**FINAL PROJECT REPORT**

**Team: Hey Prabhu!  
ISEN 613 (Engineering Data Analysis) Spring 2017**



**TITLE**

**Analysis by classification of Radar Signals Returned from Ionosphere**

Instructor: Dr. Li Zeng

*Submitted by: Aasheet Kumar*

*Abhijeet Shinde*

*Dwarkanath Prabhu*

*Shreyas Sasle (624009630)*

Contents:

Abstract:

In the study of radar signals in ionospheric research it is necessary to check whether the signals returned from radar are useful for further analysis or not. Based on the available results obtained in this dataset the main objective is to find out which of these signals can be termed as significant. While some signals get lost through the ionosphere, some of them contains vague and irrelevant information in the form of noise. For this distinction of the bad signals from good ones the classification artifacts can be applied on the radar generated signals. To evaluate the performance characteristics of these signals and their usefulness for further analysis the model diagnostics can be applied to check the adequacy of data signals. With these measures of accuracy, sensitivity, specificity and proportion of variance identified and proved to be viable enough, these signals can be used for further analysis.

Problem Statement:

To develop a model to be able to classify radar returns from the ionosphere as either “Good” or “Bad” radar signals, “"Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not”. Looking at the response, we continue with classification methods on the dataset.

The Dataset:

The dataset consists of 34 attributes “There were 17 pulse numbers for the Goose Bay system. Instances in this database are described by 2 attributes per pulse number,” with 351 replications of data, which the response being a binary one with outputs either “good” or “bad”.

Literature Review:

The data set used has been used by various researchers as an experimental data set to train and test their methodologies. The methods used by the researchers are conventional models covered in ISEN 613 or their extensions. The data set is mainly of a signal processing unit and hence perform best when combined with statistical approach of ACF as proposed by “Sigillito, V. G., Wing, S. P., Hutton, L. V., \& Baker, K. B. (1989)”. Hyunsoo Kim and Se Hyun Park propose an alternative approach for data reduction in SVM to save computational time “Data Reduction in Support Vector Machines by a Kernelized Ionic Interaction Model. SDM. 2004”. This is applicable to very large data sets having a greater count of attributes and needs to be feature extracted. The chosen “Ionosphere” data set has comparatively smaller and with fewer attributes to have high computational time. Additional combinational classification techniques are proposed by Marina Skurichina and Ludmila Kuncheva and Robert P W Duin in “Bagging and Boosting for the Nearest Mean Classifier: Effects of Sample Size on Diversity and Accuracy. Multiple Classifier Systems. 2002” and “Boosting in Linear Discriminant Analysis. Multiple Classifier Systems. 2000”. These techniques provide good insight of different classifiers used for analysis along with their combined effects. The former gives statistical overview and application algorithm to study the sample size effect on classifier accuracy while later provides an approach to boosting in LDA and corresponding statistical improvements. The training data for “Ionosphere” data set has good number of observations as compared to predictors and can be evaluated with or without combinational techniques.

There is abundant literature available for readers for learning exploratory data analysis followed by classification models and their improvements. However, the approach followed by this project group is in accordance to learning of ISEN 613 course and pertinent to the chosen data set.

The Proposed Approach:

The first step in working with a dataset is making sure it is clean and fit to work with. All the observation from the radar signals are not useful. Here we make sure there weren’t any missing values and deal with any which come up. Then reduce the number of variables using PCA, we also removed an attribute which had the same values for all the observations/replications. After we have the appropriate number of attributes we need we continue to perform various classification methods on the dataset like Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), K nearest neighbors (KNN), logistic regression, bagging, random forest, support vector machine [SVM (polynomial and radial)].