ESA Challenge: Collision Risk Predictions

Mar Cols-Margenet

Aerospace Engineering Sciences, University of Colorado Boulder

Abstract

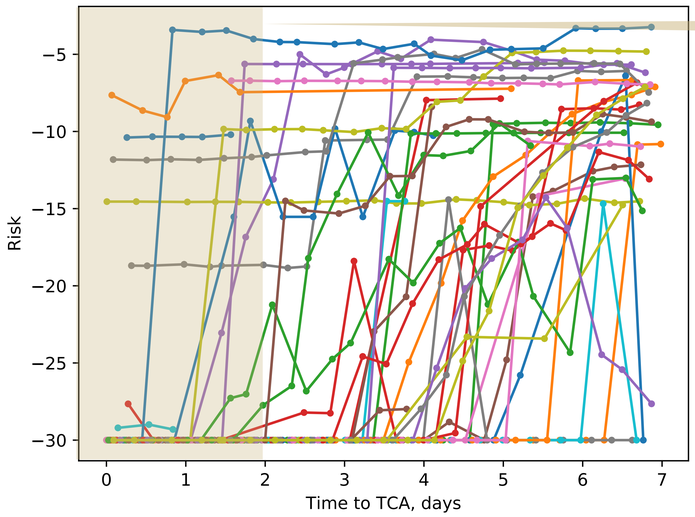
This projects consists on **building a model to predict the final collision risk estimate between a given satellite and a space object** (e.g. another satellite, space debris, etc). To do so, ESA provides access to a **database of real-world conjunction data messages** (CDMs). Main objectives are: 1) find out which CDM features are most relevant for risk prediction, 2) train a model with accurately known risks, 3) test the classification performance with a realistic test set.

Motivation & Background

This is a collision avoidance challenge posted by the European Space Agency (ESA), with the aim of leveraging complex manoeuvring decisions through machine learning. Link: https://kelvins.esa.int/collision-avoidance-challenge/. Only high risk approaches will trigger a manoeuvre.

Collision risks at time of closest approach (TCA) are to be predicted, 2 or more days in advance, using previous available data messages (CDMs).

- Each CDM (data row) contains **103 features** (columns)
- Each collision event consists of a CDMs time-series
- Important conceptual features:
  - **Event ID**: unique number for each close approach
  - **Time to TCA**: days left until closest approach [0, 7]
  - **Risk**: current risk estimate [-1, -30] — more negative = more risky
- **Plot**: evolution of risk vs. time-to-TCA for several events



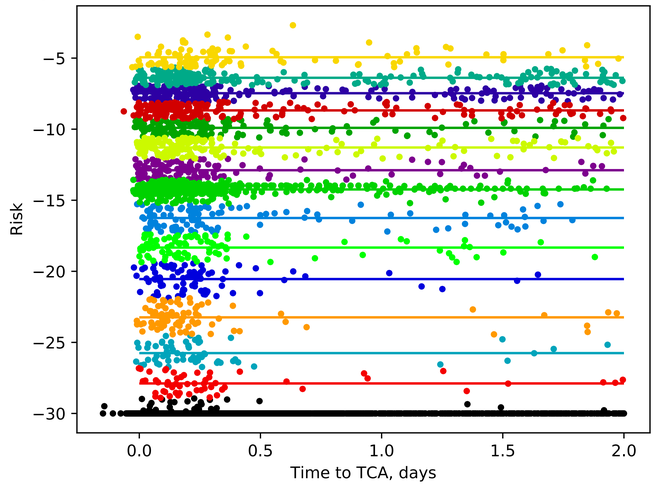
Can you predict latest risks based on **prior history**?

Method

Feature selection: k-means clustering, tree-classifier & Gini importance. **Dataset condensation**: data filtering & time-series condensation. **Model training & risk classification**: nonlinear SVM and neural network.

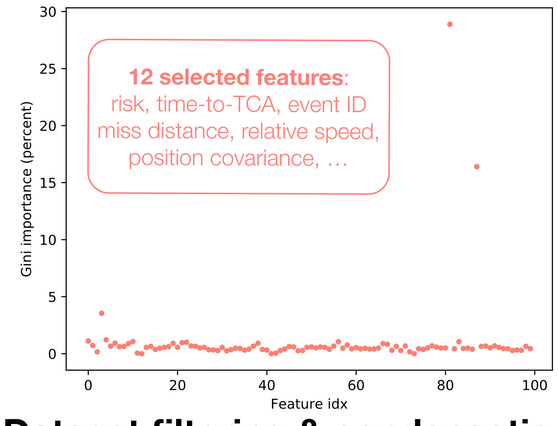
Feature selection

K-means clustering: data = latest available CDM for each event, feature = risk, k = 15 clusters. Clustering generates the labels to train a tree.



Tree-based classifier: data = latest available CDM for each event, features = all, labels = cluster groups.

Gini index: which of the 102 features (apart from the risk itself) are relevant for classification?



Dataset filtering & condensation

Only events with more than one CDM are kept. For each CDM, only the 12 selected features are preserved

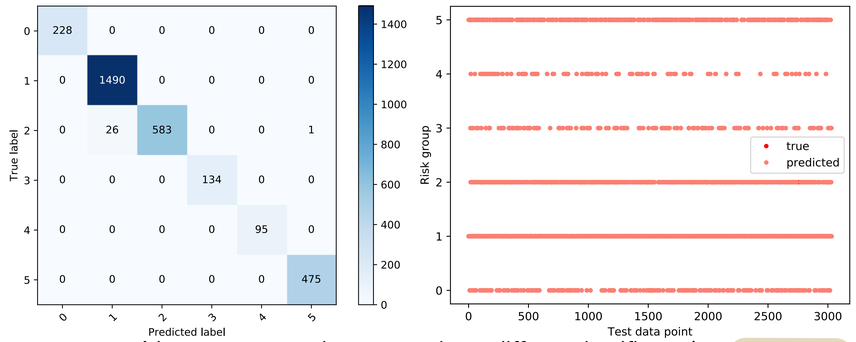
	CDM last	CDM last-1	CDMs previous
x = [x1, x2] **y = label**	risk defines label **y**	**x1** = selected features	**x2** = mean(risk), var(risk)

Model selection

To facilitate interpretation results, final risks are clustered into 6 classes.

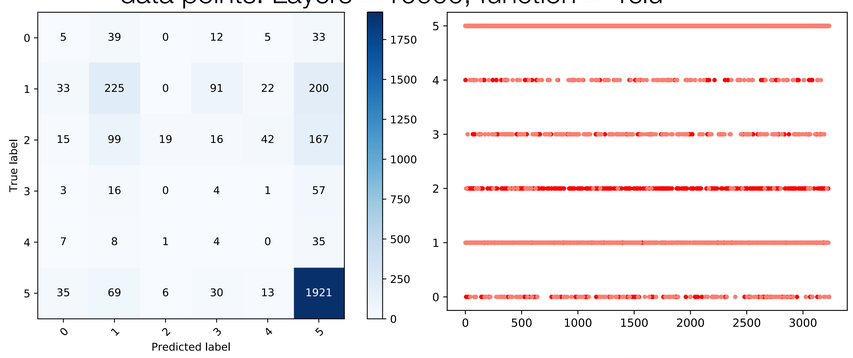
Non-linear SVM with RBF kernel:

- Great when test data is **similar** to train data



- Not so great when test data differs significantly — not shown

Neural Network: better than SVM at extrapolating to **new** data points. Layers = 10000, function = "relu"



- NN is biased towards groups 1 and 5. These are indeed the most represented groups in the training set

Results & Conclusions

The CDMs database provided by ESA has been successfully filtered, condensed and used for training a model that is capable of predicting final risks.

- The features selected by Gini importance dovetail well with field-specific knowledge
- For each time-series, previous risk history is captured in terms of mean and covariance
- Alternative to condensation would be slicing time windows (future work)
- While SVM works really well in specific cases, NN is better at handling novelty
- NN performance could be increased by boot-strapping samples from underrepresented groups