文章编号：2095-4980(2021)02-0308-05

基于改进 AdaBoost 的密度峰值聚类法

王伟光，刘绍翰，胡　文，李梦霞

(南京航空航天大学 电子信息工程学院，江苏 南京 211106)

摘　要：针对雷达数据集中目标和杂波点迹的聚类不平衡问题，提出一种基于改进AdaBoost的密度峰值聚类法。介绍密度峰值聚类法的思想，基于不对称误分代价改进AdaBoost的误差函数，提高正类错分代价权重，将改进AdaBoost和密度峰值聚类结合，对由目标和杂波点迹组成的不平衡雷达数据集聚类。仿真实验结果表明，该算法在保证总体聚类性能的同时提高对正类的识别。

关键词：不平衡数据；目标和杂波点迹；AdaBoost 算法；密度峰值聚类

中图分类号：TN951　　**文献标志码：**A　　**doi：**10.11805/TKYDA2019407

Density Peaks Clustering based on optimized AdaBoost algorithm

WANG Weiguang，LIU Shaohan，HU Wen，LI Mengxia

(College of Electronic and Information Engineering，Nanjing University of Aeronautics & Astronautics，Nanjing Jiangsu 211106，China)

Abstract：A Density Peaks Clustering algorithm based on optimized AdaBoost-DPC is proposed to deal with the class-imbalanced question of target and clutter in radar data sets. The method of density peaks clustering is introduced, and the clutter dots are under-sampled. The error function of AdaBoost algorithm is improved based on the asymmetric misclustering cost, which raises the weight of positive misclassification cost. Then the improved AdaBoost algorithm is combined with density peaks clustering method to cluster the imbalanced radar data sets consisting of the targets and clutter dots. The experimental results show that the optimized method can effectively improve the identification of target.

Keywords：imbalanced data；targets and clutter clustering；AdaBoost algorithm；density peaks clustering

不平衡数据集是指数据集中包含不同的类，某些类的数量远远大于其他类别的数量，其中类别数量多的为多数类，反之为少数类[1-2]。在二聚类问题中，少数类和多数类也被称为正类和负类。现实应用领域中，不平衡数据集广泛存在，如雷达点迹信息、医疗诊断信息、网络入侵检测数据等[3]。在雷达数据中，杂波点迹数量众多，探测目标是人们关注的重点，需要通过聚类实现杂波抑制。传统的聚类方法可以取得较高的聚类精确度，往往是通过将少数类误分到多数类获得，实际中对少数类聚类效果并不理想。对不平衡数据集的挖掘问题是机器学习领域的热点问题，对于雷达杂波抑制具有重要意义[4]。

国内外学者针对不平衡数据集进行大量研究，目前，不平衡数据处理方法主要有 2 个方面：第 1 种是算法层面，通过设计新算法或考虑不同误分情况代价的差异性优化原有算法，使之更适用于不平衡数据，如代价敏感学习、集成学习、主动学习等[5-6]；第 2 种是数据层面，通过样本重采样优化数据集，降低其不平衡度，主要包括：过采样、欠采样和两者相结合的方法[7]。过采样通过不断复制少数类使数据规模变大，却并未增加有效的新信息，从而容易导致过拟合[8]。欠采样由于抽取部分多数类样本，重要信息样本容易删除。改进的数据采样方法，如少数过采样技术 SMOTE 算法，通过采用少数类样本合成技术产生新的样本，该算法使少数类样本增加，调节不平衡度有限，会产生噪音样例或者边际样例[9-11]。在雷达领域，聚类大多用于 SAR 图像中进行目标识别，针对点迹的不平衡数据聚类方法具有较大研究空间[12]。本文提出一种基于改进 AdaBoost 的密度峰值聚类方法，对雷达数据集中的目标和杂波点迹聚类，区分二者，实现智能化的雷达处理模式。本文中，数据集是由目标和杂波点迹构成，目标为正类是关注的重点，杂波为负类，正负类的类别标签取值分别为 {+1,-1}。本文提出的 AdaBoost-DPC 方法结合数据和算法方面进行优化，借鉴代价敏感学习，对 AdaBoost 错分的正类样本赋予更大的错分代价，修改各个基聚类模型的输出决策权重，最终得到集成聚类模型。

收稿日期：2019-10-20；**修回日期：**2019-12-29

作者简介：王伟光(1994–)，男，在读硕士研究生，主要研究方向为雷达信号与数据处理。email:13813018269@163.com

1 密度峰值聚类法

1.1 算法思想

聚类关键步骤是根据不同的定义或者概念确定聚类中心，密度峰值聚类法提出一种新的选取聚类中心的方法。聚类中心的特征：聚类中心被比其局部密度低的数据点包围，这些中心点与其他比其局部密度高的任何点之间的距离较大，分别用局部密度和相对距离来表示这 2 个特征，选择二者数值相对较大的点作为聚类中心，归并其他非聚类中心点，剔除噪声点，从而实现聚类[13]。

1.2 算法原理

考虑本文的点迹数据集 $\boldsymbol{X}=\{x_i \mid i=1,2,\cdots,N\}$，相应的下标集合 $I_{\boldsymbol{X}}=\{1,2,\cdots,N\}$，则数据点 x_i 和 x_j 的距离为 d_{ij}：

$$d_{ij}=dist(x_i,x_j)=\left\|x_i-x_j\right\|_2 \tag{1}$$

基于上述 2 个特征，对于每个数据点 i，提出了 2 个变量：局部密度 ρ_i、到较高密度点的最小距离 δ_i。这 2 个量取决于数据点两两距离 d_{ij}。局部密度计算方法：

1) 截断核

$$\rho_i=\sum\nolimits_j \chi(d_{ij}-d_{\mathrm{c}}), j\in I_{\boldsymbol{X}}, \text{且} j\neq i \tag{2}$$

式中 d_{c} 为截断距离，需事先指定。其中函数：

$$\chi(x)=\begin{cases}1, & x<0\\ 0, & x\geqslant 0\end{cases} \tag{3}$$

ρ_i 代表除了 i 点外，其他所有样本点和 i 点距离在 d_{c} 之内的样本点的数量。

2) 高斯核

$$\rho_i=\sum\nolimits_j \mathrm{e}^{-\left(\frac{d_{ij}}{d_{\mathrm{c}}}\right)^2},\quad j\in I_{\boldsymbol{X}}, \text{且} j\neq i \tag{4}$$

由式(2)~式(3)可得，截断核处理数值离散型数据和高斯核处理数值连续型数据。若 $i\neq j$，但 $\rho_i=\rho_j$，该现象称为数据点冲突。通过高斯核计算局部密度产生冲突的概率较截断核较小。本文选取高斯核来计算局部密度值。

最短距离 δ_i 表示点 i 和其他具有更高密度的点之间的最小距离，最小距离对应的数据点被称为数据点 i 的最近邻点。计算如下：

$$\delta_i=\min_{j:\rho_j>\rho_i}(d_{ij}) \tag{5}$$

设 $\{q_i\}_{i=1}^{s}$ 表示 $\{\rho_i\}_{i=1}^{s}$ 的一个降序排列下标序，满足 $\rho_{q_1}\geqslant\rho_{q_2}\cdots\geqslant\rho_{q_s}$，式(5)进一步转换为：

$$\delta_{q_i}=\begin{cases}\min\limits_{j<i,q_j}\{d_{q_iq_j}\}, & i\geqslant 2\\ \max\limits_{j\geqslant 2}\{\delta_{q_j}\}, & i=1\end{cases} \tag{6}$$

当数据点 x_i 具有最大值密度时，δ_i 是 $\boldsymbol{X}$ 中与 x_i 距离最大的数据点与其之间的距离，即 $\delta_i=\max_j(d_{ij})$。式(11)的定义可以很好地避免将一个簇类拆成 2 个的情况，在 ρ 值相等的情况下，经过降序排列之后，数据点 x_i 和 x_j 有先后顺序，保证了聚类中心只取到排在前面的数据点。对于数据集中的每一个数据点 x_i，都有相应的二元对 (ρ_i,δ_i)。可以把数据集中的所有二元对 $\{(\rho_i,\delta_i)\mid i=1,2,\cdots,N\}$ 在平面上以 ρ 为横轴，δ 为纵轴画出来，该图称之为决策图，聚类中心是同时具有较大 ρ 和 δ 的数据点，直观上可以根据决策图判断选择聚类中心。为了使该过程更加自动化，清晰并正确地确定聚类中心，综合考虑 ρ_i 和 δ_i，将其乘积作为指标：

$$\gamma_i=\rho_i\delta_i,\ i\in I_X \tag{7}$$

显然，γ 值越大，越有可能是聚类中心。只需要对其进行降序排列，并把排序后的 γ 在坐标平面内以下标为横轴，值为纵坐标画出，选取 γ 值较大的 2 个数据点作为聚类中心。

2 基于改进 AdaBoost 算法的密度峰值聚类法

2.1 基于误分代价的 AdaBoost 算法

AdaBoost 算法原理：开始时用相同的权重初始化数据集中的每一个样本，通过一轮聚类算法，得到某个聚

类模型，之后根据该模型结果重点关注被错误聚类的样本，增加其权重，对于已经被正确聚类的样本，减少其权重，经过不断的迭代，最终得到的结果是每个模型的加权统计值，这样使得最终聚类算法的性能更优。

考虑给定的样本集 $\boldsymbol{X}=\{x_i \mid i=1,2,\cdots,N\}$，$w_{ti}$ 表示第 t 轮迭代第 i 个样本点的权重，第一轮初始化权重 $w_{1i}=1/N$，第 t 轮的误差定义为：

$$\varepsilon_t=\sum_{i=1}^{N} w_{ti}\cdot[\![y_{\mathrm{cl}}(t,i)\neq y_{\mathrm{label}}(i)]\!],\ i=1,2,\cdots,N \tag{8}$$

式中：$[\![\zeta]\!]$为满足 ζ 的条件下输出 1，否则输出为 0；$y_{\mathrm{cl}}(t,i)$ 为经过第 t 轮得到的样本点 i 所属的簇类；$y_{\mathrm{label}}(i)$ 为第 i 个样本点的实际簇类，式(8)表示将聚类结果和实际值不同的样本点的权重求和。当 $\varepsilon_t=0$ 或者 $\varepsilon_t\geqslant 0.5$ 时停止迭代过程。每一轮聚类后得到模型 Cl_t，下一轮即 $t+1$ 轮迭代中，按照式(9)更新样本权重：

$$w_{(t+1)i}=\frac{w_{ti}\exp[-\alpha_t y_{\mathrm{label}}(i)y_{Cl}(t,i)]}{Z_t},\ i=1,2,\cdots,N \tag{9}$$

式中 α_t 是第 t 轮模型 Cl_t 的决策权重，表示为：

$$\alpha_t=\log\frac{1-\varepsilon_t}{\varepsilon_t} \tag{10}$$

Z_t 为归一化常数：

$$Z_t=\sum_{i=1}^{N} w_{ti}\exp[-\alpha_t y_{\mathrm{label}}(i)y_{\mathrm{cl}}(t,i)] \tag{11}$$

最后的结果是每一轮迭代模型的加权和：

$$output=sign\left[\sum_{t=1}^{T}\alpha_t y_{\mathrm{cl}}(t,i)\right] \tag{12}$$

AdaBoost 算法决策过程见图 1，$\boldsymbol{X}$ 是样本集，$w_1,w_2,\cdots,w_{T-1},w_T$ 代表每一轮样本的权重，共 T 轮迭代，$Cl_1,Cl_2,\cdots,Cl_{T-1},Cl_T$ 代表每一轮迭代完得到的聚类模型，Cl_{output} 标识最终输出的判别结果。

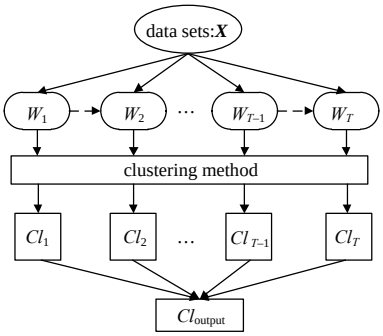


Fig.1 Decision-making process based on AdaBoost
图 1 基于 AdaBoost 的决策过程

本文的实测数据集中目标点迹的数量远不如杂波点迹，提高总体的识别率并不意味着提高了目标的识别。通过总体误分情况更新权重不可取。本文通过引入对每一簇类的误分代价，对上述算法中式(8)计算的误差进行修改，更加关注被错误聚类的正类样本，增加其误分代价，并以此来更新样本权重。假设原始训练集的不平衡度 IR 为：

$$IR=S_{\mathrm{neg}}/S_{\mathrm{pos}} \tag{13}$$

式中 S_{pos} 和 S_{neg} 分别代表原始数据集中目标和杂波的样本数目。将式(8)修改为：

$$\varepsilon_t=\sum_{i=1}^{N} IR^{\frac{1+y_{\mathrm{label}}(i)}{2}} w_{ti}\cdot[\![y_{Cl}(t,i)\neq y_{\mathrm{label}}(i)]\!],\quad i=1,2,\cdots,N \tag{14}$$

在每一轮迭代时，每个样本点的误分代价多乘一个系数 $IR^{\frac{1+y_{\mathrm{label}}(i)}{2}}$，若样本为负类时，则 $y_{\mathrm{label}}(i)=-1$，那么系数 $IR^{\frac{1+y_{\mathrm{label}}(i)}{2}}=1$，误分代价与原来的误分代价相同；若样本为正类时，则 $y_{\mathrm{label}}(i)=+1$，那么系数 $IR^{\frac{1+y_{\mathrm{label}}(i)}{2}}=IR$，此时的误分代价需要乘以不平衡度 IR。由上述分析可知，若正类样本误分较多，其误分代价也会随之变大，则最终输出的权重减小，提高对正类目标的识别。

2.2 改进 AdaBoost 的密度峰值聚类

本节将 2.1 节中介绍的改进 AdaBoost 算法和密度峰值聚类结合，算法总体流程如图 2 所示。考虑样本集 $\boldsymbol{X}=\{x_i \mid i=1,2,\cdots,N\}$，将式(1)计算数据点的距离修改为：

$$d_{ij}=dist(s_i,s_j)=\left\|w_i s_i-w_j s_j\right\|_2=\sqrt{\sum_{k=1}^{m}\left(s_{ik}w_i-s_{jk}w_j\right)^2},\quad i=1,2,\cdots,N \tag{15}$$

根据 d_{ij} 按照密度峰值聚类法实施聚类，并进入迭代的过程，具体的迭代流程如图 2 所示。

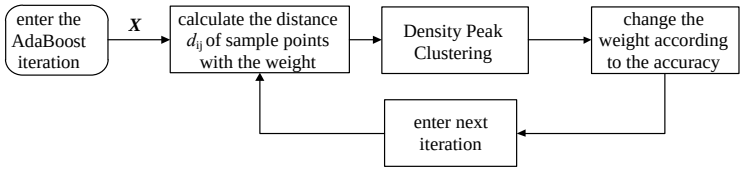


Fig.2 Iteration process of density peak clustering based on advanced AdaBoost
图 2 基于改进 AdaBoost 的密度峰值聚类方法迭代过程

2.3 评价指标

对于不平衡样本集，需要关注正类样本聚类性能，平衡样本集的准确率等评价指标对于不平衡数据意义不大。针对不平衡数据，本文采用建立在匹配矩阵上的 F-measure 和 G-mean 作为聚类评价指标，见表 1。

表 1 聚类结果匹配矩阵

Table1 Matching matrix for clustering results

actual class	class of prediction	
	minority class	majority class
minority class	*TP*	*FP*
majority class	*FP*	*TP*

F-measure 定义为：

$$F\text{-}measure=\frac{2\times recall\times precision}{recall+precision} \tag{16}$$

式中：*recall* 为查全率；*precision* 为精确率。分别表示为：

$$recall=\frac{TP}{TP+FN},\quad precision=\frac{TP}{TP+FP} \tag{17}$$

F-measure 是查全率和查准率的加权调和均值，衡量的是少数类样本的聚类效果。G-mean 综合考虑样本集正负类聚类效果，反映了样本集的整体聚类性能：

$$G\text{-}mean=\sqrt{acc^{+}\times acc^{-}} \tag{18}$$

式中 acc^{+} 和 acc^{-} 分别表示正类和负类被正确归类的比值。

3 实验结果和分析

3.1 数据集

表 2 实测数据集基本参数

Table2 Basic parameters of measured data set

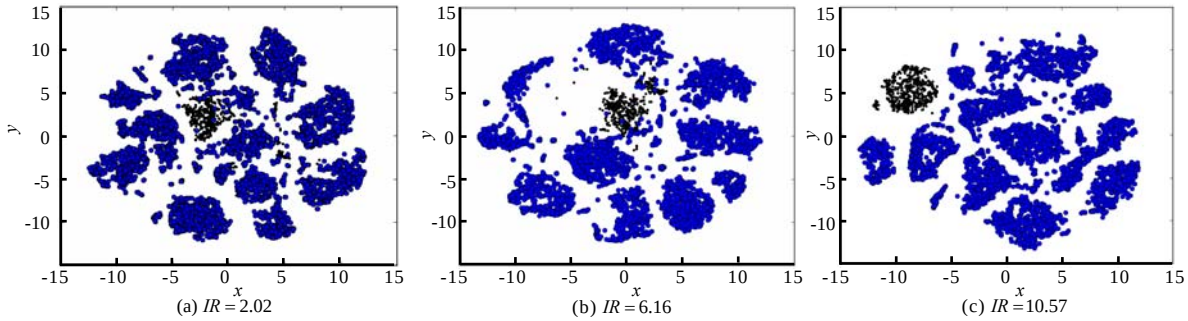
minority class N^{+}	majority class N^{-}	degree of unbalance
1 655	3 345	2.02
698	4 302	6.16
432	4 568	10.57

本文的数据集由某型低慢小目标探测雷达外场试验获得。通过信号处理后得到原始点迹 $x=(dis,azim,elev,v,E_{\Sigma},E_{azim},E_{elev},E_{pro})$，其中 dis 为距离，$azim\in[0,2\pi)$ 为方位角，$elev\in[0,\pi/2]$ 为俯仰角，v 为速度，$E_{\Sigma},E_{azim},E_{elev},E_{pro}$ 分别为 4 个通道的各自能量。故样本集为 $\boldsymbol{X}=\{x_i\,|\,i=1,2,\cdots,N\}$，其中 $x_i=\left(dis,azim,elev,v,E_{\Sigma},E_{azim},E_{elev},E_{pro}\right)$。

不平衡度可分为低度、中度和高度不平衡，其范围分别为 $[1.5,3.5)$，$[3.5,9.5)$，$[9.5,+\infty)$。选用不同平衡度下的实测数据集，样本点个数为 5 000，见表 2。

3.2 结果分析

利用上述介绍的 AdaBoost–DPC 方法聚类，不同不平衡度下的聚类结果如图 3 所示，其中 x，y 轴为高维数据降维处理后的特征维度，黑色点迹为目标点，蓝色点迹为杂波点，在高度不平衡条件下本方法对目标和杂波有较理想的聚类效果。



(a) $IR=2.02$ (b) $IR=6.16$ (c) $IR=10.57$

Fig.3 Clustering results in different IRs
图 3 不同不平衡度下的聚类结果

为了评估改进算法的性能，在不同不平衡度下，分别用 F-measure 和 G-mean 评价改进前后的算法对少数类的识别和整体聚类性能，如图 4~图 5 所示。由图 4 可以看出，在少数类目标识别的评价 F-measure 方面，改进的算法具有明显的优势，中低平衡度下，提升了 1%~2%，特别是在不平衡度较高的情况下，有 5%的精确度提升。由图 5 可以看出，对数据集整体聚类效果的 G-mean 值，在低度和中度不平衡条件下，改进的算法在整体上聚类效果达到最优，在高度不平衡条件下，AdaBoost-DPC 算法的 G-mean 值低于原始算法。改进的 AdaBoost-DPC 算法在严重不平衡的数据集上，通过牺牲整体的聚类性能 G-mean 来提升正类目标的识别 F-measure。

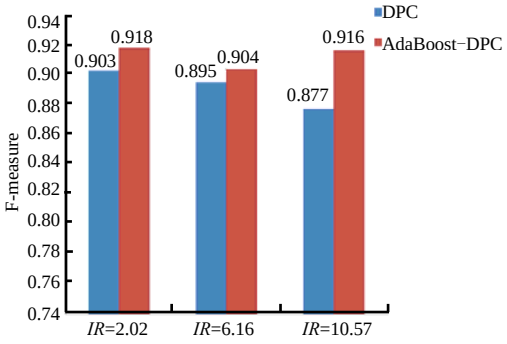


Fig.4 F-measure value in different IRs
图 4　不同不平衡度下的 F-measure 值比

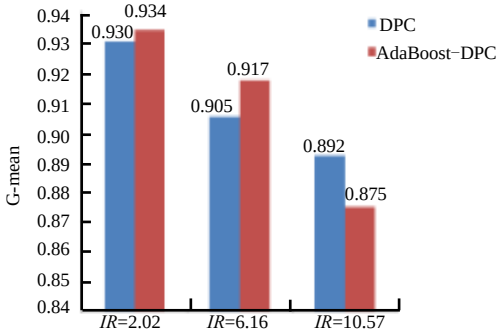


Fig.5 G-mean value in different IRs
图 5　不同不平衡度下的 G-mean 值比较

4　结论

本文针对样本集类别不平衡问题，借鉴代价敏感学习优化 AdaBoost 算法，提出了一种基于改进 AdaBoost 的密度峰值聚类法。引入不对称误分代价，赋予正类样本更大的误分代价，改进 AdaBoost 算法，更符合实际应用。根据每一轮样本的权重计算样本点距离，进行密度峰值聚类。最后，通过 2 个指标 F-measure 和 G-mean 对聚类效果评价，对比改进前后的聚类性能，验证了改进算法对少数类样本目标识别的提升效果。

参考文献：

[1] XU D,TIAN Y. A comprehensive survey of clustering algorithms[J]. Annals of Data Science, 2015,2(2):165-193.
[2] MATHEWS L,HARI S. Learning from imbalanced data[M]. Commonwealth of Pennsylvania,USA:IGI Global, 2018.
[3] 李军. 不平衡数据学习的研究[D]. 长春:吉林大学, 2011. (LI Jun. Research on the imbalanced data learning[D]. Changchun, China:Jilin University, 2011.)
[4] 向鸿鑫,杨云. 不平衡数据挖掘方法综述[J]. 计算机工程与应用, 2019,55(4):1-16. (XIANG Hongxin,YANG Yun. Survey on imbalanced data mining methods[J]. Computer Engineering and Applications, 2019,55(4):1-16.)
[5] LIN W C,TSAI C F,HU Y H,et al. Clustering-based under-sampling in class-imbalanced data[J]. Information Sciences, 2017(409):17-26.
[6] KUMAR G. A survey on clustering-a data mining technique[Z]. 2015.
[7] RIVERA W A. Noise reduction a priori synthetic over-sampling for class imbalanced data sets[J]. Information Sciences, 2017(408):146-161.
[8] ABDI L,HASHEMI S. To Combat multi-class imbalanced problems by means of over-sampling techniques[J]. IEEE Transactions on Knowledge & Data Engineering, 2015,19(12):3369-3385.
[9] DEMIDOVA L,KLYUEVA I. SVM classification:optimization with the SMOTE algorithm for the class imbalance problem[C]// Mediterranean Conference on Embedded Computing(MECO). Bar:[s.n.], 2017:1-4.
[10] LIN C T,HSIEH T Y,LIU Y T,et al. Minority oversampling in kernel adaptive subspaces for class imbalanced datasets[J]. IEEE Transactions on Knowledge & Data Engineering, 2017(30):950-962.
[11] PAN Z,QIU X,HUANG Z,et al. Airplane recognition in terra SAR-X images via scatter cluster extraction and reweighted sparse representation[J]. IEEE Geoscience and Remote Sensing Letters, 2017,14(1):112-116. doi:10.1109/LGRS.2016.2628162.
[12] LI Caoyuan,DING Gangyi,WANG Dakui,et al. Clustering by fast search and find of density peaks with data field[J]. Chinese Journal of Electronics, 2016,25(3):397-402.
[13] ZHANG X,YAN Z,WEI W,et al. Transfer boosting with synthetic instances for class imbalanced object recognition[J]. IEEE Transactions on Cybernetics, 2016,39(99):1-14.