**Method**:

**Datasets and preprocess**:

We use the dataset created and made available by Tsagkogeorga *et al*. Original dataset collected from Harmonizome website(<https://maayanlab.cloud/Harmonizome>). 15 one-hot-encoded datasets were selected to construct our dataset. The dataset was initially standardized to continuous-value ranged from 0 to 1, or -1 to 1 where 1 indicated strong positive gene-feature association, 0 indicates no gene-feature association observed and -1 indicates strong negative gene-feature association.

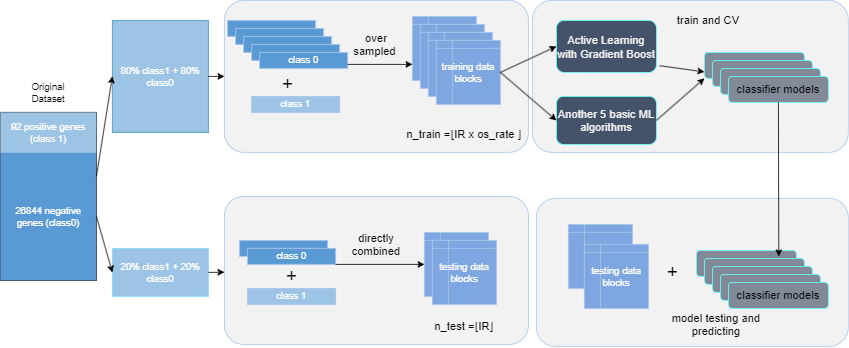
To gain high-informative features and reduce feature dimensions, following 2 kinds of features will be removed: (i) zero values more than 70%. (ii)variance less than 16% of whole.

**Model definition**:

To predict genes associates with RNA methylation pathways, we took 80% positive and negative samples for training and CV while other 20% samples remaining for testing. Next, calculated the imbalance ratio (IR), i.e., the ratio between negative and positive examples.

Realizing dataset’s IR is higher than most public datasets’, reaching more than 290, oversample algorithm was applied to alleviate imbalance problem: Assuming parameter *OSRATE*, a decimal number between 0 and 1 used to determine oversample rate. This parameter can be obtained through cross-validation. We next divided majority training samples (negative training samples) into *NTRAIN = IR\* OSRATE* (round down) equal pieces, combining each negative piece with the same positive sample to get *NTRAIN* copies of data blocks. Then, each data block was balanced by oversample methods. Values in generated samples was initiated to discrete or continuous value between -1 and 1.

For test set, calculate test set’s IR, it should be equal to the training set’s. Then, directly divide negative test set into IR(rounded down) pieces, each data piece combined with the same positive test sample. As shown in FIG 1.



【FIG1】

**Result:**

We applied active-learning (based on gradient-boost and uncertainty-sampling) and 5 machine-learning classifiers: SVM, GB, GNB, RF, LR. Trained on NTRAIN=173 balanced data blocks. Notice: Hyperparameters in GB based active-learning model were default assigned while grid-search technique was used to select best (accuracy) hyperparameters among given parameter combinations in 3 machine-learning classifiers (SVM, GB, RF). 3-fold cross-validation was used in grid-search in order to select best hyperparameters. None-sample method(OSRATE = 1) was set as baseline.

**Oversample methods successfully improve predictive accuracy:**

We use 10-fold cross validation to evaluate model performance. In general, oversampling method successfully improved the predictive accuracy of genes associates with RNA methylation pathways. Active-learning method hits average accuracy 89.74%, reaching accuracy 91% after oversampling. GB hits accuracy 87% in ML methods (Figure 3) and this value reached 93% of best after oversampling (Figure 2). It should be noted that the final training sets of AL and ML are not the same after oversampling, this is because the training scripts for AL and ML are different.

**Different performance between SMOTE and GAN:**

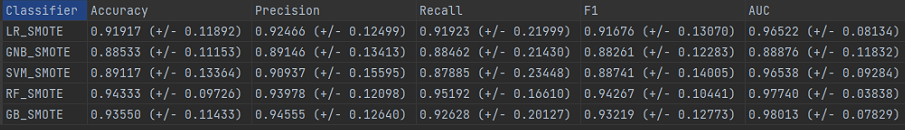
(…..)

**Influence of parameter OSRATE:**

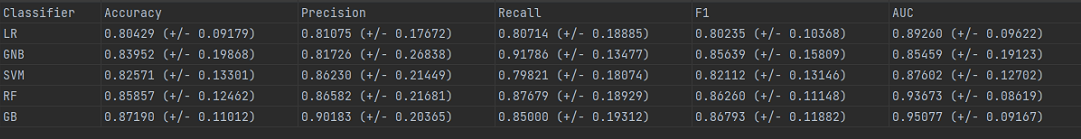
(…..)

**Influence of datablock:**

(…..)



【FIG2】



【FIG3】

TO DO

(i)I didn’t apply hybrid methods yet, considering combining oversample and undersample algorithm to our study.

(ii)Other ways like GAN and cost-sensitive learning may also works.

(iii)Try some fancy evaluate metrics to prove our result.

(iv)Adjust OSRATE to get more comparative results.

(v) More experiments should be added in order to enrich paper.