DMW Assignment-8

SVM-Boosting based on Markov resampling: Theory and algorithm

Submitted By - [Akhil Shukla, IIT2018112] [Akhil Singh, IIT2018198][Javed Ali, IIT2018501][Manan Bajaj, IIT2018502][Lokesh, IIT2018503]

6 th Semester, B.Tech, Department of Information Technology, IIIT Allahabad

You have to understand and implement the algorithm proposed in the paper "SVM-Boosting based on Markov resampling: Theory and algorithm☆ ".

The paper introduces Markov Boosting for linear SVM and comparison with AdaBoost and XGBoost Classifiers. The boosting methods mentioned in paper are for linear kernel SVM only so we implement AdaBoostClassifier, XGBoost CLassifier, Markov Boosted SVM (BM-SVM) and it's improved version (IBM-SVM) separately and test their performance on the same dataset with the same train to test sample ratio.

We used following datasets -
Letter Dataset[2], it has 16 different features relating to alphabets A and B (for forming a binary classifier as given in paper) to be recognized. First we segment the dataset into a train and test set with 1088 samples for training and 467 for testing. We use markov sampling (explained next) to choose samples from the training set that forms a markov chain.

COD-RNA Dataset[3], it has 7 different features relating to RNA type(-1 and 1 for non-coding and coding RNA) to be recognized. First we segment the dataset into a train and test set with 7000 samples for training and 3000 for testing. We use markov sampling (explained next) to choose samples from the training set that forms a markov chain.

Markov Boosting Algorithm
It is based on Markov Resampling which is in principle Batch Learning of SVM on N samples from training dataset repeatedly for finding Markov Chains in the Training Set. These Markov Chains Samples are then used to train a final SVM.[1]

Improved Markov Boosting Algorithm
The improvisation is using the support vectors instead of the entire weight matrix for classification. Support vectors are supposed to be only few in number regardless of number of training samples, which gives it a much better speed without sacrificing much accuracy.[1]

Observation

On COD-RNA Dataset ->
All Boosted Classifiers are run using 7000 training samples and 3000 test samples.

Accuracy on AdaBoosted SVM Classifier = 76.433 %

Misclassification Rate on AdaBoosted SVM Classifier = 23.567 %

Accuracy on XGBoosted Classifier = 92.466 %
Misclassification Rate on XGBoosted Classifier = 7.543 %

Accuracy on Linear Kernel SVM using BM-SVM = 89.966 %
Misclassification Rate on Linear Kernel SVM using BM-SVM = 10.034 %

Accuracy on Linear Kernel SVM using IBM-SVM = 90.800 %
Misclassification Rate on Linear Kernel SVM using IBM-SVM = 9.200 %

On Letter Dataset ->
All Boosted Classifiers are run using 1088 training samples and 467 test samples.

Accuracy on AdaBoosted SVM Classifier = 98.501 %
Misclassification Rate on AdaBoosted SVM Classifier = 01.499 %

Accuracy on XGBoosted Classifier = 99.357 %
Misclassification Rate on XGBoosted Classifier = 0.643 %

Accuracy on Linear Kernel SVM using BM-SVM = 99.143 %
Misclassification Rate on Linear Kernel SVM using BM-SVM = 0.857 %

Accuracy on Linear Kernel SVM using IBM-SVM = 99.571 %
Misclassification Rate on Linear Kernel SVM using IBM-SVM = 0.429 %

The IBM-SVM performs slightly better than simple Markov Boosting.

References

[1] Jiang H, Zou B, Xu C, Xu J, Tang YY. SVM-Boosting based on Markov resampling: Theory and algorithm. Neural Networks : the Official Journal of the International Neural Network Society. 2020 Nov;131:276-290. DOI: 10.1016/j.neunet.2020.07.036.

[2] Letter Dataset - https://archive.ics.uci.edu/ml/datasets/Letter+Recognition

[3] COD-RNA Dataset - https://www.openml.org/d/351

[4] XGBoost Classifier Documentation
https://xgboost.readthedocs.io/en/latest/parameter.html

[5] sklearn Ensemble AdaBoostClassifier
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html