AI project summary– PCA and feature selection

Greeshma Gopal(gopalg)

Data set description

The data set chosen is related to the detection of high energy seismic bumps in the coal mines. One of the major threats faced in these coaling mines are the seismic hazards. Seismic hazard is nothing but the probability of earthquake occurrence in a particular geographical area at a given time. Predicting the seismic hazard is highly significant to help make a decision in a given area of mine. If the high energy seismic bumps are expected to happen in a specific area its crucial to evacuate the crew from the endangered zones or stop the mining process. The data has been fetched from the UCI machine learning repository. The data acquisition is carried out at a geophysical station (Fig 1).

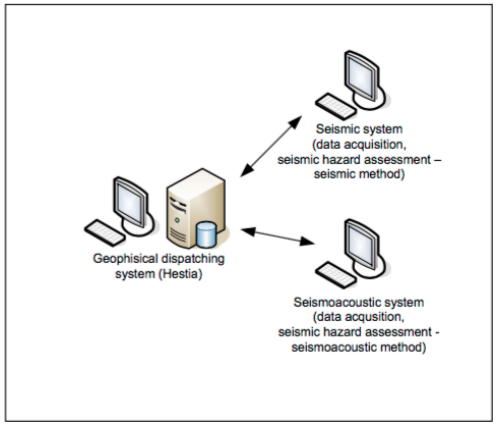

Fig 1: Data acquisition

There are 3604 data sets and 19 attributes associated with the same. The decision/target attribute is the 'CLASS' which is the supposed to be predicted. '1' means that high energy seismic bump occurred in the next shift ('hazardous state'), '0' means that no high energy seismic bumps occurred in the next shift ('non-hazardous state'). This data if predicted will aid in lowering the degree of hazard. The important factors affecting this hazard include the genergy(Seismic energy recorded within previous shift by the most active geophone (GMax) out of geophones monitoring the longwall), gpuls(a number of pulses recorded within previous shift) and gdenergy(a deviation of energy recorded within previous shift by GMax from average energy recorded during eight previous shifts). Also the nbumps which is the number of the seismic bumps based on the energy range.

Experiments conducted

Classification approach was adopted and an initial base model fitting considering all attributes was done to find the overall accuracy score. The main dataset was split into test (25%) and train (75%) data before the data was fitted. After each model was fit the graph was plotted for features and accuracy scores. Confusion matrix was also derived after each fit.
PCA and backward selection were done with the below model fits:

- K- nearest neighbors classifier(KNN)

- Random forest classifier
- Naïve Bayes' classifier

As an initial step the above models were fit with scaled data and below are the accuracies calculated.

Base model fit	Accuracy score
Logistic regression	75.1%
K- nearest neighbors classifier(KNN)	90.8%
Random forest classifier	97.4%
Naïve Bayes' Classifier	72.0%

Results

Below are the some of the observations from the model fittings which were done:

Analyses of the PCA fit

- Out of the three classifiers implemented, it was found that Random forest showed the highest accuracy where as Naïve Bayes' was giving the lowest accuracy score (Fig 2).
- The highest accuracy score was 97.3% for the first 12 features after which the accuracy started dropping. It can be concluded that these features (13-18) are impacting the dataset.
- On the other hand, KNN was giving the highest accuracy up to 14 features and the accuracy score began to decline.
- Naïve Bayes' had the lowest accuracy from the start and was not a good model to fit in the given dataset.

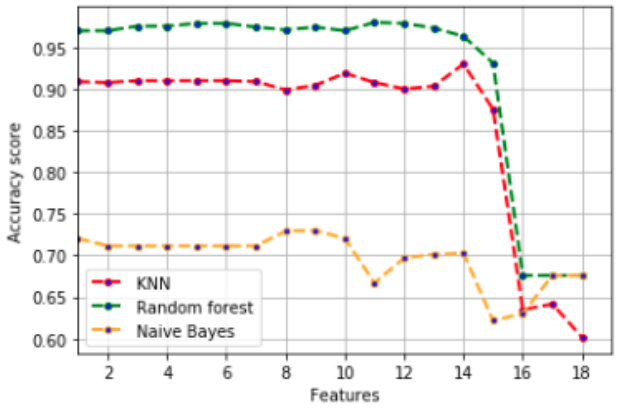

Fig 2: Model fit(PCA)

AI project summary– PCA and feature selection

Greeshma Gopal(gopalg)

Analyses of the backward feature selection

- When the dimensions were reduced using backward feature selection, Random forest fit showed the highest accuracy score compared to the others. The score was high (97%) when the number of features were reduced to 13 after which the score started dropping down (Fig 3).
- It can be also inferred that Naïve Bayes' shows the least accuracy compared to the other models.
- On the other hand, KNN was showing good accuracy (maximum of 93.0%) but started declining when the features were reduced up to 5 in backward dimensionality reduction.

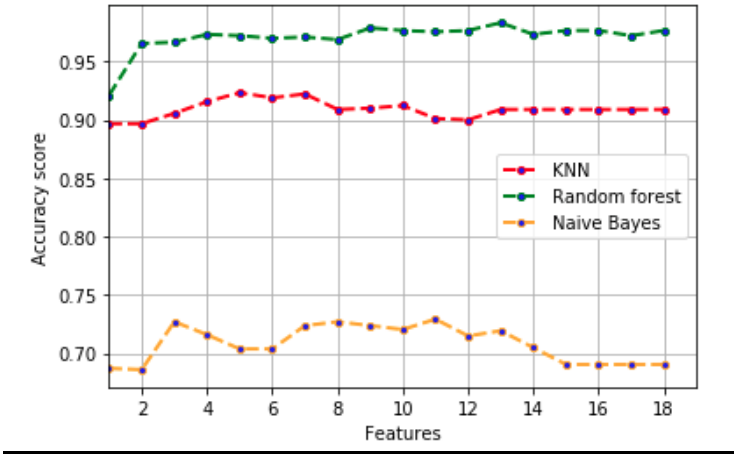

Fig 3: Model fit (Backward selection)

Test and train computational times

- The computational time was calculated for each model when fitting the test and train data. The times were calculated for both PCA and backward feature selection.
- In KNN, it was observed that the test data computational time is higher than the train data. This is because train data is used to store data, while in test data it will classify the labels which are more frequent among training samples.
- However, in Random forest model fitting and Naïve Bayes' model fits the test time data is lesser than train data.
- The trend is the same for all the models during PCA and the backward selection.
- We can also observe that there is a slight increase in the computational time as the number of components increases in every classifier model fit. This is one of the major lessons learnt from this project.

AI project summary– PCA and feature selection

Greeshma Gopal(gopalg)

References

- https://pdfs.semanticscholar.org/3e13/0e2ab1fa869725b4ad18ca100132328c6e55.pdf
- https://www.sciencedirect.com/science/article/pii/S1365160915001240
- https://link.springer.com/article/10.1007/s11600-016-0002-9