Fast Maritime Anomaly Detection using Kd-Tree Gaussian Processes

By Julia Will, Leto Peel, Christopher Claxton
Advanced Technology Centre, BAE Systems, Filton, Contact: julia.will@baesystems.com

Abstract

This work considers the challenge of scalability in the detection of anomalies in world-wide shipping traffic. A model of normal vessel behaviours is useful for detecting abnormal, illegal, suspicious, or unsafe behaviour; such as vessel theft, drugs smuggling, people trafficking or poor sailing. This work presents a state-of-the-art non-parametric regression model, based on Gaussian Processes, and is demonstrated to model normal shipping behaviour. This model is constructed from Automatic Identification System data and allows a measure of normality to be calculated for each newly-observed transmission according to its speed given its current latitude and longitude. Using this measure of normality, ships can be identified as potentially "anomalous" and prioritised for further investigation.

1. Introduction

Detecting anomalous behaviour is important as it can be indicative of nefarious behaviours, such as piracy, drug smuggling, arms trading, people trafficking and illegal immigration. In some cases, anomalous behaviour could pose a safety risk. The large numbers of vessels on the seas makes inspecting vessel tracks for anomalous behaviour time consuming and error prone for human analysts. Furthermore, with satellite-based Automatic Identification System (AIS) receivers coming online (as reported in Davenport, M. (2008)), the amount of data will only increase. Automated tools are essential to reduce the cognitive workload of the analysts.

One of two routes can be taken for automatically detecting anomalous behaviour: Either the behaviours can be codified by Subject Matter Experts (SMEs) based on their experience and domain knowledge; or the behaviours can be learnt from data.

Davenport, M. (2008) ran a workshop with SMEs to elicit abnormal behaviours that were codified as rules. Roy, J. & Davenport, M. (2010) represented that expert knowledge as an ontology and developed a system to perform reasoning using Description Logics to classify vessels of interest and detect anomalies. A rule-based fuzzy expert system was illustrated by Jasinevicius, R. & Petrauskas, V. (2008) that considered the vessel type, persons on board and the riskiness of the cargo.

Rhodes et al. (2006) developed a prototype system to detect unsafe, illegal and threatening vessel activity that learns from operator-labelled track reports and their responses to automatically generated alerts. A modified Fuzzy ARTMAP neural network classifier is employed to learn models of vessel behaviour. Longitude, latitude, speed and course are input to the system and the classes employed are normal, anomalous and unknown.

Li et al. (2006) detected suspicious or anomalous behaviour by dividing the tracks into "motifs" (such as straight line, u-turn or loop) and examples of anomalous behaviour are used to train the classifier. Johansson, F. & Falkman, G. (2007) took a Bayesian network

approach where a model of normal vessel behaviour was learnt and anomalies were then detected by the system by comparing the data to the model.

In this work, a model of normality is created from historical AIS data using Gaussian Processes (GPs), thus codified expert knowledge is not required. An advantage with GPs is that the model is non parametric so it is not necessary to build in features of anomalous behaviour. A limitation of this approach is that although GPs provide a flexible and robust approach to anomaly detection, they are not typically suitable for large datasets due to high computational complexity in training ($O(n^3)$) and prediction ($O(n)$). In order to scale to the tens of thousands of ships broadcasting at any moment in time, Kd-Trees are used for efficient data representation in conjunction with an approximation heuristic to minimise computational complexity.

The contribution in this paper is a GP based model for normal behaviour combined with a Kd-Tree approximation for training and prediction. The speed and accuracy of the approximation is reported along with the results of anomaly detection.

2. Gaussian Process Regression

A Gaussian Process is a stochastic process which generates Normally distributed samples which have a multivariate Normal joint distribution. GPs have often been used for Bayesian non-parametric regression. This section will briefly describe the GP regression model; for a more detailed explanation see Rasmussen & Williams (2006).

2.1. *Gaussian Process Regression Model*

A standard regression model consists of a function, f, on a set of d covariates, X, such that $y = f(X) + \varepsilon$, where y represents the regression target variable and ε is independent normally distributed noise with zero-mean and variance σ^2. For a GP regression model $f(X)$ is a (zero-mean) Gaussian Process with covariance function $K(X, X')$. Predicting the value of new test point y_* given a training set $\mathcal{D} = \{X_i, y_i\}_{i=1}^N$ and an input vector X_* requires inferring the posterior distribution:

$$p(f_*|X_*, \mathcal{D}) \sim \mathcal{N}(\overline{f}_*, \sigma_*^2), \tag{2.1}$$

where the mean and variance are defined as:

$$\overline{f}_* = k_*^T M^{-1} y, \tag{2.2}$$

$$\sigma_*^2 = K(X_*, X_*) - k_*^T M^{-1} k_* \tag{2.3}$$

and $M = (K + \sigma^2 I)$ and $k_* = [K(X_*, X_1), ..., K(X_*, X_N)]^T$.

2.2. *Application*

In this paper the GP is applied to AIS data. The covariates X were chosen as the latitude and longitude from the raw (unfused) AIS data and the target vector consisted of the speed of the vessels at those positions. For the purposes of training the model, all the AIS data points are assumed to be normal.

3. Kd-Tree Gaussian Process

Kd-Trees are an efficient data structure for storing a finite set of points from a k-dimensional space Moore, A. W. (1991). A Kd-Tree is a binary tree that splits the widest dimension of the input space such that each child node contains the relevant half of the data. The data is split until the leaf nodes only contain one data point. This data

structure can be used in a GP to speed up the calculation of the predicted mean (Shen, Ng & Seeger (2005)). The GP prediction in (2.2) can be considered as a weighted sum:

$$\overline{f}_* = k_*^T M^{-1} y = k_*^T p = \sum_{i=1}^{n} w_i p_i, \tag{3.1}$$

where $w_i = K(X_*, X_i)$. It is this weighted sum which can be approximated by the tree to gain an increase in speed. Each node in the tree contains:

- The number of data points contained in the node (N_{ND})
- The unweighted sum of p over all leaf nodes contained in the branch

The speed increase is achieved by approximating the weighted sum. The weights are calculated from the covariance function of the GP. This function has the property that the further away two points are the smaller the value of the weight between them. This means that points which are further away contribute less to the weighted sum. The tree method takes advantage of this by searching the tree for points which are close to the query point but once the magnitude of the weights drop below a defined threshold (i.e. the distance between the points is large) the contribution from the remaining points is estimated as:

$$\sum_{i=1}^{n} w_i p_i = w \sum_{i=1}^{n} p_i, \tag{3.2}$$

where $w = \frac{1}{2}(w_{min} + w_{max})$ and w_{min} and w_{max} are the minimum and maximum weights contained within the node. The cut off rule suggested in Shen, Ng & Seeger (2005) is:

$$N_{ND}(w_{max} - w_{min}) \leq 2\epsilon(W_{SoFar} + N_{ND} w_{min}). \tag{3.3}$$

where W_{SoFar} is the "accumulated weight so far in the computation" and $W_{SoFar} + N_{ND} w_{min}$ "serves as a lower bound on the total sum of weights involved in the regression" (Shen, Ng & Seeger (2005)). The parameter ϵ controls the amount of the tree for which an exact weighted sum is calculated and can be altered to approximate more of the tree. This approximation provides an increase in speed without significantly impacting the performance of the GP as long as ϵ is not too large. This implementation is referred to as KDGP. A similar Kd-Tree approximation has been developed for the variance of the prediction. However due to space limitations this is not reported.

3.1. *Training Approximation*

It is known that GPs do not scale well when considering training. The training time has cubic complexity, ($O(n^3)$), where n is the number of training samples. Shen, Ng & Seeger (2005) proposes an additional method to increase the speed of training by using Conjugate Gradient (CG). This method speeds up the training time while avoiding the need to invert a large matrix.

The CG method searches for the solution to

$$p = M^{-1} y \tag{3.4}$$

as required in (3.1) by maximizing the quadratic function

$$q(z) = y^T z - \frac{1}{2} z^T M z \tag{3.5}$$

where z converges to p after n steps. Following the suggestion by Shen, Ng & Seeger (2005) an intermediate z was used in this work as a further approximation.

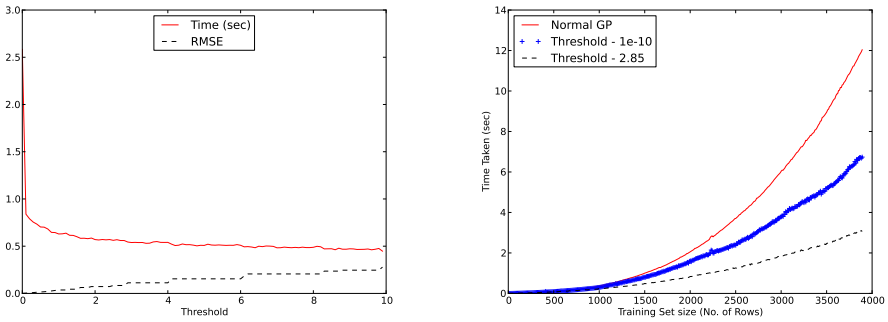

FIGURE 1. The left graph show the affect on the RMSE and time taken to estimate p as the break threshold increases. The right graph shows a comparison of training times against the number of training samples for a normal GP and the KDGP with two different thresholds in the calculation p

4. Experimental Results

Three main experiments were run during this investigation. First, a comparison between the GP and the KDGP training speeds was made. Second, the value of ϵ was varied to determine how the error and speed of the predictions were affected. Third, the ability to detect abnormal behaviour was investigated.

4.1. *Training speed*

To investigate the training approximation the threshold on the change in q for breaking the CG loop was varied and the affect on the time taken and the error in p was investigated. Figure 1(Left) shows how the time taken to calculate p varies as the threshold increases for a training set of 4000. It also shows the affect on the root mean squared error (RMSE) of the calculated p. For this problem it was decided that an RMSE of 0.1 was acceptable, providing a speed up of 4 times for the largest dataset. A comparison between the training times for the normal GP and the KDGP for two thresholds can be seen in Figure 1(Right) for different sample sizes.

4.2. *Error and Prediction Speed vs.* ϵ

As has been discussed in Section 3 the Kd-Tree provides an approximation to the mean prediction of a normal GP. The amount of approximation can be controlled by varying ϵ. How the value of ϵ affects the time for the prediction and the accuracy of the results was investigated to find the optimum value of ϵ. The KDGP was trained on 4000 data points and used for prediction on 200 test points. The 200 predictions were repeated with increasing values of ϵ. The RMSE of the predictions was calculated along with the mean time it took for one prediction and the results can be seen in Figure 2. It can be seen that the smaller the value of ϵ the more accurate the prediction but the longer the computation takes. From this graph the value of ϵ chosen was 0.001 as this was thought to give a reasonable trade off between speed and accuracy. It is noted that at $\epsilon = 0$ the error is not zero. This residual error is introduced through approximations during training.

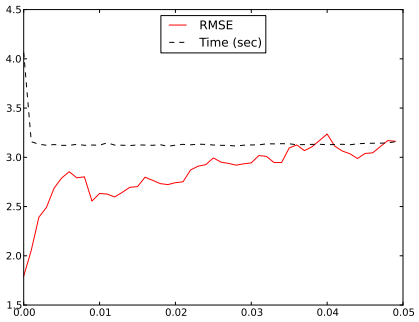

FIGURE 2. A plot of prediction the mean time and the root mean squared error of the predictions in knots versus the threshold

4.3. *Abnormality detection*

The difference between the GP model prediction and the new AIS reading are used to assign a measure of abnormality. Three metrics have been identified to generate this measure. The first two methods, squared residual (SR) and predictive log-likelihood (score), are defined in Loy, Xiang & Gong (2009) as

$$SR = (y_* - \overline{f}_*)^2 \tag{4.1}$$

$$score = \frac{1}{2} log(2\pi\sigma_*^2) + \frac{(y_* - \overline{f}_*)^2}{2\sigma_*^2} \tag{4.2}$$

The third is a simple threshold of the variance. If a test point lies beyond 3 standard deviations from the predicted mean it is considered to be abnormal. For this experiment both the KDGP and GP were trained on 4000 data points. The test points were made up of 200 points from the remaining data and 200 artificially created anomalous points. This was done as the data from AIS is assumed to contain no examples of anomalous behaviour. The anomalous data was created by taking 200 points from the AIS data that were not part of the training set or the set already chosen for testing. These points then had a small amount of Gaussian noise added as well as a constant offset, which may be characteristic of AIS spoofing. For each experiment the threshold for the metrics and the size of the constant offset added to the anomalous data was incremented and receiver operating characteristic (ROC) curves created for each metric. These curves are shown in Figure 3. The graphs show that the larger the offset added to the speed component the better the system determines whether the vessel is anomalous. The approach taken could be used to detect behavioural anomalies, such as vessel loitering (where the speed is too low) and sailing too fast e.g. in a port.

5. Conclusions and Further Work

This paper has developed a GP based model for normal AIS data and applied Kd-Tree approximation to speed up training and prediction. It has shown anomaly detection to be

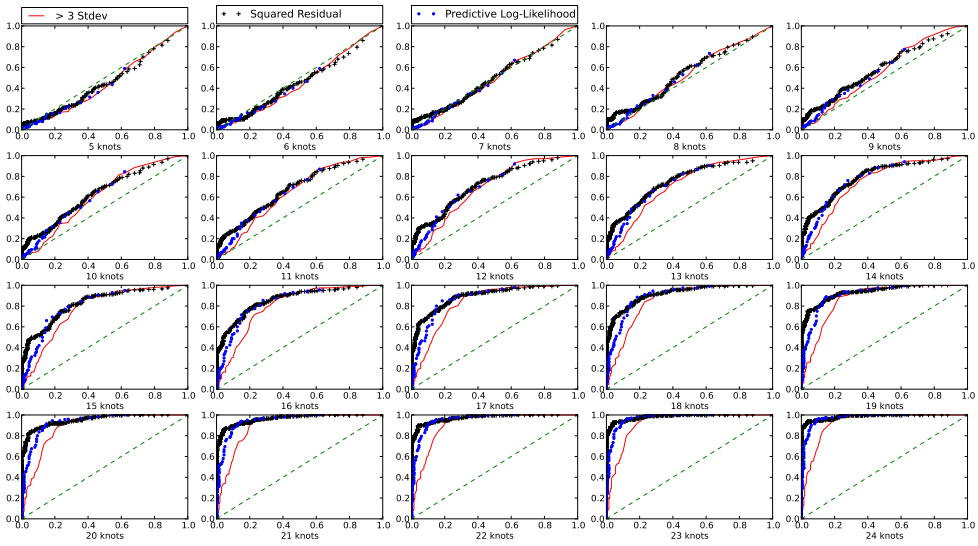

FIGURE 3. ROC curves showing the abnormality detection performance for all three metrics with growing amounts of abnormal data. The closer to the top left corner the curve is the better the performance

feasible on large datasets with a 4 times improvement in time taken with 4000 training samples. Further work is being undertaken to extend these techniques to produce a system to work online with a dataset covering the whole world.

REFERENCES

DAVENPORT, M. 2008 Maritime Anomaly Detection Workshop Report and Analysis *MacDonald Dettwiler and Associates Ltd., DRDC Scientific Authority, CR 2008-275*

JASINEVICIUS, R., PETRAUSKAS, V. 2008 Fuzzy expert maps for risk management systems *US/EU-Baltic International Symposium, 2008 IEEE/OES, Tallinn*

JOHANSSON, F., FALKMAN, G. 2007 Detection of vessel anomalies - a Bayesian network approach 2007 *3rd International Conference on Intelligent Sensors, Sensor Networks and Information*

LI, X., HAN, J., KIM, S. 2006 Motion-alert: automatic anomaly detection in massive moving objects *Proceedings of the 2006 IEEE Intelligence and Security Informatics Conference (ISI 2006), San Diego, CA*

LOY, C. C., XIANG, T., & GONG, S. 2009 Modelling Multi-object Activity by Gaussian Processes. *School of EECS, Queen Mary University of London*

MOORE, A. W. 1991 An introductory tutorial on kd-trees, Carnegie Mellon University, extract from A. W, Moore's PhD Thesis: Efficient Memory-based Learning for Robot Control. *Computer Laboratory, University of Cambridge.*

RASMUSSEN, C. E. & WILLIAMS, C. K. I. 2006 Gaussian Processes for Machine Learning. *The MIT Press*

RHODES, B.J., BOMBERGER, N.A., SEIBERT, M., WAXMAN, A.M. 2006 SeeCoast: Automated Port Scene Understanding Facilitated by Normalcy Learning *IEEE Military Communications Conference, 2006. MILCOM 2006, 23-25 Oct., pp1-7*

ROY, J., DAVENPORT, M. 2010 Exploitation of maritime domain ontologies for anomaly detection and threat analysis *2010 International Waterside Security Conference (WSS), pp.1-8*

ROY, J. 2008 Anomaly detection in the maritime domain *Proc. SPIE, Vol. 6945, Optics and Photonics in Global Homeland Security IV, Orlando, FL, USA*

SHEN, Y., NG, A. Y. & SEEGER M. 2005 Fast Gaussian Process Regression using KD-Trees. *In Advances in Neural Information Processing Systems* **18**