Data Mining and Warehousing - Assignment 2

Indian Institute of Information Technology, Allahabad

IIT2018176	*IIT2018178*	*IIT2018179*	*IIT2018202*
Milan Bhuva	*Manav Agarwal*	*Mohammed Aadil*	*Ankit Rauniyar*

Abstract: Support Vector Machine is a supervised machine learning algorithm that can be used for both classification and regression. However, it is mostly used in classification problems. SVM works on the concept of finding the optimal hyperplane that best classifies the given data. An optimal hyperplane is the one that maximizes margins from both classes. SVMs can be used against both linear and non-linear data. SVM makes use of tuning parameters namely Kernels, Regularization, Gamma, and Margin. These parameters can be tuned in order to receive high accuracy

I. INTRODUCTION

Basically, support vector machines are a way of classifying points by building separating hyperplanes between them. Kernels determine how SVMs work because they allow the SVM to consider candidates separating hyperplanes in very high dimensions. A **kernel** is a specialized kind of similarity function. It takes two points as input, and returns their similarity as output, just as a similarity metric does.

This is what our assignment mainly focuses on. We, as a group, were given the task to classify and compute the accuracy of SVM using Gaussian Kernels performed on two datasets: PASCAL VOC 2012 and Letter Recognition. It is important to get a grip on SVM using gaussian kernels before moving on. We have also utilised Tsybakov's noise assumption and local Rademacher averages to establish learning rates up to the order of 1/n for nontrivial distributions.

Tsybakov's noise condition is used to describe the amount of noise in the labels. The function $|2\eta - 1|$ can be used to describe the noise in the labels of a distribution P. Indeed, in regions where this function is close to 1 there is only a small amount of noise, whereas function values close to 0 only occur in regions with a high level of noise.

Now what we understand by the rademacher average of a class of functions is, it is a capacity measure that can be interpreted as the ability of the functions in the class to estimate random noise. Using these 2 variables on SVM with gaussian kernels we can obtain a faster and a further accurate estimate than we can without these 2.

II. DATASET DESCRIPTION

We have used 2 datasets for training and classification purposes. One of these papers is provided by UCI and the other one dataset was created for Visual object classes challenge and is commonly referred to as Pascal2.

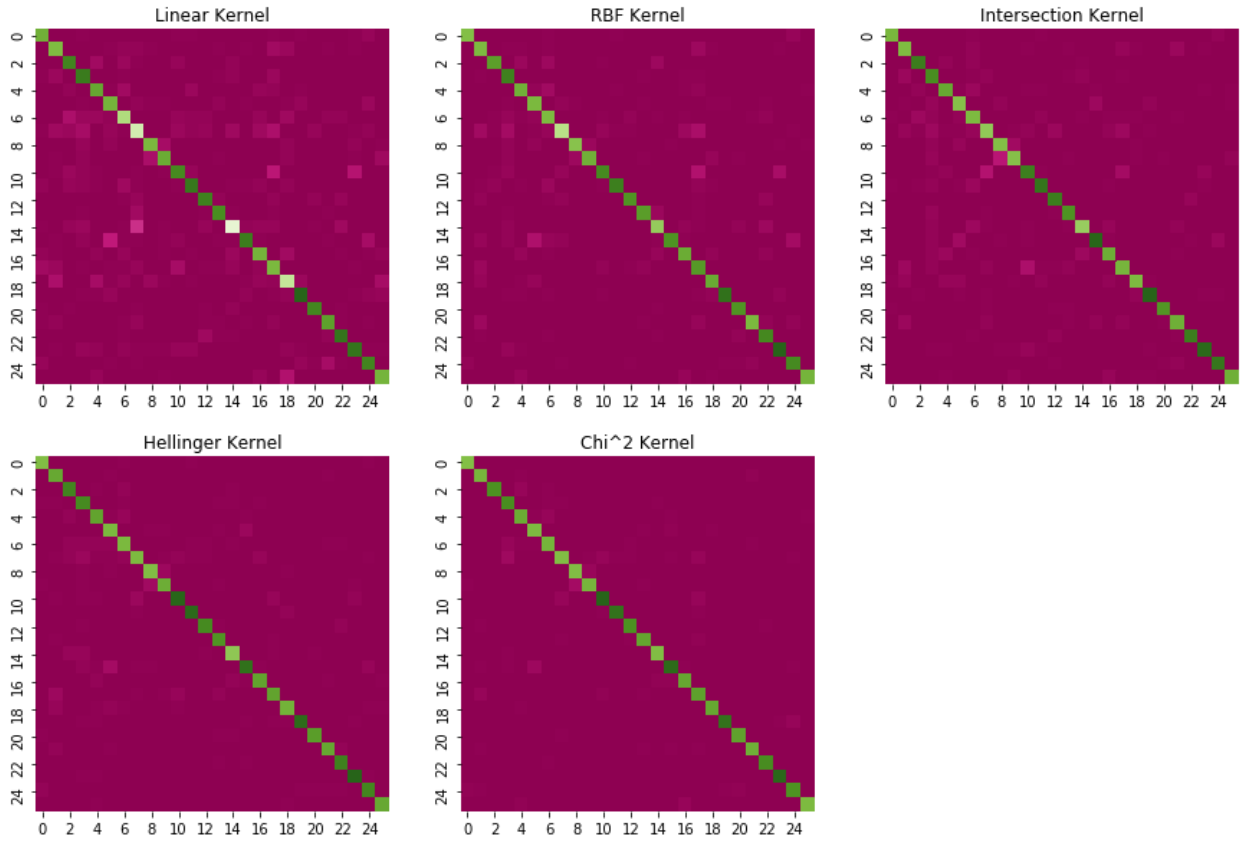
The first dataset consists of 20000 instances of 26 Capital letters in the english alphabet. The images are based on 20 different fonts where each letter within these 20 fonts is randomly distorted to produce 20,000 unique stimuli. We train our SVM on the first 16000 stimuli and use the rest 4000 to test our Model.

The second dataset consists of random objects in images. This dataset 's main purpose is to teach the model to recognise objects in an image. The training dataset consists of 20 classes of images and 11,530 images in total and 27,450 ROI annotated objects with 6,929 segmentations.

III. RESULTS

Kernel	KPCA	SVDD	OCSVM	OCSSVM	OCSSVM with SMO	MS_SVM
Linear	0.02	0.09	0.01	0.07	0.04	0.14
RBF	0.05	0.07	0.14	0.09	0.04	0.07
Intersection	0.18	0.01	0.04	0.26	0.22	0.07
Hellinger	0.01	0.02	0.02	0.13	0.10	0.04
χ2	0.18	0.00	0.02	0.18	0.17	0.02

IV. CONCLUSION



As mentioned in assignment 1, The above image confusion matrices is obtained from using different kernels on Fast SVM. And the linear kernel again consists of a lot of false positives and false negatives, however the chi squared kernel gives a confusion matrix with very fewer number of false positives and negatives, i.e. more accurate than the linear kernel. Also, using Fast SVM along with chi squared kernel has improved the accuracy over traditional models while working faster as noticed in results.