INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

VI Semester BTech in Information Technology

Report - Group Assignment 5

Data Mining and Warehousing

Group Members:

- **Mudit Goyal (IIT2018132)**
- **Shiv Kumar (IIT2018134)**
- **Moksh Grover (IIT2018186)**
- **Karan Chatwani (IIT2018194)**
- **Shubham (IIT2018200)**

Title: SVM-Boosting based on Markov resampling: Theory and algorithm
Author: Hongwei Jiang, Bin Zou , Chen Xu , Jie Xu, Yuan Yan Tang

1. INTRODUCTION:

With the advent of the high-tech era, the capacity of data is growing rapidly, and the value density of data is usually very low, which implies that there are many noise examples in big data. The main idea of the AdaBoost algorithm is to adjust the weights of training examples so that the examples misclassified by the last classifier will be focused in the next train.

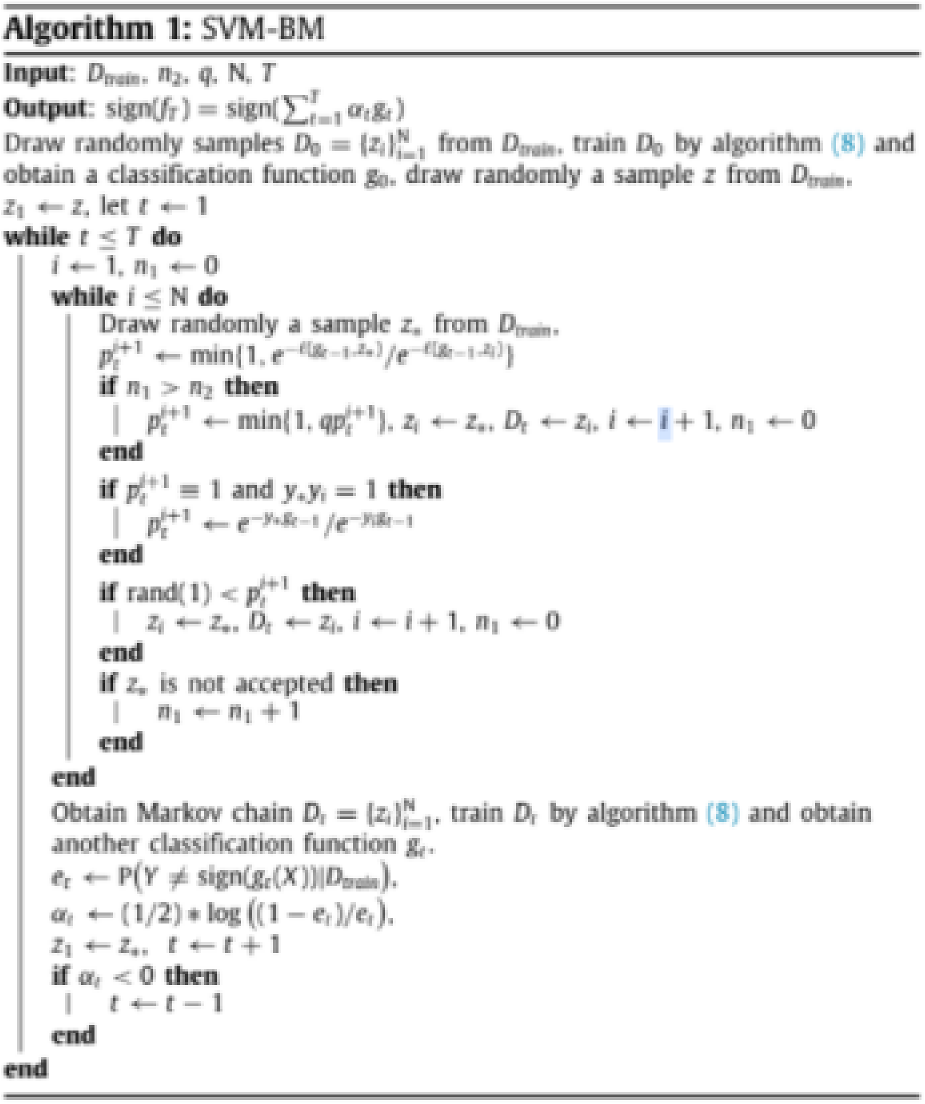
Thus AdaBoost algorithm will be very time-consuming or hard to implement as the size of data is very big. In addition, many experiments of machine learning indicate that the noise example not only leads to an increase in the amount of storage space, algorithms. We highlight some contributions of this paper.

• The Boosting algorithm with general convex loss function based on u.e.M.c. examples are proved to be consistent and its fast convergence rate is established. • Two new SVM-Boosting algorithms based on Markov resampling, SVM-BM and ISVM-BM are proposed. The numerical experiments based on benchmark data show that the proposed algorithms have better classification performance compared to the classical AdaBoost, XGBoost and SVM-AdaBoost algorithms.

2. PROPOSED PROBLEM :

In this problem , We have to apply Boosting algorithm based on Markov resampling to Support Vector Machine (SVM), and introduce two new resampling based Boosting algorithms: SVM-Boosting based on Markov resampling (SVM-BM) and improved SVM-Boosting based on Markov resampling (ISVM-BM).

3. ALGORITHM:

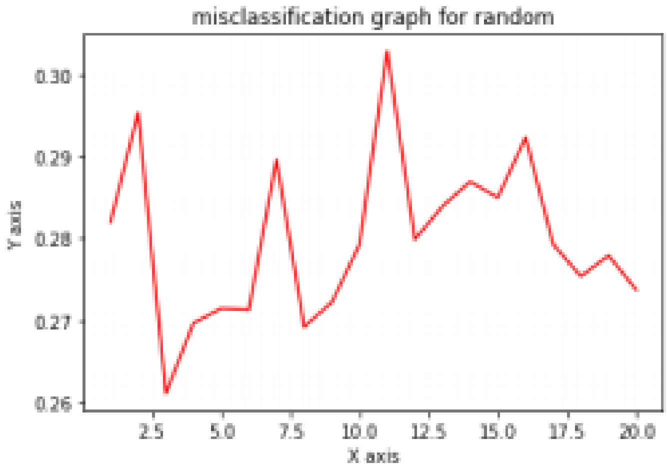


Algorithm 1: SVM-BM

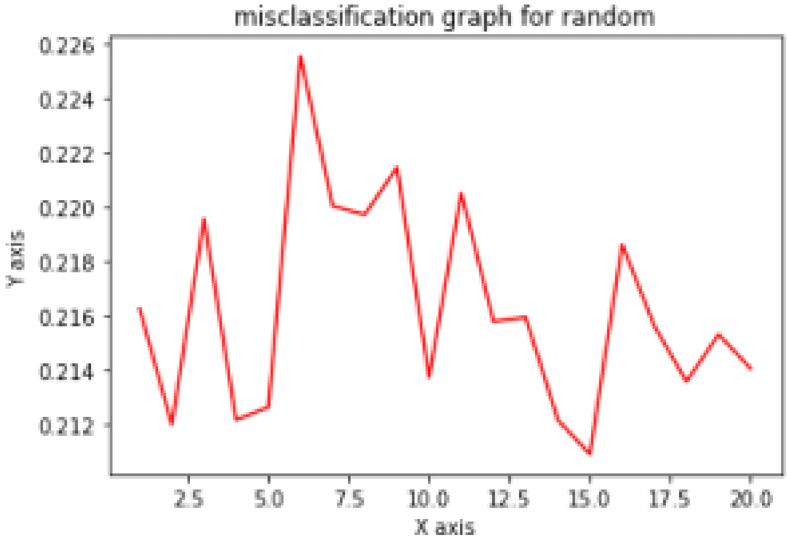
```
Input: $D_{train}$, $n_2$, $q$, N, $T$
Output: $\mathrm{sign}(f_T) = \mathrm{sign}(\sum_{t=1}^{T} \alpha_t g_t)$
Draw randomly samples $D_0 = \{z_i\}_{i=1}^{N}$ from $D_{train}$, train $D_0$ by algorithm (8) and
obtain a classification function $g_0$, draw randomly a sample $z$ from $D_{train}$.
$z_1 \leftarrow z$, let $t \leftarrow 1$
while $t \leq T$ do
    $i \leftarrow 1$, $n_1 \leftarrow 0$
    while $i \leq$ N do
        Draw randomly a sample $z_*$ from $D_{train}$.
        $p_t^{i+1} \leftarrow \min\{1, e^{-\ell(g_{t-1}, z_*)}/e^{-\ell(g_{t-1}, z_i)}\}$
        if $n_1 > n_2$ then
            $p_t^{i+1} \leftarrow \min\{1, qp_t^{i+1}\}$, $z_i \leftarrow z_*$, $D_t \leftarrow z_i$, $i \leftarrow i + 1$, $n_1 \leftarrow 0$
        end
        if $p_t^{i+1} = 1$ and $y_* y_i = 1$ then
            $p_t^{i+1} \leftarrow e^{-y_* g_{t-1}}/e^{-y_i g_{t-1}}$
        end
        if rand(1) $< p_t^{i+1}$ then
            $z_i \leftarrow z_*$, $D_t \leftarrow z_i$, $i \leftarrow i + 1$, $n_1 \leftarrow 0$
        end
        if $z_*$ is not accepted then
            $n_1 \leftarrow n_1 + 1$
        end
    end
    Obtain Markov chain $D_t = \{z_i\}_{i=1}^{N}$, train $D_t$ by algorithm (8) and obtain
    another classification function $g_t$.
    $e_t \leftarrow \mathrm{P}(Y \neq \mathrm{sign}(g_t(X)) | D_{train})$.
    $\alpha_t \leftarrow (1/2) * \log((1 - e_t)/e_t)$.
    $z_1 \leftarrow z_*$,  $t \leftarrow t + 1$
    if $\alpha_t < 0$ then
        $t \leftarrow t - 1$
    end
end
```

4. RESULT :

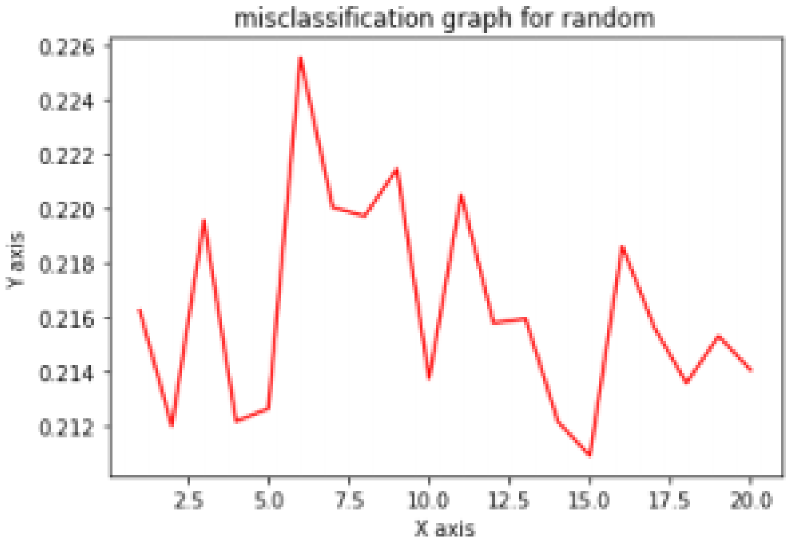
- `mean misclassification rate for random sample with std for letter (27.993100344982754, 0.9862973533986148)`



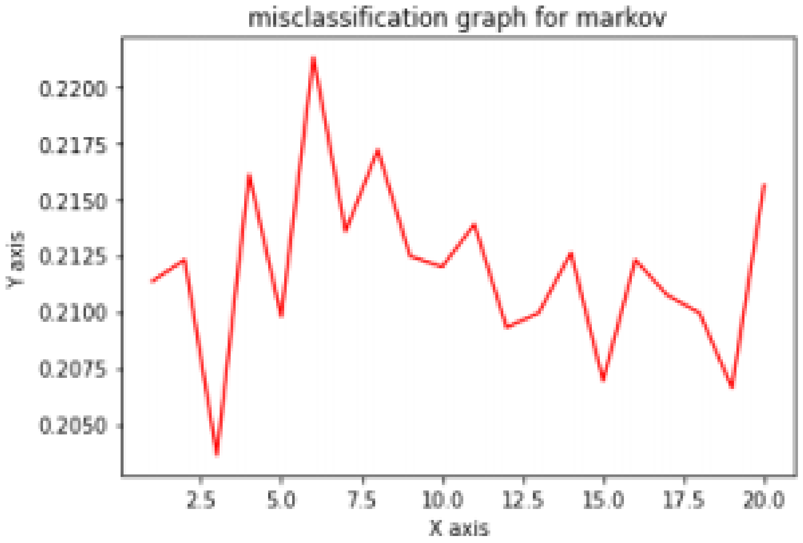
- mean misclassification rate for boosted with markov sample with std for letter (28.01184940752962, 0.7155204324623777)



- mean misclassification rate for random sample with std for magic (21.62697160883281, 0.37980453556690263)



- `mean misclassification rate for markov sample with std for magic (21.187697160883282, 0.3852941367498738)`



5. CONCLUSION:

In this paper, we introduced the idea of resampling for Boosting algorithms. We firstly proved that the resampling-based Boosting algorithm with general convex loss function is consistent and established the fast learning rate for resampling-based Boosting algorithm. To our knowledge, these results are the first results on this topic. We also applied the Boosting algorithm based on resampling to the classical classification algorithm, SVM, and proposed the SVM Boosting based on Markov resampling algorithm.(SVM-BM).

Along the line of the present work, there are several open problems worth further study. For example, improving the proposed algorithms such that it has less sampling and training total time compared to XGBoost, at the same time keeping its smaller classification rates. Applying the resampling-based Boosting algorithm to other problems such as regression estimation and multi-class. All these problems are under our present investigation.