**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD**

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

**Report - Group Assignment 3**

**Data Mining and Warehousing**

**Group Members:**

**1. G. Shashank(IIT2018106)**

**2. M J Akhil Naik(IIT2018143)**

**3. A Rahul Naidu(IIT201192)**

**4. Gurtej Singh(IIT2018193)**

**Title:**  Generalization Ability of the SVM Classification Based on the Markov Sampling . **Author:** Jie Xu, Yuan Yan Tang, Fellow, IEEE, Bin Zou, Zongben Xu, Luoqing Li, Yang Lu, and Baochang Zhang.

**INTRODUCTION**

In the given research paper, authors considered using of the markov sampling for the improvement in the accuracy of SVM-based classifiers in the case of dataset with a large number of features. Paper is actually based on the premise that the data fed to an SVC is transmitted individually and identically (i.i.d.),and claims that this is not 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 is created by Vapnik, based on the Structural Risk Minimization principle. It is a binary classification algo that will transform data and finds the best boundary between the possible outputs depending on the transformations. This is also known as kernel trick.

SVMs are important because:

• It is successful because there are a large no.of features and limited sample size.

• In this, it is possible to learn both the basic and complex classification models.

• In this, we use complex statistical methods to prevent overfitting.

**MARKOV CHAIN:**

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

**MARKOV SAMPLING:**

Markov chain Monte Carlo methods are family of algos for sampling from a probability distribution. 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 algorithms making Markov's flexible chains is the Metropolis-Hastings algorithm. We can’t take a direct sample in background (unknown) distribution, so this algorithm will use 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