PROJECT PROPOSAL

Analysis by classification of RADAR signals returned from ionosphere.

Introduction and motivation:

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 “good”.

While some signals get lost through the ionosphere, some of them contain vague and irrelevant information in the form of noise. Thus, “good” radar signals are those showing evidence of some type of structure in the ionosphere, “bad” signals are those that do not. For this distinction of the bad signals from the good ones, classification methods can be applied on the radar generated signals.

Thus, the problem then becomes to identify the classification model that most accurately predicts quality of signal based on pulses received from the ionosphere by antennas in the radar system.

Deliverables:

The data set we will use consist of 34 attributes. These describe the pulses received by antennas in the radar system at Goose Bay, Labrador. “There were 17 pulse numbers for the Goose Bay system. Instances in this database are described by 2 attributes per pulse number”. There are 351 observations of data, with the response being either “g” or “b” standing for “good” or “bad”.

As this is a classification problem, we would use all the classification algorithms on the data set and conclude which algorithm would provide the max accuracy with least error rate.

Methodology:

1. The first step in working with a dataset is making sure it is clean and fit to work with. All the observations from the radar signals are not useful. Hence we first make sure there weren’t any missing values and deal with any which come up.
i.e. we would apply Principal Component Analysis to determine the major attributes that explain most of the variance in the data.

2. After we have the appropriate number of attributes, we continue to perform various classification methods on the dataset viz.

- K Nearest Neighbors (KNN)

- Logistic Regression
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Bagging
- Random Forest
- Support Vector Classifier (SVC)
- Support Vector Machine [SVM (polynomial and radial)].

Resources:

1. https://archive.ics.uci.edu/ml/datasets/ionosphere to provide the training dataset.

2. Understanding of the different classification algorithms used.

Milestones:

1.Importing the data set, running Principal Component Analysis on the data and implementing KNN and Logistic Regression on the data (by end of October).

2.Implementing and evaluating Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Bagging algorithms on the dataset. (By mid-November).

3.Implemeting Random Forest, Support Vector Classifier (SVC) and Support Vector Machine [SVM (polynomial and radial)] on the dataset, comparing the prediction accuracy of all the classification algorithms and stating the best classification model to be used on the ionosphere dataset. (By end of November)

References:

1. Sigillito, V. G., Wing, S. P., Hutton, L. V., & Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, *10*(3), 262-266
2. Kim, H., & Park, H. (2004, April). Data reduction in support vector machines by a kernelized ionic interaction model. In *Proceedings of the 2004 SIAM International Conference on Data Mining* (pp. 507-511). Society for Industrial and Applied Mathematics.
3. Skurichina, M., Kuncheva, L. I., & Duin, R. P. (2002, June). Bagging and boosting for the nearest mean classifier: Effects of sample size on diversity and accuracy. In *International Workshop on Multiple Classifier Systems* (pp. 62-71). Springer Berlin Heidelberg.

4. Skurichina, M., & Duin, R. P. (2000, June). Boosting in linear discriminant analysis. In *International Workshop on Multiple Classifier Systems* (pp. 190-199). Springer Berlin Heidelberg.