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 SVMCs, 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:

1. Let m be the size of training samples and m%2 be the remainder of m divided by 2. m+ and m−denote the size of training samples which label are +1 and −1, respectively. Draw randomly N1(N1 ≤ m) training samples {zi} N1 i=1 from the dataset Dtr. Then we can obtain a preliminary learning model f0 by SVMC and these samples. Set m+ = 0 and m− = 0.
2. Draw randomly a sample from Dtr and denote it the current sample zt. If m%2 = 0, set m+ = m+ + 1 if the label of zt is +1, or set m− = m− + 1 if the label of zt is −1.
3. Draw randomly another sample from Dtr and denote it the candidate sample z∗.
4. Calculate the ratio P of e−(f0,z) at the sample z∗ and the sample zt, P = e−(f0,z∗)/e−(f0,zt)
5. If P = 1, yt = −1 and y∗ = −1 accept z∗with probability P = e−y∗f0 /e−ytf0 . If P = 1, yt = 1 and y∗ = 1 accept z∗ with probability P = e−y∗f0 /e−ytf0 . If P = 1 and yty∗ = −1 or P < 1, accept z∗ with probability P. If there are k candidate samples z∗ can not be accepted continuously, then set P = qP and with probability P accept z∗. Set zt+1 = z∗, m+ = m+ +1 if the label of zt is +1, or set m− = m− + 1 if the label of zt is −1

[if the accepted probability P (or P, P) is larger than 1, accept z∗ with probability 1].

6. If m+ < m/2 or m− < m/2 then return to Step 3, else stop it.

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	0.79
RBF	0.05	0.07	0.14	0.09	0.04	0.83
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	0.71