

1 Score (without probability)

CV method: RepeatedStratifiedKFold(n_splits=5, n_repeats=6). So, training_data are 0.8 of data and testing_data are 0.2 of data.

While computing the training time, training_data = data.

1.1 Standard models

Data info: data.shape = (2000, 100)

	SVC	RFC	KNC	LGB
Best HP	kernel = 'rbf' gamma = 0.0151 C = 1.45	n_estimators = 400 min_samples_split = 3 bootstrap = False	n_neighbors = 1 algorithm = 'ball_tree', leaf_size = 10 p = 8	best_learning_rate = 0.112 best_min_data_in_leaf = 7 best_num_leaves = 30
Score	0.98054 ± 0.00627	0.91433 ± 0.01130	0.90033 ± 0.01501	0.91175 ± 0.00405
Training time	0.506 s (1 thread)	13.519 s (4 thread)	0.106 s (4 thread)	0.785 s (4 thread)

1.2 Data augmentation

Data info: data.shape = (400000, 100)

	SVC	SVC bagging	RFC	KNC	LGB
Best HP	kernel = 'rbf' gamma = 0.0145 C = 0.8	kernel = 'rbf' gamma = 0.0145 C = 0.8 n_estimators = 4 max_samples = 0.95	n_estimators = 400 min_samples_split = 3 bootstrap = False	n_neighbors = 1 algorithm = 'ball_tree' leaf_size = 10 p = 8	best_learning_rate = 0.112 best_min_data_in_leaf = 7 best_num_leaves = 30
Score normal dataset	0.98966 \pm 0.00214	0.98967 \pm 0.00204	0.99608 \pm 0.00118	0.98917 \pm 0.00224	0.95367 \pm 0.00394
Score augment. dataset	0.98966 \pm 0.00214	0.98980 \pm 0.00187	0.99587 \pm 0.00118	0.98917 \pm 0.00224	0.95430 \pm 0.00372
Training time	1 h 53 m 50.236 s (1 thread)	1 h 21 m 1.844 s (4 thread)	5 h 9 m 18.714 s (4 thread)	45.476 s (4 thread)	10.325 s (4 thread)

1.3 Sorted models

Data info: data.shape = (2000, 100)

	SVC	RFC	KNC	LGB
Best HP	kernel = 'linear' C = 0.15	n_estimators = 400 min_samples_split = 3 bootstrap = False	n_neighbors = 1 algorithm = 'ball_tree' leaf_size = 10 p = 8	best_learning_rate = 0.177 best_min_data_in_leaf = 24 best_num_leaves = 120
Score	0.99932 \pm 0.00124	0.98992 \pm 0.00518	0.99842 \pm 0.00177	0.99617 \pm 0.00184
Training time	0.017 s (1 thread)	0.899 s (4 thread)	0.097 s (4 thread)	0.627 s (4 thread)

2 Score (with probability) + computational time

Data info: `data.shape = (2000, 100)`

Every model has been tested using `data` and repeating the test 2000 times. So, the testing sample is about 4 million elements. The final time reported is the sum of every single testing time.

N.B.: the training time is different only for the SVC model, because of the hyper parameter `probability = True`; the other models always calculate the probability.

2.1 Standard model

	SVC	SVC bagging	RFC	KNC	LGB
Training time	2.639 s (1 thread)	2.692 s (4 thread)	13.519 s (4 thread)	0.106 s (4 thread)	0.802 s (4 thread)
Testing time	12 m 47.225 s (1 thread)	9 m 57.627 s (4 thread)	6 m 44.651 s (4 thread)	4 h 0 m 39.048 s (4 thread)	16.325 s (4 thread)

2.2 Sorted model

	SVC	RFC	KNC	LGB
Training time	0.766 s (1 thread)	0.899 s (4 thread)	0.097 s (4 thread)	0.643 s (4 thread)
Testing time	3 m 9.198 s (1 thread)	3 m 26.799 s (4 thread)	1 h 23 m 32.290 s (4 thread)	9.788 s (4 thread)

3 Real life test

Test models on OD from 91 to 105 (data already cleaned).

3.1 Standard model

	SVC	SVC bagging	RFC	KNC	LGB
Total time (only model)	26 m 27.209 s (1 thread)	20 h 29 m 44.908 s (4 thread)	6 d 19 h 29 m 5.175 s (4 thread)	6 d 18 h 20 m 54.573 s (4 thread)	14 m 32.227 s (4 thread)
Total time (all)	38 m 22.223 s (1 thread)	20 h 56 m 47.242 s (4 thread)	6 d 20 h 9 m 27.406 s (4 thread)	6 d 18 h 46 m 23.865 s (4 thread)	27 m 43.873 s (4 thread)

3.2 Sorted model

	SVC	RFC	KNC	LGB
Total time (only model)	8 m 17.891 s (1 thread)	6 d 19 h 19 m 5.913 s (4 thread)	6 d 18 h 43 m 17.681 s (4 thread)	13 m 42.721 s (4 thread)
Total time (all)	21 m 8.052 s (1 thread)	6 d 19 h 58 m 32.617 s (4 thread)	6 d 19 h 13 m 18.166 s (4 thread)	27 m 59.176 s (4 thread)