of the 19th IEEE Int'l Conf. on Image Processing (ICIP). 2012. 2873−2876. [doi: 10.1109/ICIP.2012.6467499]

[9] Li H, Peng SF, Samet H. Streaming news image summarization. In: Proc. of the 23rd Int'l Conf. on Pattern Recognition (ICPR). 2016. 1279−1284. [doi: 10.1109/ICPR.2016.7899813]

[10] Xu H, Wang JD, Hua XS, *et al*. Hybrid Image Summarization. In: Proc. of the 19th ACM Int'l Conf. on Multimedia. 2011. 1217−1220. [doi: 10.1145/2072298.2071978]

[11] Zhang YP, Zhang H. Image clustering based on SIFT-affinity propagation. In: Proc. of the 11th Int'l Conf. on Fuzzy Systems and Knowledge Discovery (FSKD 2014). 2014. 358−362. [doi: 10.1109/FSKD.2014.6980860]

[12] Frey BJ, Dueck D. Clustering by passing messages between data points. Science, 2007,315(5814):972−976. [doi: 10.1126/SCIENCE.1136800]

[13] Li KH. Reservoir-sampling algorithms of time complexity O(n(1+log(N/n))). ACM Trans. on Mathematical Software, 1994,20(4): 481−493. [doi: 10.1145/198429.198435]

[14] Babcock B, Datar M, Motwani R. Sampling from a moving window over streaming data. In: Proc. of the 13th Annual ACM-SIAM Symp. on Discrete Algorithms. 2002. 633−634. [doi: 10.1145/545381.545465]

[15] Shang FH, Jiao LC, Shi JR, *et al*. Fast affinity propagation clustering: A multilevel approach. Pattern Recognition, 2012,45(1): 474−486. [doi: 10.1016/J.PATCOG.2011.04.032]

[16] Sun LL, Guo CH. Incremental affinity propagation clustering based on message passing. IEEE Trans. on Knowledge and Data Engineering, 2014,26(11):2731−2744. [doi: 10.1109/TKDE.2014.2310215]

[17] Li P, Ji HF, Wang BL, *et al*. Adjustable preference affinity propagation clustering. Pattern Recognition Letters, 2017,85(85):72−78. [doi: 10.1016/J.PATREC.2016.11.017]

[18] Zhang J, He MY, Dai YC. Modified affinity propagation clustering. In: Proc. of the 2014 IEEE China Summit & Int'l Conf. on Signal and Information Processing (ChinaSIP). 2014. 505−509. [doi: 10.1109/CHINASIP.2014.6889294]

[19] Weiss Y, Freeman WT. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. IEEE Trans. on Information Theory, 2001,47(2):736−744. [doi: 10.1109/18.910585]

[20] Wang CD, Lai JH, Suen CY, *et al*. Multi-exemplar affinity propagation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013,35(9):2223−2237. [doi: 10.1109/TPAMI.2013.28]

[21] Yang HF, Liu Y, Xie ZP, *et al*. Efficiently training ball vector machine in online way. Journal of Computer Research and Development, 2013,50(9):1836−1842 (in Chinese with English abstract).

[22] Xie ZP, Sun J, Palade V, *et al*. Evolutionary sampling: A novel way of machine learning within a probabilistic framework. Information Sciences, 2015,299:262−282. [doi: 10.1016/J.INS.2014.12.001]

[23] Li XN, Yin MJ, Li MT, *et al*. Hierarchical affinity propagation clustering for large-scale data set. Computer Science, 2014,41(3): 185−188 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-137X.2014.03.040]

[24] Dyk DAV, Meng XL. The art of data augmentation. Journal of Computational and Graphical Statistics, 2001,10(1):1−50. [doi: 10.1198/10618600152418584]

[25] Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 2017,24(3):279−283. [doi: 10.1109/LSP.2017.2657381]

[26] Cui XD, Goel V, Kingsbury B. Data augmentation for deep convolutional neural network acoustic modeling. In: Proc. of the 2015 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). 2015. 4545−4549. [doi: 10.1109/ICASSP.2015.7178831]

附中文参考文献:

[7] 齐美彬,朱俊俊,纪平,等.大规模图像集中的代表性图像选取.自动化学报,2014,40(4):706−712. [doi: 10.3724/SP.J.1004.2014.00706]

[21] 杨海峰,刘渊,谢振平,等.球向量机的快速在线学习.计算机研究与发展,2013,50(9):1836−1842.

[23] 刘晓楠,尹美娟,李明涛,等.面向大规模数据的分层近邻传播聚类方法.计算机科学,2013,41(3):185−188. [doi: 10.3969/j.issn.1002-137X.2014.03.040]

陈晓琪(1994－),女,硕士生,主要研究领域为大数据知识发现.

刘渊(1967－),男,教授,博士生导师,CCF 高级会员,主要研究领域为数字媒体,网络安全.

谢振平(1977－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为知识建模,认知计算,智能系统软件.

詹千熠(1989－),女,博士,副教授,CCF 专业会员,主要研究领域为数据挖掘,社交网络分析.