

**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD**

**VI Semester B.Tech in Information Technology**

**Report - Group Assignment 3**

Data Mining and Warehousing

**Generalization ability of SVM classification based on Markov Sampling**

**By :-**

Palak Mittal (IIT2018117)

Nehal Singh (IIT2018119)

Abhishek Kumar Gupta (IIT2018187)

Puja Kumari (IIT2018191)

Prabha Kumari (IIT2018195)

### **Abstract**

The previously known works studying the generalization ability of support vector machine classification (SVMC) algorithms are usually based on the assumption of independent and identically distributed samples. In this paper,we go far beyond this classical framework by studying the generalization ability of SVMC based on uniformly ergodic Markov chain (u.e.M.c.) samples. We analyze the excess misclassification error of SVMC based on u.e.M.c. samples, and obtain the optimal learning rate of SVMC for u.e.M.c. samples.

### **Why do we do Markov Sampling?**

There are many methods of sampling which are dependent (e.g., α mixing, β mixing and φ mixing) learned in machine textbooks. In this article, we focus only on the analysis where input samples are Markov's chains, the reasons are as follows. First, in real-world problems, Markov chain samples appear frequently and naturally in applications, such as biological analysis (DNA or protein), web-based content search, tag prediction, and so on. Second, ample strong evidence suggests that learning algorithms often work well with Markov chain samples (e.g., biological sequencing analysis, speech recognition). The reason for this, however, was unknown (in particular, it is not known how well it works in terms of general learning and practice). Or we can say that it is a normal approach.

**Algorithm**

For a given original training sample set Dtr, the new Markov sampling algorithm for SVMC is stated as follows.

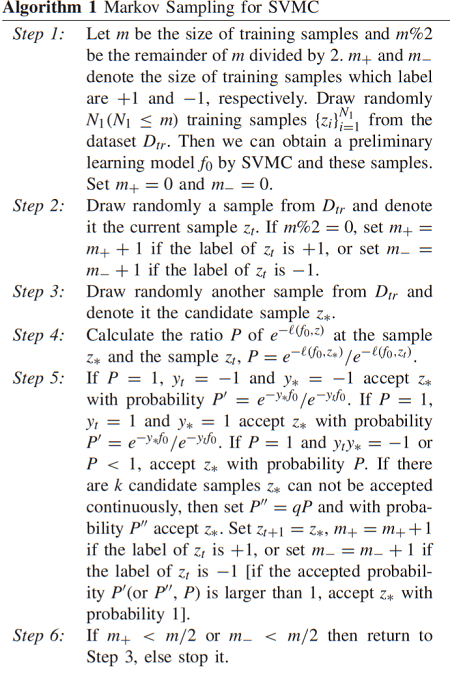
Remark 3: To have better understanding Algorithm 1, we present the following remarks.

1) We introduce the notions m%2, such that Algorithm 1 is suit to all the cases (even or odd) of training sample sizes. To generate quickly Markov chain samples, inAlgorithm 1 we introduce the continuously rejected number k and the constant q. Since as the loss (f,zt) of sample zt is smaller, the acceptance probability P =e−(f0,z∗)/e−(f0,zt) will be smaller. This implies that the

candidate sample z∗ will always be rejected, and generating u.e.M.c. samples is very time-consuming. In the following experiments, we take k = 5 and q = 1.2.

2) Since we have only the dataset Dtr, to define the transition probability of Markov chain, in Algorithm 1 we introduce the preliminary learning model f0. The reasonis that under the technical condition, we can compute easily the transition probabilities P (or P, P), and P, P

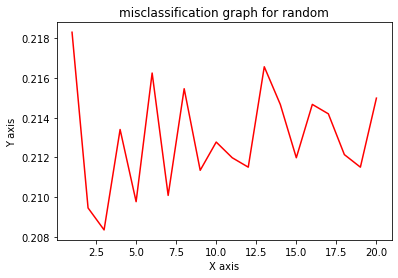
,P are always positive. Thus, by the theory of Markov Chain in [37] (if the size of state space of Markov chain is finite, and the transition probabilities of any two states are always positive, then the Markov chain is u.e.M.c.,please, we can conclude that {zi}ti=1 generated by Algorithm 1 is a u.e.M.c. sequence.

****

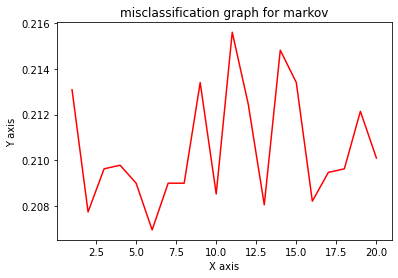
3) Different from MCMC method in, Algorithm 1 is a method of generating u.e.M.c. samples from a given dataset, and in Algorithm 1 we did not use the information of distribution of training samples since the distribution of samples is unknown. While MCMC is a sampling method of using the probability distribution of training samples. Compared random sampling with

Algorithm 1, we can find that random sampling can be regarded as the special case of Algorithm 1, that is, the acceptance probabilities P, P, P defined in Algorithm 1 are equal to 1.

**RESULTS**

****

Mean misclassification rate for random sample with std (21.297318611987382, 0.2561270889440842)



Mean misclassification rate for markov sample (21.04968454258675, 0.24574694356243504)

# **Citations**

[1] J. Xu et al., "The Generalization Ability of SVM Classification Based on Markov Sampling," in IEEE Transactions on Cybernetics, vol. 45, no. 6, pp. 1169-1179, June 2015, doi: 10.1109/TCYB.2014.2346536.