KA-Ensemble

Towards Imbalanced Image Classification Ensembling Over-samling and Under-sampling

Hao Ding

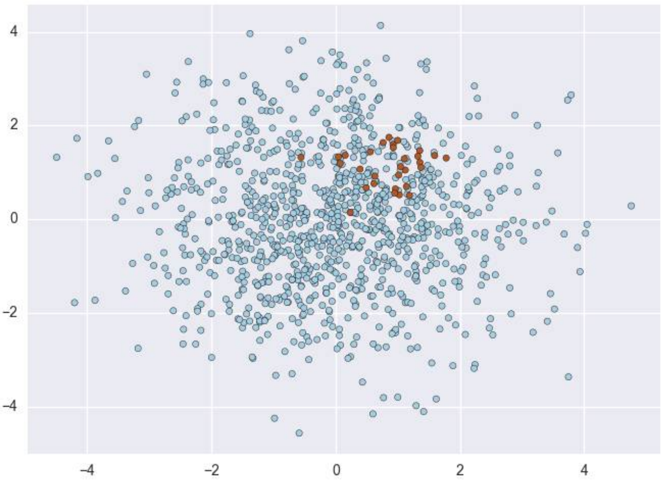
September 23, 2018

Ocean Univercity of China

Table of contents

Imbalanced learning

Imbalanced learning



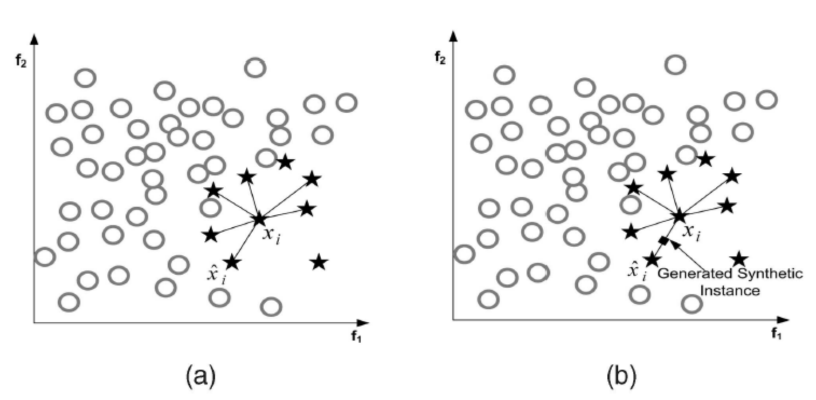
Sampling method

According to the different sort of samples, sampling methods can be roughly classified into three classes:

- over-sampling
- under-sampling
- hybrid-sampling

Method

SMOTE



f2
f1
x_i
$\hat{x}_i$
(a)
f2
f1
x_i
$\hat{x}_i$
Generated Synthetic Instance
(b)

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

Pros: Adaptively determine the frequency of each minority sample as the primary sample and focus the attention on the boundary regions of the minority class.

Cons: The anti-noise performance of the algorithm is poor, which will amplify the range of small-scale noise information to a certain extent, resulting in a decline in classification quality.

Kernal ADASYN

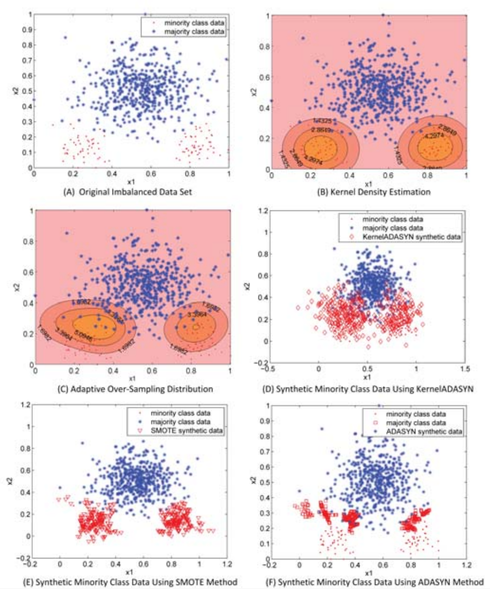
kernal density estimation:

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

Not only adaptively shifting the classification decision boundary toward the difficult examples.

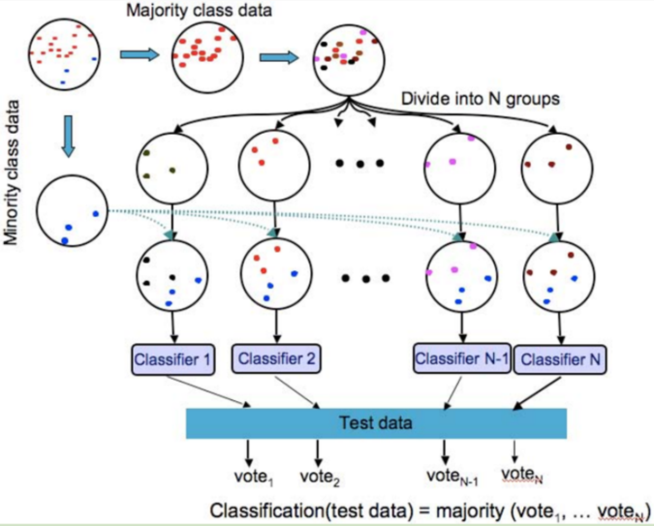
But also construct an adaptive over-sampling distribution to generate synthetic minority class data.

Kernal ADASYN

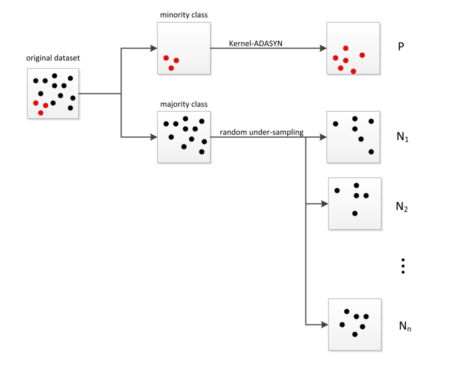


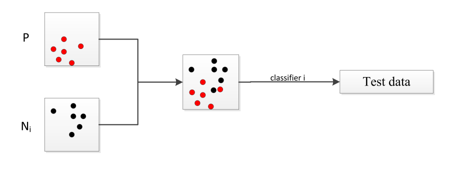
(A) Original Imbalanced Data Set
(B) Kernel Density Estimation
(C) Adaptive Over-Sampling Distribution
(D) Synthetic Minority Class Data Using KernelADASYN
(E) Synthetic Minority Class Data Using SMOTE Method
(F) Synthetic Minority Class Data Using ADASYN Method

EasyEnsemble



KA-Ensemble


minority class
Kernel-ADASYN
P
original dataset
majority class
random under-sampling
N_1
N_2
N_n


P
N_i
classifier i
Test data

Experiment

Datasets

	Sample	IR
balance	625	11.8
car	1728	3.5
Colon	62	1.82
cmc	1473	3.4
Glioma	50	2.57
haberman	306	2.8
mf-morph	2000	9.0
mf-zernike	2000	9.0
vehicle	846	3.0
ZooScan	2000	15.25

Evaluation Criteria

TN : true negatives FP : false positives

TP : true positives FN : false negatives

$Precision : p = \frac{TP}{TP+FP}$

$Recal : r = \frac{TP}{TP+FN}$

acc $= \frac{TP+TN}{TP+FN+TN+FP}$

G-means $= \sqrt{TPR * TNR}$

F-measure $= \frac{1}{\frac{1}{2}(\frac{1}{p}+\frac{1}{r})} = \frac{2pr}{p+r}$

$Micro - F = \frac{1}{k}\sum_{i=1}^{k} F_i$

AUC: Area under the ROC curve

SVM based on Gaussian radial basis kernel function

parameters of SVM	value
σ: the width of Gaussian radial basis kernel function	5
C: Penalty factor	500

parameters of ACOSampling	value
ant_n	50
ITA	50
ITP	100
dispose	0.8
ph_initial	1
ph_{min}	0.5
ph_{min}	0.5
α, β, γ	$\frac{1}{3}$

Result

Colon

	ORI	ROS	RUS	SMOTE	BSO1	BSO2
Acc	83.23	84.19	85.48	85.48	83.07	84.03
F-measure	75.24	76.78	79.31	79.37	74.99	76.91
G-mean	80.23	81.54	84.17	83.83	80.01	81.68
AUC	87.23	87.76	89.16	89.13	88.20	88.61
	OSS	**ADA-SYN**	**SBC**	**ACOSampling**	**Kernel-ADASYN**	**KA-Ensemble**
Acc	85.65	85.65	83.23	85.63	85.63	**86.03**
F-measure	81.13	79.76	78.95	81.13	83.02	**85.82**
G-mean	85.76	84.21	84.25	85.92	**86.10**	85.99
AUC	91.33	88.82	90.19	94.18	95.67	**97.25**

Glioma

	ORI	ROS	RUS	SMOTE	BSO1	BSO2
Acc	92.80	94.00	92.20	93.60	94.00	93.40
F-measure	87.08	89.35	87.54	88.56	89.32	88.80
G-mean	90.94	92.71	93.16	91.97	92.47	93.19
AUC	98.71	98.93	98.75	98.87	**99.15**	98.73

	OSS	ADA-SYN	SBC	ACOSampling	Kernel-ADASYN	KA-Ensemble
Acc	68.21	65.38	67.18	71.79	72.00	**75.21**
F-measure	62.50	54.79	60.43	67.86	67.84	**69.10**
G-mean	68.06	62.67	66.74	72.32	72.78	**73.25**
AUC	73.53	68.00	73.22	77.42	**79.02**	78.00

ZooScan

	ORI	ROS	RUS	SMOTE	BSO1	BSO2
Acc	53.23	64.19	55.48	55.48	63.07	64.03
F-measure	46.10	43.30	50.56	45.58	55.40	49.39
G-mean	53.46	51.48	56.89	53.35	52.93	56.16
AUC	57.75	57.36	61.22	57.92	58.14	59.78
	OSS	**ADA-SYN**	**SBC**	**ACOSampling**	**Kernel-ADASYN**	**KA-Ensemble**
Acc	58.16	56.25	55.48	66.12	66.66	**71.26**
F-measure	52.83	51.99	50.56	50.82	52.64	**57.33**
G-mean	54.36	58.37	56.89	53.89	60.09	**65.24**
AUC	57.35	55.83	51.22	57.93	62.37	**68.80**

Q & A