Kernel ADASYN

Kernel Based Adaptive Synthetic Data Generation for Imbalanced Learning

Hao Ding

Ocean University of China

May 25, 2018

Content

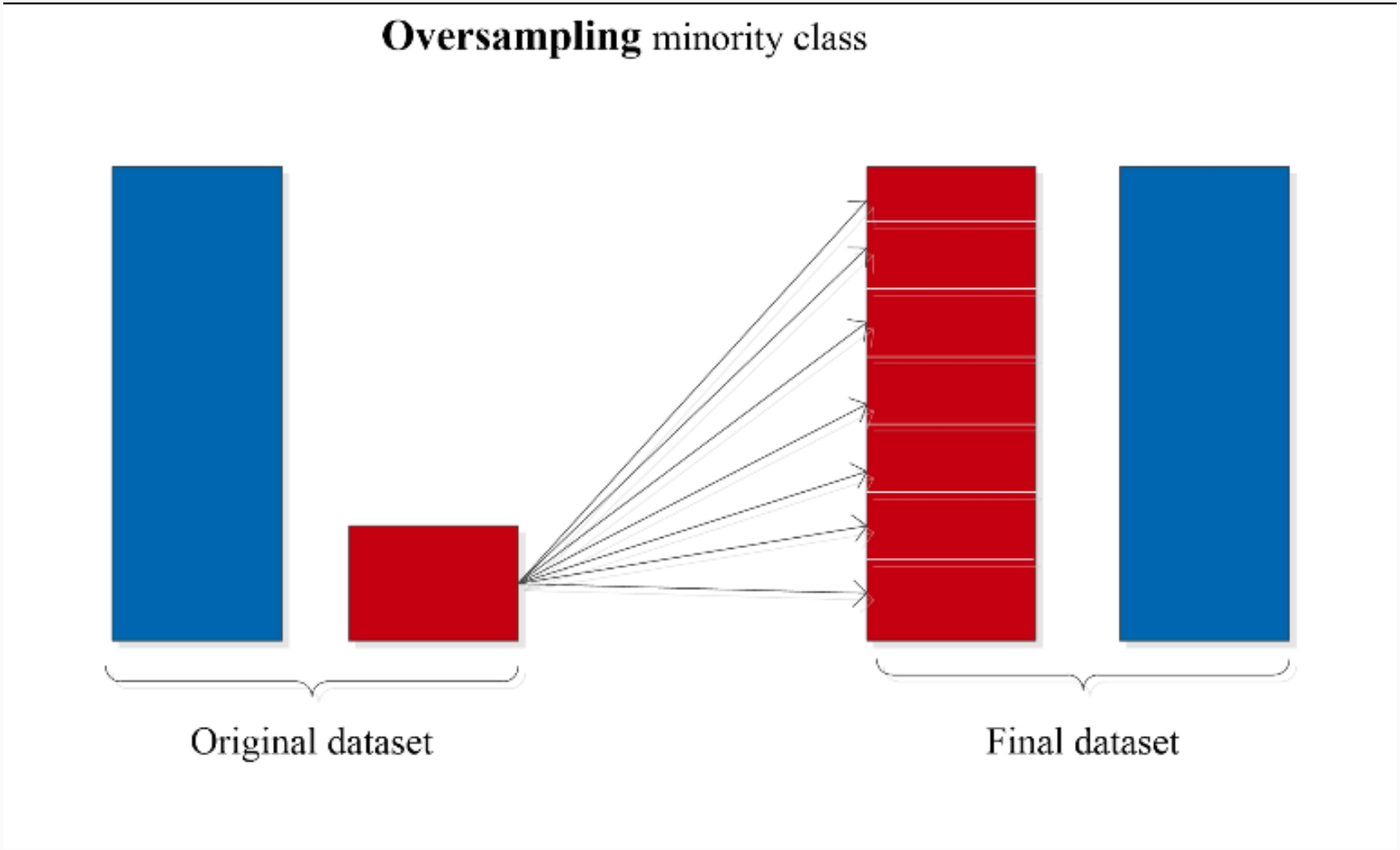
Over-sampling method

SMOTE

ADASYN

Kernal ADASYN

Over-sampling method



SMOTE

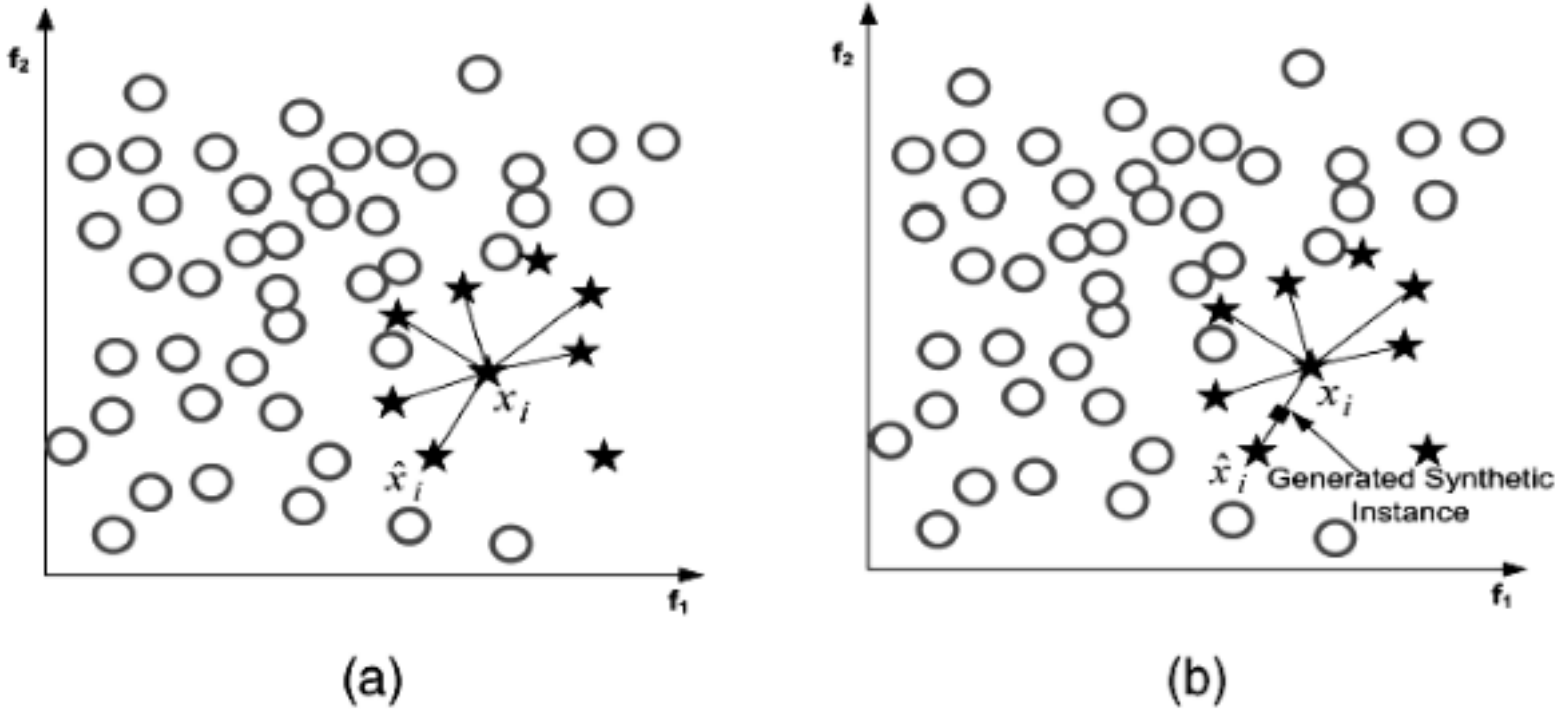


Fig. 3. (a) Example of the K-nearest neighbors for the $\boldsymbol{x_i}$ example under consideration ($K = 6$). (b) Data creation based on euclidian distance.

ADASYN

Calculate how many minority samples to generate ($N^+ * SR$).

For each minority sample $x_i^+, i = 1, 2, ..., N^+$, find its K-nearest neighbors, N_i^{maj} of which from the majority.

$\Gamma_i = \frac{\frac{N_i^{maj}}{K}}{Z}$, Z is a standardization factor to make sure $\sum \Gamma_i = 1$

$g_i = \Gamma_i * N^+ * SR$

ADASYN

Algorithm ADASYN.

Input: Training set $S = \{(x_i, y_i), i = 1, 2, ..., N, y_i \in \{+, -\}\}$, majority N^-, minority N^+, $N = N^+ + N^-$, SR, K

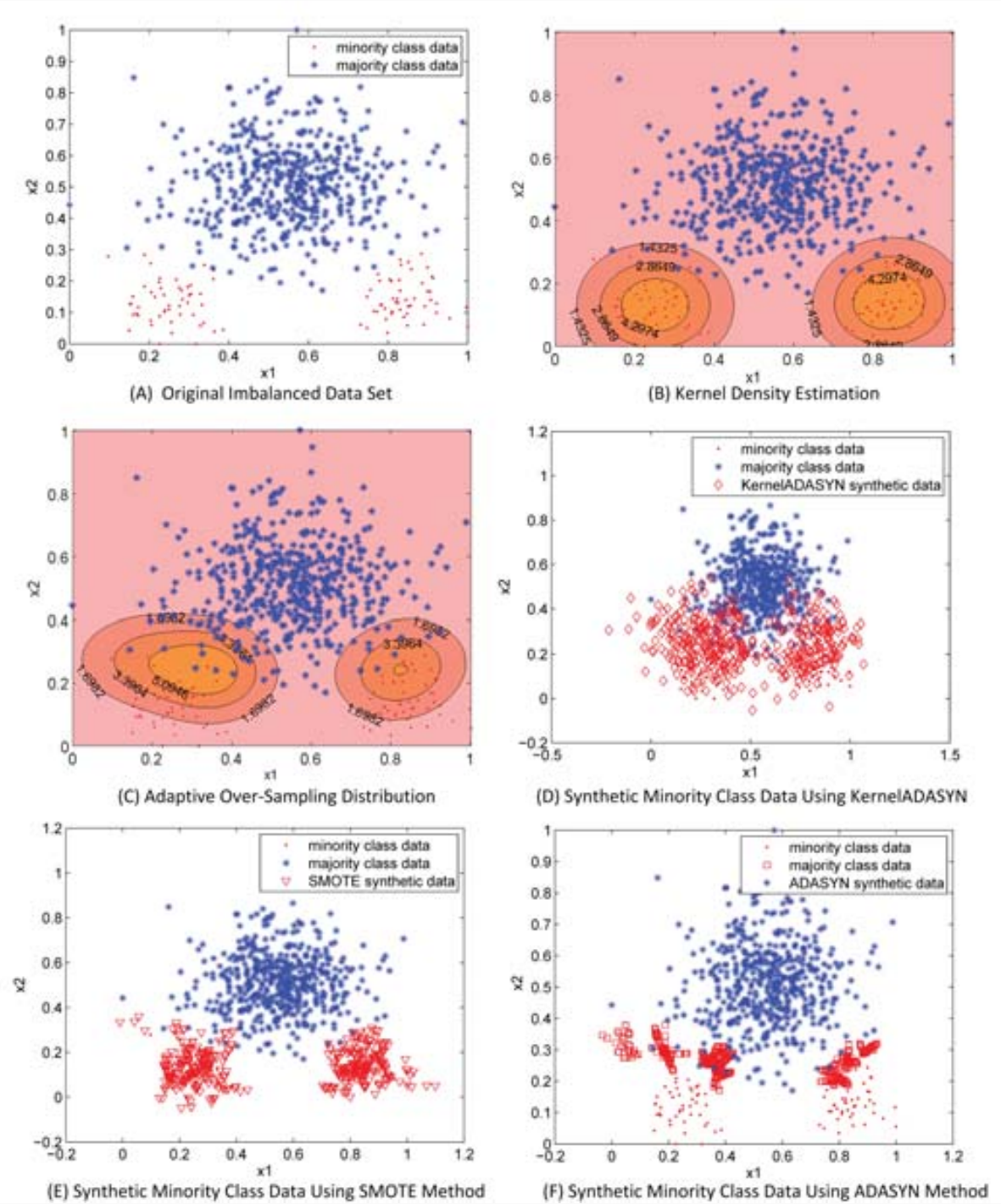
Output: $S' = \{(x_i, y_i), i = 1, 2, ..., N + N^+ * SR$ $SR, y_i \in \{+, -\}\}$

Kernal ADASYN

kernal density estimation:

$$\hat{p}(x) = \frac{1}{N+h} \sum_{i \in I_{+1}} \hat{r}_i \frac{1}{(\sqrt{2\pi}h)^n} exp\left(-\frac{1}{2}\frac{|x-x_i|}{h}\right)$$

Not only adaptively shifting the classification decision boundary toward the difficult examples.
But also construct an adaptive over-sampling distribution to generate synthetic minority class data.


minority class data
majority class data
x1
x2
(A) Original Imbalanced Data Set
(B) Kernel Density Estimation
(C) Adaptive Over-Sampling Distribution
minority class data
majority class data
KernelADASYN synthetic data
(D) Synthetic Minority Class Data Using KernelADASYN
minority class data
majority class data
SMOTE synthetic data
(E) Synthetic Minority Class Data Using SMOTE Method
minority class data
majority class data
ADASYN synthetic data
(F) Synthetic Minority Class Data Using ADASYN Method

Datasets	Methods	OA	Precision	Recall	F-Measure	G-Mean
Pima Indians Diabetes	Decision Tree	0.7049	**0.5803**	0.5642	0.5696	0.6612
	SMOTE	0.6750	0.5284	**0.6627**	0.5867	0.6707
	ADASYN	0.6799	0.5374	0.5881	0.5606	0.6535
	KernelADASYN	**0.7089**	0.5731	0.6560	**0.6112**	**0.6950**
ILPD	Decision Tree	0.6515	0.3858	0.3687	0.3741	0.5274
	SMOTE	0.6326	0.3797	0.4518	0.4116	0.5629
	ADASYN	0.6357	0.3841	0.4542	0.4139	0.5642
	KernelADASYN	**0.6571**	**0.4004**	**0.4598**	**0.4187**	**0.5722**
Parkinsons	Decision Tree	0.8237	0.6611	0.6417	0.6398	0.7478
	SMOTE	0.8361	0.6627	**0.7500**	0.6938	0.8012
	ADASYN	0.8216	0.6456	0.6958	0.6611	0.7723
	KernelADASYN	**0.8433**	**0.6672**	**0.7500**	**0.7000**	**0.8059**
Vertebral Column	Decision Tree	0.7903	**0.6711**	0.7040	0.6837	0.7631
	SMOTE	0.7845	0.6491	0.7480	0.6910	0.7720
	ADASYN	**0.8013**	0.6703	0.7800	**0.7195**	0.7944
	KernelADASYN	0.7935	0.6559	**0.7820**	0.7194	**0.7984**
Breast Cancer	Decision Tree	0.9367	0.9117	0.9058	0.9077	0.9287
	SMOTE	0.9476	**0.9171**	0.9333	0.9244	0.9439
	ADASYN	0.9401	0.9080	0.9200	0.9133	0.9349
	KernelADASYN	**0.9481**	0.9115	**0.9408**	**0.9258**	**0.9463**
Breast Tissue	Decision Tree	0.8962	0.7179	**0.8600**	0.7644	0.8766
	SMOTE	0.9115	**0.8530**	0.6900	0.7391	0.8074
	ADASYN	**0.9288**	0.8172	0.8400	**0.8196**	**0.8907**
	KernelADASYN	0.8788	0.7313	0.7200	0.7038	0.8070
SPECT	Decision Tree	0.8143	0.5605	0.5037	0.5270	0.6695
	SMOTE	0.7820	0.4759	0.5593	0.5105	0.6830
	ADASYN	0.7902	0.4978	**0.5595**	0.5227	**0.6880**
	KernelADASYN	**0.8174**	**0.5513**	0.5185	**0.5280**	0.6755
Winning Time	Decision Tree	0	2	1	0	0
	SMOTE	0	2	2	0	0
	ADASYN	2	0	1	2	2
	KernelADASYN	**5**	**3**	**4**	**5**	**5**

Questions

THANK YOU