SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Important Sampling Based Active Learning for Imbalance Classification

Xinyue Wang[1,2], Bo Liu[3], Siyu Cao[1], Liping Jing.[1,2*] & Jian Yu.[1,2]

[1]*Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing,* 100044, *China;*
[2]*Beijing Key Lab of Traffic Data Analysis and Mining, No.3 Shangyuancun, Haidian District, Beijing,* 100044, *China;*
[3]*College of Information Science and Technology, Hebei Agricultural University, Baoding, Hebei, China*

Appendix A ALIS framework

To give a more comprehensive understanding of proposed framework ALIS, the graphical representation of ALIS is shown in Figure A1. Circles and triangles denote positive and negative set respectively, and negative class instances outnumbers the positive ones as FigureA1(a). At the beginning of the training process, the initial two class instances are tinted by solid as FigureA1(b). After important undersampling, we get selected negative class set $\mathcal{N}^j_{active}$ represented by black solid and red outline triangles in Figure A1(c), and after important oversampling, we get generated synthetic positive class set $\mathcal{P}^j_{active}$ represented by black solid and red outline circles in Figure A1(d). It can be seen that after important undersampling and oversampling, the decision boundary is adjusted from the solid black line to the dot red line in Figure A1(e). Thus, as the selection and generation continue, a precise decision boundary is determined with the aid of informative and representative boundary instances.

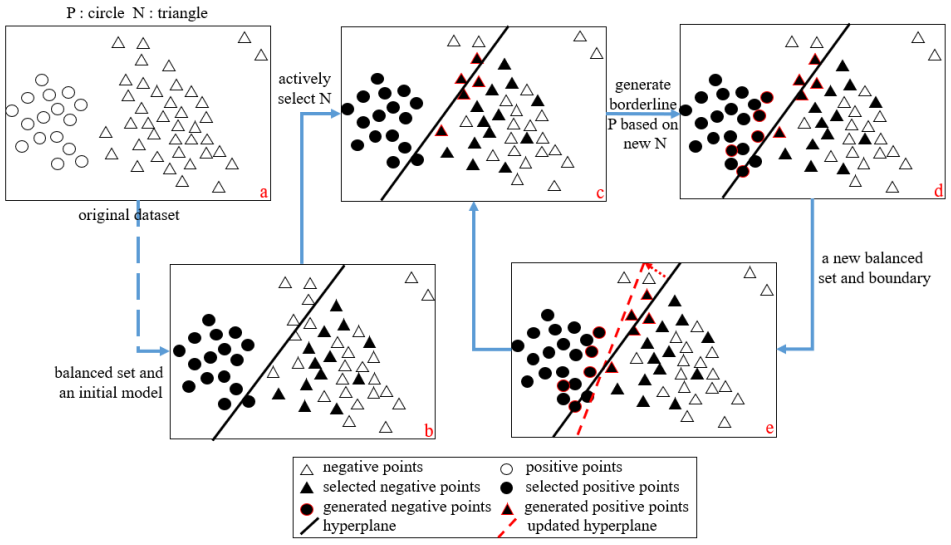

Figure A1 The schematic diagram of ALIS framework

Appendix B Ablation Test

ALIS consists of three components, i.e., important undersampling, decision boundary detection and important oversampling. Thus, in this section, to study the effect of each component individually, we replace the proposed method for each part with the existing one and check the results. More specifically, the important undersampling is replaced with borderline sampling

* Corresponding author (email: lpjing@bjtu.edu.cn)

(BorS) (Ertekin et al., 2007;), denoted as ALIS_{under}. The important oversampling is replaced with SMOTE (N. Chawla and Kegelmeyer, 2002), called as ALIS_{over}. The decision boundary detection method is replaced by the traditional SVMs, denoted as ALIS_{SVM}.

Table B1 lists the comparison results on twelve datasets in terms of three evaluation metrics. The results (mean and standard deviation) are recorded under the 5-fold cross validation. As expected, ALIS outperforms over any other situation which replaces one component with the existing method. This result further confirms that the proposed framework benefits from each component.

Appendix C Performance on Multi-class Datasets

As we known, a multi-class classification problem can be decomposed into several binary classification problems, such as multi-class SVM. Thus, the proposed method can be extended for multi-class classification task.

Actually, the experimental datasets adopted in this paper originally contain multiple classes. Take *balance* and *ecoli* dataset as examples. The dataset *balance* contains 625 examples divided into three classes (left (L), right (R) and balance (B)) which contain 288 / 288 / 49 instances respectively. Following AdaS (Peng, 2015), the binary data in our experiments is formed by merging the first two classes as majority class, and the third class is taken as the minority class. The dataset *ecoli* is for predicting localization site of protein, and 336 examples are actually divided into eight classes which contain 147 / 77 / 52 / 35 / 20 / 5 / 2 / 2 instances respectively. Following Yan et al.,2017, the binary data in our experiments is formed by taking the fourth class as the minority class and merging the other classes as the majority class.

For multi-class classification on *ecoli* dataset, we built three classifiers for each class. For multi-class classification on *ecoli* dataset, we built five classifiers for the top 5 largest classes. The experimental setiing is the same as described in Section 4.2. The overall performance in terms of each metric is listed in Table C1, which shows that ALIS scales to multi-class problems via decomposition into binary-class classification problems.

Appendix D The Effect of the Size of Neighbors in Oversampling in ALIS

A larger neighbors set $NN(\mathbf{x_i})$ for the instance $\mathbf{x_i}$ may finally result in more minority class instances being selected as sampling seeds. However, not all minority class instances is important for generating the synthetic instances because some of them can not provide much discriminative information in learning process. A smaller neighbors set $NN(\mathbf{x_i})$ for the instance $\mathbf{x_i}$ may finally result in less minority class instances being selected as sampling seeds. In this case, the designed generator may be incapable of capturing the appropriate distribution of the boundary minority class instances and further deteriorate the subsequent learning. Different data sets may have different characteristics (such as various amount, dimensionality, imbalance ratio and etc.). To demonstrate this point, a series of experiments are conducted on *Letter-a* dataset under varying $k = |NN(\mathbf{x_i})|$ from 4 to 15 with step 1. Figure D1 shows the experimental results in terms of F_{macro}. It can be seen that proper k is necessary to be selected. Fortunately, there is not a big difference among results with different k values. Thus, in experiments, k is tuned in $\{5, 6, 7, 8, 9\}$ for each dataset with the aid of validation set, and the proper k is used to train the learning model.

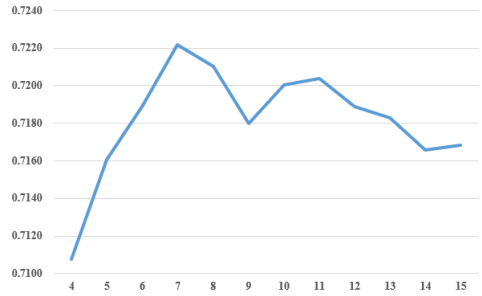

Figure D1 The effect of $k \in \{4, 5, \ldots, 15\}$ on the performance of ALIS in terms of F_{macro}

Table B1 Investigating ALIS by Replacing Important Oversampling, Important Undersampling, Decision Boundary Detection

Dataset	Metric	ALIS_{over}	ALIS_{under}	ALIS_{SVM}	ALIS_{LDM}
harbeman	Recall-minority	0.6926 ± 0.09	0.5801 ± 0.11	0.7294 ± 0.06	**0.7676 ± 0.07**
	Precision-majority	0.8783 ± 0.03	0.8381 ± 0.03	0.8190 ± 0.05	**0.8856 ± 0.04**
	F_{macro}	**0.7195 ± 0.02**	0.6663 ± 0.06	0.5108 ± 0.04	0.7114 ± 0.05
	AUC	0.7102 ± 0.10	0.7155 ± 0.10	0.5665 ± 0.05	**0.7192 ± 0.09**
libra	Recall-minority	0.6410 ± 0.14	0.6238 ± 0.14	0.5724 ± 0.16	**0.8067 ± 0.14**
	Precision-majority	0.9099 ± 0.03	0.9036 ± 0.03	0.8555 ± 0.06	**0.9457 ± 0.03**
	F_{macro}	0.7474 ± 0.03	0.7379 ± 0.08	0.5634 ± 0.11	**0.7520 ± 0.11**
	AUC	0.7898 ± 0.06	0.7265 ± 0.07	0.6426 ± 0.12	**0.8313 ± 0.07**
glass6	Recall-minority	0.9333 ± 0.09	0.9333 ± 0.09	0.9333 ± 0.09	**0.9668 ± 0.06**
	Precision-majority	0.9884 ± 0.02	0.9889 ± 0.02	**0.9895 ± 0.01**	0.9947 ± 0.01
	F_{macro}	0.6762 ± 0.20	0.8256 ± 0.14	0.8763 ± 0.10	**0.9368 ± 0.06**
	AUC	0.8495 ± 0.23	0.9369 ± 0.05	**0.9757 ± 0.05**	0.9676 ± 0.03
ecoli3	Recall-minority	0.9429 ± 0.08	0.9429 ± 0.08	**0.9714 ± 0.06**	**0.9714 ± 0.06**
	Precision-majority	0.9924 ± 0.01	0.9926 ± 0.01	0.9959 ± 0.01	**0.9964 ± 0.01**
	F_{macro}	0.5430 ± 0.37	0.4230 ± 0.33	0.6904 ± 0.04	**0.7554 ± 0.05**
	AUC	0.9184 ± 0.05	0.9241 ± 0.04	0.9152 ± 0.04	**0.9289 ± 0.02**
yeast0256vs3789	Recall-minority	0.7153 ± 0.18	0.6663 ± 0.22	0.6763 ± 0.15	**0.9103 ± 0.07**
	Precision-majority	0.9641 ± 0.01	0.9637 ± 0.02	0.9638 ± 0.02	**0.9764 ± 0.01**
	F_{macro}	0.6694 ± 0.23	0.6814 ± 0.18	**0.7722 ± 0.06**	0.6400 ± 0.06
	AUC	0.8161 ± 0.09	0.7950 ± 0.08	**0.8336 ± 0.05**	0.8320 ± 0.03
satimage	Recall-minority	0.0000 ± 0.00	0.0000 ± 0.00	0.2555 ± 0.03	**0.9170 ± 0.04**
	Precision-majority	0.9027 ± 0.00	0.9027 ± 0.00	0.9252 ± 0.00	**0.9831 ± 0.01**
	F_{macro}	0.4744 ± 0.00	0.4744 ± 0.00	**0.6709 ± 0.01**	0.4825 ± 0.02
	AUC	**1.0000 ± 0.00**	**1.0000 ± 0.00**	0.9092 ± 0.01	0.7283 ± 0.02
balance	Recall-minority	0.4800 ± 0.43	0.3044 ± 0.23	**0.7600 ± 0.15**	0.6600 ± 0.43
	Precision-majority	0.9494 ± 0.04	0.9352 ± 0.01	0.9663 ± 0.02	**0.9700 ± 0.03**
	F_{macro}	0.4797 ± 0.03	**0.5201 ± 0.03**	0.4950 ± 0.06	0.5193 ± 0.06
	AUC	0.4846 ± 0.05	0.5143 ± 0.06	**0.7056 ± 0.09**	0.5184 ± 0.06
shuttlec0vsc4	Recall-minority	**1.0000 ± 0.00**	**1.0000 ± 0.00**	0.4637 ± 0.16	**1.0000 ± 0.00**
	Precision-majority	**1.0000 ± 0.00**	**1.0000 ± 0.00**	0.9791 ± 0.01	**1.0000 ± 0.00**
	F_{macro}	**0.8545 ± 0.03**	0.8175 ± 0.06	0.7344 ± 0.07	0.8506 ± 0.03
	AUC	0.9829 ± 0.01	0.9827 ± 0.01	0.9848 ± 0.00	**0.9858 ± 0.00**
Letter-a	Recall-minority	0.9050 ± 0.02	0.9151 ± 0.03	**0.9937 ± 0.01**	0.9454 ± 0.01
	Precision-majority	0.9959 ± 0.00	0.9964 ± 0.00	**0.9997 ± 0.00**	0.9979 ± 0.00
	F_{macro}	0.7924 ± 0.02	0.7872 ± 0.01	**0.9864 ± 0.00**	0.7222 ± 0.02
	AUC	0.9801 ± 0.00	0.9822 ± 0.00	**0.9999 ± 0.00**	0.9845 ± 0.00
yeast4	Recall-minority	0.5855 ± 0.18	0.6636 ± 0.30	0.8218 ± 0.09	**0.9400 ± 0.09**
	Precision-majority	0.9851 ± 0.01	0.9879 ± 0.01	0.9924 ± 0.00	**0.9963 ± 0.00**
	F_{macro}	**0.7056 ± 0.07**	0.5592 ± 0.26	0.5620 ± 0.03	0.4980 ± 0.01
	AUC	0.8635 ± 0.07	**0.8793 ± 0.05**	0.8642 ± 0.05	0.7472 ± 0.12
yeast6	Recall-minority	0.8571 ± 0.14	0.8571 ± 0.10	0.9143 ± 0.13	**0.9435 ± 0.08**
	Precision-majority	0.9964 ± 0.00	0.9963 ± 0.00	**0.9974 ± 0.00**	0.9973 ± 0.00
	F_{macro}	0.4306 ± 0.37	0.5599 ± 0.27	0.5321 ± 0.02	**0.6413 ± 0.07**
	AUC	0.9262 ± 0.06	**0.9371 ± 0.07**	0.9179 ± 0.08	0.7912 ± 0.05
abalone19	Recall-minority	0.6190 ± 0.33	0.5619 ± 0.25	**0.7286 ± 0.25**	0.6286 ± 0.21
	Precision-majority	**0.9970 ± 0.00**	0.9965 ± 0.00	0.9964 ± 0.00	0.9965 ± 0.00
	F_{macro}	0.4379 ± 0.24	**0.5403 ± 0.02**	0.3731 ± 0.02	0.5061 ± 0.04
	AUC	0.7172 ± 0.11	**0.8123 ± 0.08**	0.7121 ± 0.09	0.7177 ± 0.11

Table C1 Comparing ALIS with baselines on *balance* and *ecoli* dataset

Dataset	Metric	RoSR-LL	KernelADASYN	AdaS	AQM	BorS	ALIS_{LDM}
balance	Recall	0.7018 ± 0.07	0.6340 ± 0.01	0.7607 ± 0.05	0.8172 ± 0.07	0.6215 ± 0.02	**0.8554 ± 0.16**
	Precision	0.9608 ± 0.02	0.9413 ± 0.02	**0.9765 ± 0.00**	0.9520 ± 0.02	0.9360 ± 0.02	0.9639 ± 0.02
	F_{macro}	0.8766 ± 0.02	0.9541 ± 0.02	0.8748 ± 0.02	0.9358 ± 0.03	0.9419 ± 0.01	**0.9597 ± 0.01**
	AUC	0.6758 ± 0.03	-	-	0.1024 ± 0.04	0.0857 ± 0.01	**0.8308 ± 0.03**
ecoli	Recall	0.9611 ± 0.07	0.3201 ± 0.05	0.8371 ± 0.08	0.9345 ± 0.08	0.7316 ± 0.14	**0.9860 ± 0.02**
	Precision	0.9925 ± 0.01	0.9042 ± 0.02	0.9879 ± 0.01	0.9832 ± 0.02	0.9528 ± 0.02	**0.9947 ± 0.01**
	F_{macro}	0.8703 ± 0.04	0.8784 ± 0.05	0.8284 ± 0.04	**0.8943 ± 0.03**	0.8932 ± 0.05	0.8668 ± 0.11
	AUC	0.6816 ± 0.11	-	-	0.1047 ± 0.06	0.0437 ± 0.04	**0.9156 ± 0.04**