Executive Summary Report of the Minor Research Project

Submitted to the

University Grants Commission

(Ref No. 1784-MRP/14-15/KLMG002/UGC-SWRO dated 04-02-2015)

By

BLESSON GEORGE

DEPARTMENT OF PHYSICS

CMS COLLEGE, KOTTAYAM

KERALA– 686 001, INDIA

FEBRUARY 2017

Executive Summary Report of the Minor Research Project

PI: BLESSON GEORGE

(Ref No. 1784-MRP/14-15/KLMG002/UGC-SWRO dated 04-02-2015)

Dimensionality reduction is a frontier research area at the intersection of various disciplines, including statistics, databases, data mining, text mining, simulation, pattern recognition, machine learning, artificial intelligence, climate studies, biophysics and optimization. Each of these areas approach this problem in different ways. For example, in pattern recognition the problem of dimensionality reduction is to extract a small set of features that recovers most of the veracity and variability of the data. In text mining, however, the problem is defined as selecting a small subset of words or terms. The application of dimensionality reduction also varies with the application domain. Examples of applications of dimensionality reduction techniques include: mining of text documents, gene structure discovery, image processing, statistical learning, and exploratory data analysis.

Real-world data, such as speech signals, digital photographs, or natural language processing usually has a high dimensionality. In order to handle such real-world data adequately and accurately, its number of dimension needs to be reduced. Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality.

In this work we tried to apply dimensionality reduction techniques on different datasets before classifying them using SVM and kNN classifiers. Development of algebraic machine learning algorithms for understanding symmetries in data patterns by making use of Fourier transforms and Group theory as basic tools is of great significance. Techniques such as PCA, Random Projection (RP) and Feature selection using selectkbest make use of transforms and finding variances in various cases. These dimensionality reduction techniques select the best features and helps in improving the classification

The project work verified the importance of Random Project as an effective dimensionality reduction tool by studying the results of it in various cases. The research works which dealt with wide variety of datasets and used RP as the dimensionality reduction tool proved that RP is an emergent tool for dimensionality reduction.

In the second part of the project, we compared performance of different techniques based on RP, and experimental results proved that the classification accuracy of RP can be improved by combining with other dimensionality reduction methods, such FS orPCA. However, it didn't yield better classification accuracy combining RP with PCA. RP followed by PCA outperforms other methods in classification accuracy on both the data sets with two classifiers.

Mutual Information score is the measure of mutual dependence between the output and input variables. It studies the symmetry patterns in datasets. In dimensionality reduced datasets, it is expected to have a low mutual information score. The measure is done for different cases and proved that symmetry in data sets is considered for dimensionality reduction and hence we obtained very low values of Mutual information score.