Data Science in Finance and Economics

DSFE, 1(1): 21–36.
DOI: 10.3934/DSFE.2021002
Received: 01 May 2021
Accepted: 25 May 2021
Published: 01 June 2021

http://www.aimspress.com/journal/dsfe

Research article

Learning from class-imbalanced data: review of data driven methods and algorithm driven methods

Cui Yin Huang and Hong Liang Dai*

School of Economics and Statistics, Guangzhou University, Guangzhou 510006, China

* **Correspondence:** Email: hldai618@gzhu.edu.cn.

Supplementary

Appendix

Here, we created a series of data sets, which were used to verify the following experiments:

- E1: in the case of different classes were non-overlapping, the influence of IR on standard classifiers;
- E2: in the case of equal overlapping ranges, the influence of IR on standard classifiers;
- E3: in the case of equal IR, the influence of the overlapping ranges on standard classifiers;
- E4: in the case of equal noise samples，the influence of IR on standard classifiers.

The detailed experiments were as follows. We implemented above experiments with Python 3.8.5.

Firstly, experiment 1 was designed; the Fibonacci sequence was used to set IR. Here, 8 data sets with 2 variables were created, and these variables were fitted to uniform distribution; these data sets, the size of the majority class samples all was set as 699, and the size of the minority class samples was set as 24, 39, 63, 102, 165, 267, 432, 699; all variables of the majority class obey uniform distribution of (0, 4.50); All variable of the minority class obey uniform distribution of (5.50, 10. 00). Therefore, samples of different classes were non-overlapping in sample space. These data sets were shown in Figure 1; The majority class were represented as orange “.”; The minority class were represented blue “+”. Then, the ratio of the training data set to test data set was set as 7: 3; decision tree classifier (DT), random forest classifier (RF) , naive Bayesian classifier (NB), support vector machine (SVM) classifier, adaboost-based decision tree classifier (AD) were built. These classifiers

all got consistent results that all samples were correctly classified. These results are shown in Table 1.1. Thus, when data distributions have these performances that are represented as Figure 1, IR was independent of these standard classifiers.

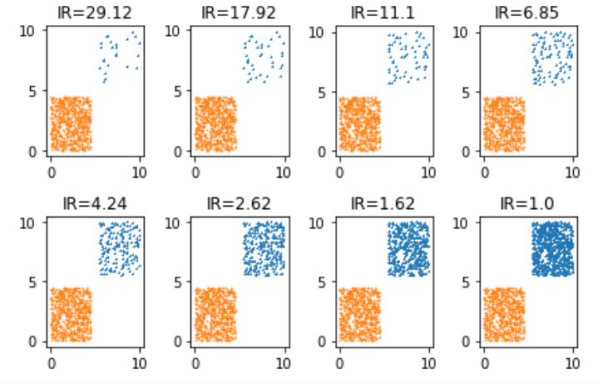

Figure 1. The scatter graphs of data sets with different IR.

Table 1.1. Results of the case of non-overlapping.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	1.0000	1.0000	1.0000
2	17.92	1.0000	1.0000	1.0000
3	11.10	1.0000	1.0000	1.0000
4	6.85	1.0000	1.0000	1.0000
5	4.24	1.0000	1.0000	1.0000
6	2.62	1.0000	1.0000	1.0000
7	1.62	1.0000	1.0000	1.0000
8	1.00	1.0000	1.0000	1.0000

Secondly, experiment 2 was designed. 8 data sets with 2 variables were created; these data sets ,the size of the majority class samples all was set as 699, and the size of the minority class samples were set as 24, 39, 63, 102, 165, 267, 432, 699; all variables of the majority class obey uniform distribution of (0, 4.00); all variable of the minority class obey uniform distribution of (6.00, 10. 00); the overlapping area was (4.00, 6.00) × (4.00, 6.00); Then, The ratio of the training data set to test data set was set as 7: 3; decision tree classifier (DT), random forest classifier (RF) , naive Bayesian classifier (NB), support vector machine (SVM) classifier, adaboost-based decision tree classifier (AD) were built; results are shown in Table 2.1. to Table2.5.; these results show that the F-measure value almost was ascending with the lower of the IR value.

Table 2.1. Results of DT under the different IR with fixed overlapping.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	0.9862	0.9214	0.6429
2	17.92	0.9775	0.9567	0.7259
3	11.10	0.9782	0.9023	0.7751
4	6.85	0.9627	0.9224	0.6806
5	4.24	0.9654	0.9284	0.7902
6	2.62	0.9448	0.9315	0.7982
7	1.62	0.9441	0.9391	0.8659
8	1.00	0.9524	0.9524	0.9057

Table 2.2. Results of RA under the different IR with fixed overlapping.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	0.9954	0.9258	0.8571
2	17.92	0.9910	0.9309	0.8667
3	11.10	0.9782	0.8790	0.7727
4	6.85	0.9793	0.9311	0.7989
5	4.24	0.9808	0.9370	0.8780
6	2.62	0.9552	0.9334	0.8291
7	1.62	0.9588	0.9497	0.8993
8	1.00	0.9476	0.9474	0.8956

Table 2.3. Results of NB under the different IR with fixed overlapping.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	0.9954	0.9258	0.8571
2	17.92	0.9775	0.8165	0.6667
3	11.10	0.9651	0.7977	0.6364
4	6.85	0.9710	0.8416	0.7083
5	4.24	0.9731	0.9218	0.8299
6	2.62	0.9724	0.9635	0.8948
7	1.62	0.9618	0.9561	0.9070
8	1.00	0.9405	0.9403	0.8821

Table 2.4. Results of SVM under the different IR with fixed overlapping.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	0.9954	0.9258	0.8571
2	17.92	0.9820	0.8563	0.7333
3	11.10	0.9651	0.7977	0.6364
4	6.85	0.9710	0.8416	0.7083
5	4.24	0.9731	0.9106	0.8293
6	2.62	0.9655	0.9449	0.8666
7	1.62	0.9471	0.9374	0.8710
8	1.00	0.9476	0.9476	0.8965

Table 2.5. Results of AD under the different IR with fixed overlapping.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	0.9862	0.9214	0.6429
2	17.92	0.9775	0.9567	0.7259
3	11.10	0.9782	0.9023	0.7751
4	6.85	0.9627	0.9224	0.6806
5	4.24	0.9654	0.9284	0.7902
6	2.62	0.9448	0.9315	0.7982
7	1.62	0.9500	0.9464	0.8803
8	1.00	0.9452	0.9451	0.8913

Thirdly, experiment 3 was designed. 6 data sets with 2 variables were created; these data sets ,the size of the majority class samples all were set as 699, and the size of the minority class samples were set as 63; these variables were fitted to uniform distribution; these overlapping ranges of these data sets were set 0.25 × 0.25, 0.50 × 0.50，0.75 × 0.75, 1.25 × 1.25 , 2 × 2, 3.25 × 3.25. These data sets were shown in Figure 2; The majority class were represented as orange “.”; The minority class were represented blue “+”. Then, The ratio of the training data set to test data set was set as 7: 3; decision tree classifier (DT), random forest classifier (RF), naive Bayesian classifier (NB), support vector machine (SVM) classifier, adaboost-based decision tree classifier (AD) were built; results are shown in Table 3.1. to Table 3.5.; these results show that the F-measure value almost was ascending with the lower of the overlapping range.

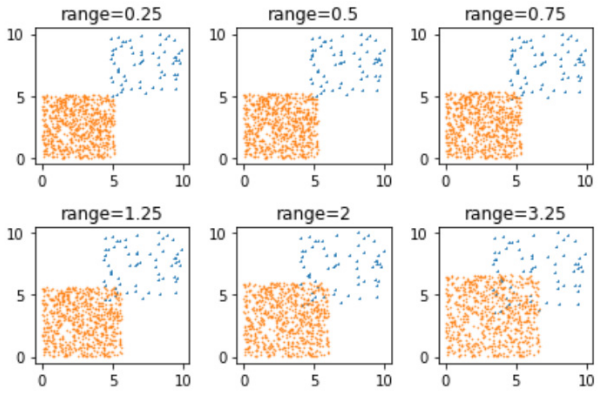

Figure 2. The scatter graphs of data sets with different overlapping ranges.

Table 3.1. Results of DT under the different overlapping ranges with fixed IR condition.

Items	Ranges	Accuracy	G-Mean	F-measure
1	0.25	1.0000	1.0000	1.0000
2	0.50	0.9956	0.9976	0.9474
3	0.75	0.9956	0.9976	0.9444
4	1.25	0.9913	0.9535	0.9091
5	2.00	0.9869	0.9512	0.8658
6	3.25	0.9825	0.9349	0.7785

Table 3.2. Results of RF under the different overlapping ranges with fixed IR condition.

Items	Ranges	Accuracy	G-Mean	F-measure
1	0.25	1.0000	1.0000	1.0000
2	0.50	1.0000	1.0000	1.0000
3	0.75	1.0000	1.0000	1.0000
4	1.25	0.9913	0.9535	0.9091
5	2.00	0.9913	0.9535	0.9091
6	3.25	0.9869	0.9075	0.8235

Table 3.3. Results of NB under the different overlapping ranges with fixed IR condition.

Items	Ranges	Accuracy	G-Mean	F-measure
1	0.25	1.0000	1.0000	1.0000
2	0.50	1.0000	1.0000	1.0000
3	0.75	1.0000	1.0000	1.0000
4	1.25	0.9869	0.9293	0.8636
5	2.00	0.9782	0.8790	0.7727
6	3.25	0.9782	0.8402	0.7059

Table 3.4. Results of SVM under the different overlapping ranges with fixed IR condition.

Items	Ranges	Accuracy	G-Mean	F-measure
1	0.25	1.0000	1.0000	1.0000
2	0.50	1.0000	1.0000	1.0000
3	0.75	1.0000	1.0000	1.0000
4	1.25	0.9913	0.9535	0.9091
5	2.00	0.9651	0.7977	0.6364
6	3.25	0.9782	0.8402	0.7059

Table 3.5. Results of AD under the different overlapping ranges with fixed IR condition.

Items	Ranges	Accuracy	G-Mean	F-measure
1	0.25	1.0000	1.0000	1.0000
2	0.50	0.9956	0.9976	0.9474
3	0.75	0.9956	0.9976	0.9444
4	1.25	0.9913	0.9535	0.9091
5	2,00	0.9869	0.9512	0.8658
6	3.25	0.9825	0.9349	0.7785

Fourthly, experiment 4 was designed. 8 data sets with 2 variables were created; these data sets ,the size of the majority class samples all were set as 699, and the size of the minority class samples were set as 24, 39, 63, 102, 165, 267, 432, 699; Almost all of these samples of the majority class obey the distribution that 2 variables obey the uniform distribution on (0,6); 2 variable of the minority class obey uniform distribution of (4 10); the number of noises of the majority class of all data sets were set to 0; the number of noises of the minority class of all data sets were set to 5; These data sets were shown in Figure 3; The majority class were represented as orange “.”; The minority class were represented blue “+”; Then, the ratio of the training data set to test data set was set as 7: 3; decision tree classifier (DT), random forest classifier (RF), naive Bayesian classifier (NB), support vector machine (SVM) classifier, adaboost-based decision tree classifier (AD) were built; results are shown in Table 4.1. to Table 4.5.; these results show that the F-measure value almost was ascending with the lower of the IR value.

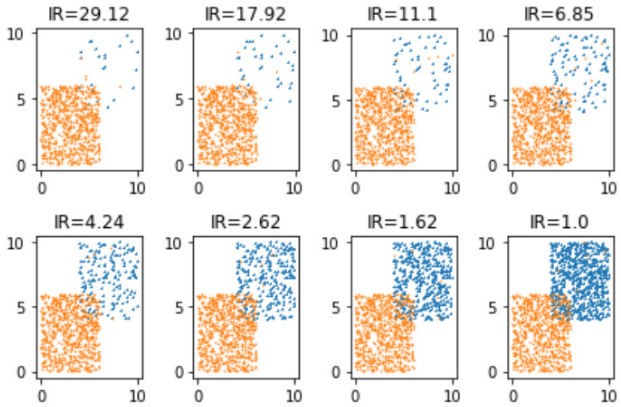

Figure 3. the scatter graphs of data sets with noise labels with different IR.

Table 4.1. Results of DT under the different IR with the same number of noises condition.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	0.9677	0.7400	0.3472
2	17.92	0.9775	0.8901	0.6857
3	11.10	0.9607	0.7849	0.5748
4	6.85	0.9502	0.8851	0.6504
5	4.24	0.9731	0.9488	0.8522
6	2.62	0.9655	0.9480	0.8624
7	1.62	0.9294	0.9211	0.8324
8	1.00	0.9381	0.9382	0.8766

Table 4.2. Results of RF under the different IR with the same number of noises condition.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	0.9770	0.7436	0.4630
2	17.92	0.9865	0.8944	0.8000
3	11.10	0.9607	0.8126	0.5833
4	6.85	0.9668	0.8937	0.7467
5	4.24	0.9731	0.9488	0.8522
6	2.62	0.9621	0.9355	0.8470
7	1.62	0.9353	0.9246	0.8449
8	1.00	0.9357	0.9356	0.8711

Table 4.3. Results of NB under the different IR with the same number of noises condition.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	0.9862	0.8165	0.6667
2	17.92	0.9820	0.8563	0.7333
3	11.10	0.9694	0.8431	0.6696
4	6.85	0.9627	0.8593	0.7110
5	4.24	0.9692	0.9281	0.8282
6	2.62	0.9552	0.9205	0.8189
7	1.62	0.9412	0.9359	0.8607
8	1.00	0.9333	0.9324	0.8646

Table 4.4. Results of AD under the different IR with the same number of noises condition.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	0.9770	0.7436	0.4630
2	17.92	0.9865	0.8944	0.8000
3	11.10	0.9782	0.8976	0.7646
4	6.85	0.9668	0.8778	0.7432
5	4.24	0.9769	0.9419	0.8701
6	2.62	0.9586	0.9227	0.8319
7	1.62	0.9412	0.9346	0.8601
8	1.00	0.9333	0.9331	0.8662

Table 4.5. Results of SVM under the different IR with the same number of noises condition.

Items	IR	Accuracy	G-Mean	F-measure
1	29.12	0.9677	0.7400	0.3472
2	17.92	0.9775	0.8901	0.6857
3	11.10	0.9651	0.8145	0.6222
4	6.85	0.9544	0.8873	0.672
5	4.24	0.9731	0.9488	0.8522
6	2.62	0.9586	0.9436	0.8381
7	1.62	0.9324	0.9248	0.8395
8	1.00	0.9381	0.9380	0.8760

Core and result for these experiments can be download from https://blog.csdn.net/weixin_42181646/article/details/116744366.

AIMS Press