

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

VI Semester BTech in Information Technology

Report - Group Assignment 3

Data Mining and Warehousing

Group Members:

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

Title: The Generalization Ability of SVM Classification Based on Markov Sampling .

Author: Jie Xu, Yuan Yan Tang, Fellow, IEEE, Bin Zou, Zongben Xu, Luoqing Li, Yang Lu, and Baochang Zhang.

INTRODUCTION

In the assigned research paper, the authors consider using markov sampling to improve the accuracy of SVM-based classifiers in the case of a dataset with a large number of features. The paper is based on the premise that the data fed to an SVC is transmitted individually and identically (i.i.d.), and claims that this isn't always the case when dealing with data that is essentially temporal, such as demand forecasting and speech recognition. To substitute the random sampling that is widely used to train SVMs, the authors use a sampling approach called markov sampling, which is inspired by Markov chain Monte Carlo (MCMC) methods.

IMPORTANT TERMINOLOGY:

SVM:

SVM was created by Vapnik based on the Structural Risk Minimization principle. It is a binary classification algorithm that transforms data and finds the best boundary between the possible outputs depending on the transformations. This strategy is known as the kernel trick.

The following are some of the reasons why SVMs are important:

- Successful because there are a large number of features and a limited sample size.
- It is possible to learn both basic and complex classification models.
- Using complex statistical methods to prevent overfitting.

MARKOV CHAIN:

A Markov chain is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules. The defining characteristic of a Markov chain is that no matter how the process arrived at its present state, the possible future states are fixed.

MARKOV SAMPLING:

Markov chain Monte Carlo (MCMC) methods are a family of algorithms for sampling from a probability distribution used in statistics. By recording states from a Markov chain that has the desired distribution as its equilibrium distribution, a sample of the desired distribution can be obtained. The further measures there are, the more precisely the sample distribution fits the ideal distribution. The Metropolis–Hastings algorithm is one of several chain-building algorithms available.

ALGORITHM:

One of the simplest and most flexible algorithms making Markov's flexible chains is the Metropolis-Hastings algorithm. Since we cannot take a direct sample in the background (unknown) distribution, this algorithm uses auxiliary distribution that is easy to sample. These samples produce a change in the state of candidates, which is accepted or rejected as much as possible.

Result:

Accuracy of SVM based on markov sampling with different Kernels for letter-recognition dataset.

Kernel	KPCA	SVDD	OCSVM	OCSSVM	OCSSVM with SMO	MS_SVM
Linear	0.02	0.09	0.01	0.07	0.04	0.8523
RBF	0.05	0.07	0.14	0.09	0.04	0.938
Hellinger	0.01	0.02	0.02	0.13	0.10	0.824
chi_square	0.18	0.0	0.02	0.18	0.17	0.884

