FLIGHT QUEST: A BETTER FLIGHT PLAN FOR PILOTS

CIS 6930: Large Advanced Data Analysis

Course Project

Spring 2013

Ashwini Ramamurthy - 11171154

Kishore Rajasekar - 34545722

Ravindra L D - 43920672

Abstract

Flights have become commonplace in our lives but the efficiency is not as good as it can be. Gate conflicts, operational challenges, air traffic management, dynamics of a flight can change quickly and lead to costly delays. Airlines spend around 22 billion dollars on improving the efficiency. Real time big data analysis is the solution to this problem. Pilots could augment their decision making process with this real time information given to them.

To reduce the delays and to make the decision making more accurate and efficient, we have come up with a module which will predict the delays in flight considering various factors. Large real time data related to flights, airports, weather are combined, analyzed and made into a model .Future predictions are made based on this model.

Contents

Introduction

Flights never stick to their schedule. The reasons for the delays vary from gate conflicts, operational challenges, air traffic management, dynamics of a flight, weather etc. All these factors come together and reduce the overall efficiency of the flight. The airlines spend a lot of money on management of flight plan and to make it more efficient. Thus, Airlines are looking for a better model that would help pilots make a better decision and reduce the overall delay of the flights.

Our main objective was to help in pilots decision making by providing real time business intelligence by predicting delays .To achieve this we removed noise from the data, we then gave a structure to the raw data and related all the tables which contribute to the delay calculation .After which we extracted the important parameters and classified the data into different classes which were defined by their delay time.

We all contributed equally to every component of the project. As a lot of ground work and research had to be done, and everything was new to us, we divided the work equally.

Background

Basic Concepts and Algorithms

Given all the relevant data, the problem can be approximated into a classification problem with delay categories. We used PCA to refine our parameters and trained a SVM classifier on the data. As this was a multi-class problem, we used the 'One vs. all' paradigm of building models.

Related Work

The technologies that are used currently do not use real time data, and the route is checked and other parameters and calculated 24-48 hours before the flight.

IFR (Instrument File Route) Planning

Special charts are made to show IFR routes from beacon to beacon with lowest safe altitudes(LST).The main disadvantage of this planning is that the route can be changed by the pilot but the pilot has to make the calculations himself which is not always efficient.

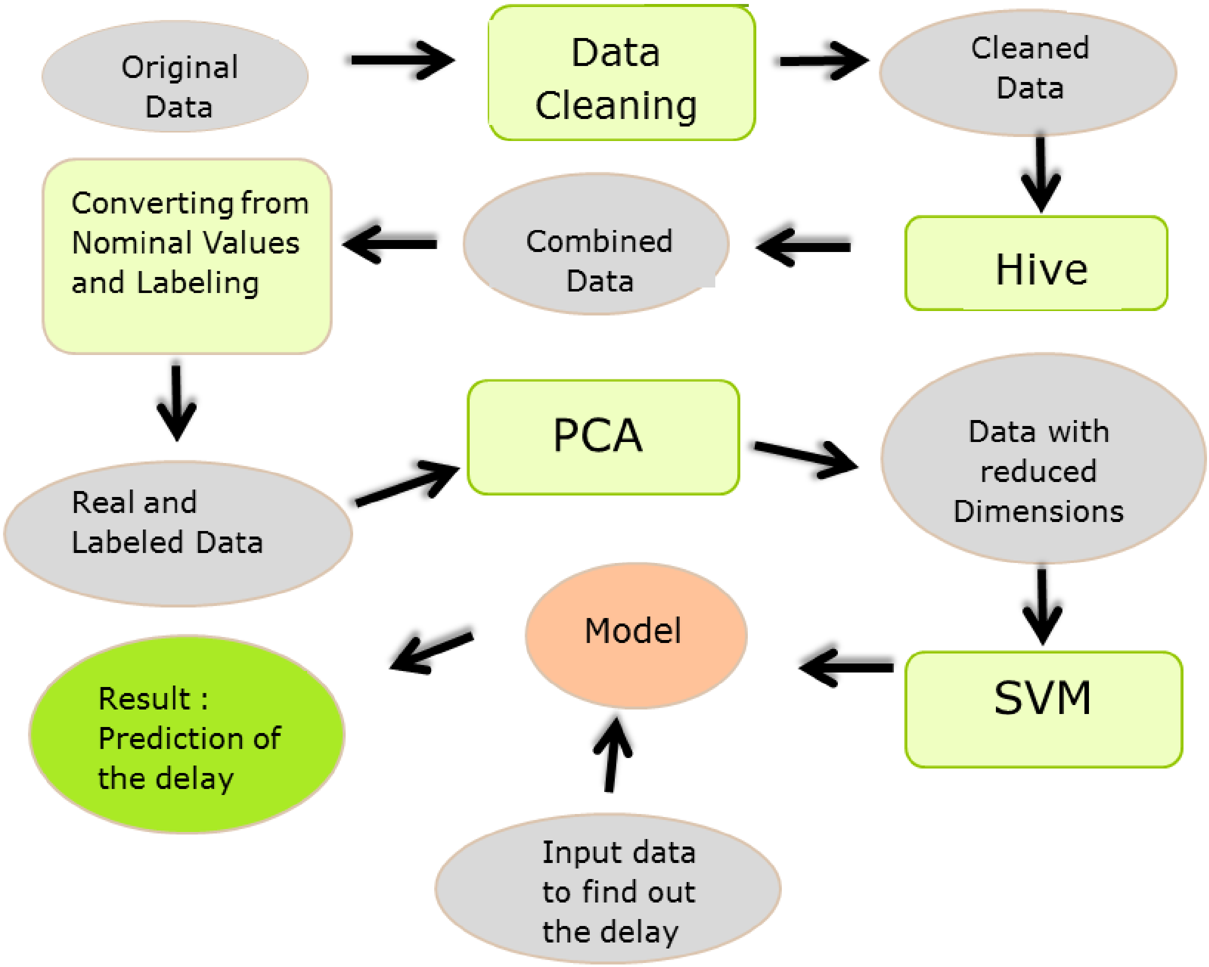
Performance Based Navigation (PBN).

This kind of planning encourages evaluation of the overall accuracy, integrity, and availability of navigation aids present. This is more recent and is better than the IFR planning as this the operator develops a route which is more time and fuel efficient.

The main disadvantage of this planning tool is that once the headings, heights, and speeds are calculated it cannot be changed by the pilot. For example if there is a weather change or some kind of delay. The pilot cannot change the route or change any other parameter to reduce the delay.

Our project /model makes a better prediction as we consider the main factors which contribute to delay and we train our model with this real time data. This model can be used 1-2 hours before the flight and by this the pilot can make better decisions.

System Description


Original Data

Data Cleaning

Cleaned Data

Hive

Combined Data

Converting from Nominal Values and Labeling

Real and Labeled Data

PCA

Data with reduced Dimensions

SVM

Model

Input data to find out the delay

Result : Prediction of the delay

Data Cleaning

Initial data contained lot of missing and hidden fields but since the data was in the structured form, we used simple unix shell scripting with rule based filtering mechanism to remove missing and hidden data, and also the data related to those sample in other relevant tables.

Hive

Hive was primarily used to combine the related information from various tables in one place, so that we can apply feature extraction and classification algorithms on the data efficiently.

We choose Hive to do this task because the samples in each of the tables were really huge and Hive does joining of data from different tables in a distributed and very efficient way.

Java module

Since all the fields in our data were in string form and the machine learning algorithms for feature extraction and classification requires the data to be in numeric form, we built our own java module to do this task.

We extracted unique values from each column of the joined data and mapped them to unique real numbers. We also maintained the reverse mapping from numbers to strings, so that once the prediction is done, we can reverse map them to the string values to get the original sample. We also labeled the samples based on the difference between scheduled arrival time and actual arrival time. There were total 11 labels like 0 to 20 minutes delay, 20 to 40 minutes delay and so on.

Feature Extraction

After cleaning and joining of data there were 23 features in total. It was difficult to classify the data with such high dimensionality. So we ran Principal Component Analysis to reduce the dimensionality of the data. We were able to reduce the features from 23 to 13 which contributed to 95% relevance.

Classification

Classification was done using Support Vector Machine algorithm. We implemented a Multi-way classifier using 11 separate models, using RBF kernel. The 11 models were based on the 11 labels we generated. Without tweaking any parameters, we were able to achieve only 28.3% accuracy in classification. When we scaled and centered the data and enforced the convexity using square transform there was increase in the accuracy. We found that input data was locally grouped, so we randomized the samples and increased the count for the positive samples and were able to achieve accuracy of 67%.

Prediction

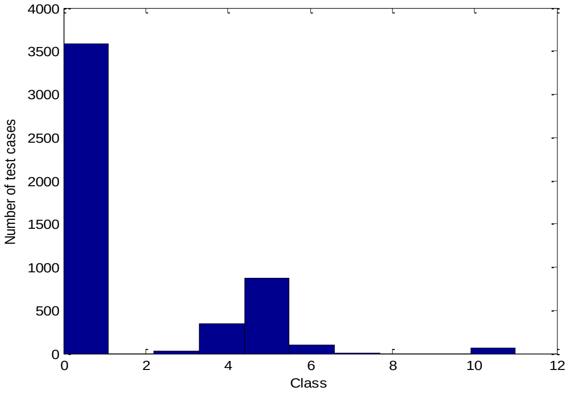
Form the output of PCA module, 20% of data was used to train the model, close to 60% of the data was used to do cross validation, and 20% of data was used to test the model .We get an average accuracy of 67% upon running model on 10 different samples of test data. The error rate was more for extreme classes.

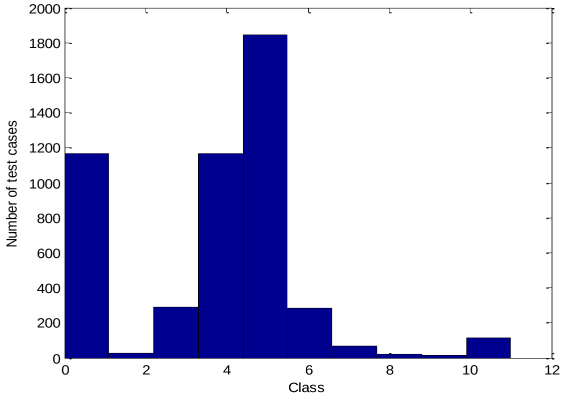
Final Data products

- Real-time calculation of delay for a given flight plan based on current weather, air-traffic, and de-icing parameters and so on.
- Usage of this product enables pilots / airlines to make plans with minimum possible delay, thereby increasing revenue and customer satisfaction.

Experiments

- Data had information related to various aspects of flights like Aircraft Situation, Flight history, Air Traffic Control, Weather.
- Data was obtained from FlightStats,Inc, U.S Department of Transportation and from METAR system.

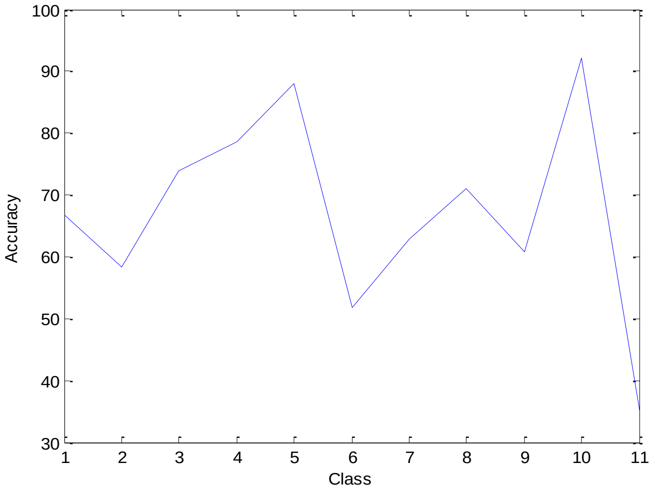




Initial accuracy was only 28.30%. The classification on test data was very poor, because training data was locally grouped.

Accuracy improved to 65.56% after tweaking some modeling parameters and random shuffling of the training sample to break the local grouping present in data.

Fringe classes still have high inaccuracy rates due to highly skewed training data (example: class 1 had <1% positive samples).



This graph shows the accuracy variation among classes (final result). Fringe class still has poor classification but we see high accuracy >80% for class 4 and 5 which are the highly populated classes.

Future work

- Improvements can be achieved by increasing training data size, as this will lead to increase in feature set and reduction in chances of over-fitting.
- Other possible routes can be calculated to reduce delay.

Conclusion

After collecting the flight related data over a long period of time and analyzing the data, we have constructed an approximate model to predict delays on particular route given the current air traffic, weather and other conditions.

This is a definite improvement over current systems and the real-time information produced is valuable in re-calculating routes for critical flights and to inform the passengers of an almost exact amount of delay if present.