Data driven Intelligence for countering crime

Panagiotis Pentaliotis (p.pentaliotis@liverpool.ac.uk) , Supervised by Prof Simon Maskell, Prof Paul Spirakis(UOFL), Dr Matt Farrow(NCA)

EPSRC Centre for Doctoral Training in Distributed Algorithms, University of Liverpool, Liverpool, UK

Introduction

The National Crime Agency (NCA) are responsible for countering Organised crime in the UK. They are always looking for innovative ideas to battle the increasing crime, The NCA has decided to fund our university through EPSRC to produce such ideas. My main interests revolve around, more informed decision making and transparent Neural Networks for future prediction. Combining these two subjects will benefit the NCA by making informed decisions about resource management, while also have the ability to argue about those decisions. Shown in this poster are two of the projects I had the pleasure of working on during my PhD. The main feature that the NCA provides is a plethora of data. This is a double edged sword, as more data requires the use of parallel algorithms and stronger hardware to process. This is an other key aspect of real world scenarios that I get to work with.

Kalman Filter/Smoother EM

Currently we are focusing on the Simplified molecular-input line-entry system(SMILES) based molecule representation. As there are different encoding algorithms for SMILES string generation, a single molecule can have multiple forms of SMILES, and training with only canonical SMILES can lead to biased models [3].The Kalman filter, smoother EM is a Bayesian algorithm that can be used to create informed, predictions since it uses the Bayesian probabilities equation to derive estimations based on previous known knowledge. Moreover it has the ability to filter the observed values that are provided to extract the noise from unforeseen circumstances. The Expectation Maximisation algorithm (EM), is a well studied algorithm. It is used for locating optimal parameters with the use of the local maximum likelihood of statistical models whose equations can not be solved directly, and optimises the parameters for the KF

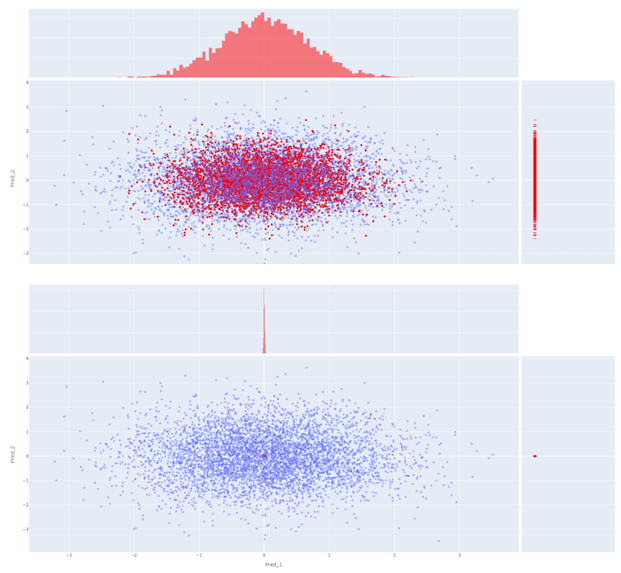

FIG 1: Prior and after EM converges results

Results

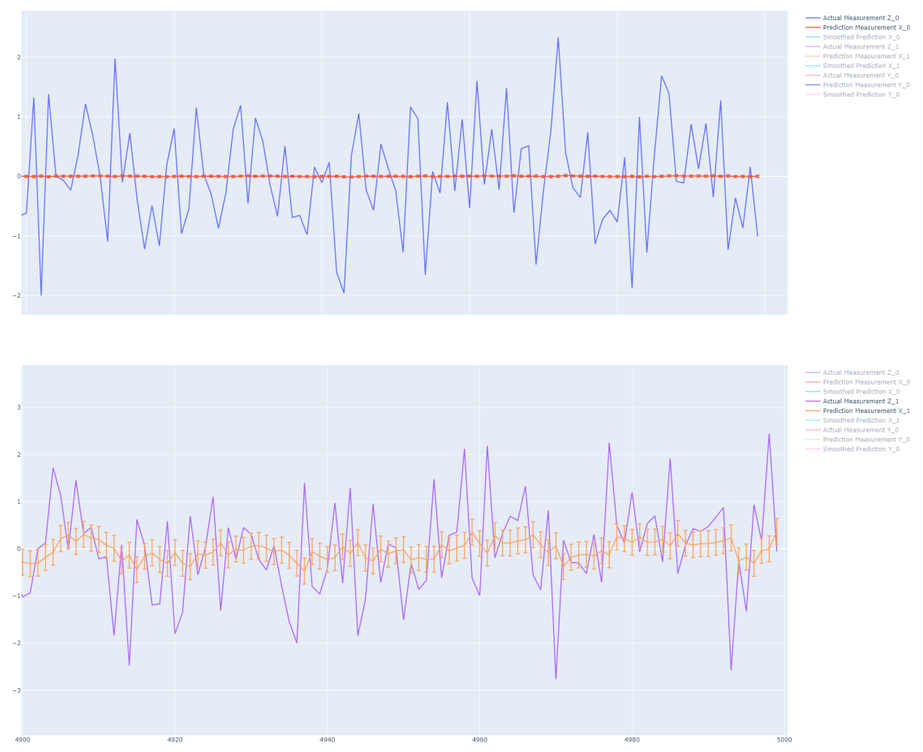

FIG 2: De-noised Measurements produced from the KF -EM, The second Graph shows the confidence of the measurements.

Tweeter Geolocation

The NCA are interested in the public opinion to keep the public safe and calm. Millions of tweets with unfiltered opinions of the public can be extracted from Tweeter every minute. Using the tweets one can monitor the public worry about crime happening all over the UK. The main issue after attempting to connect where people are when tweeting about crime is the lack of geolocation data. After extracting 100 thousand tweets only about 0.1% are geolocated, even attempting to extract tweets especially about crime the percentage is even less.

To counter this are research attempts to use the language used by each county to provide a localising belief vector of where a tweet came from. In order to increase the available data.

Each value of the vector is the ’belief’ value p(county|tweet). The probability of each county having a particular tweet that is calculated from. This allows us to observe the probability being from each of the Counties, we do this to capture outliers. For example, people that could use a strong accent from a county but have traveled and are talking about what they can see in a different county.

Results

Further enhancing the understanding of the classifications we used the Entropy measurement on each word. By collecting each p(word) for each of the counties for a word we can create a word vector that monitors the probability and importance of each word for each county. Using entropy we can distinguish between words that are helpful or not. A helpful word is a word with low entropy, meaning high information gain for the classifications

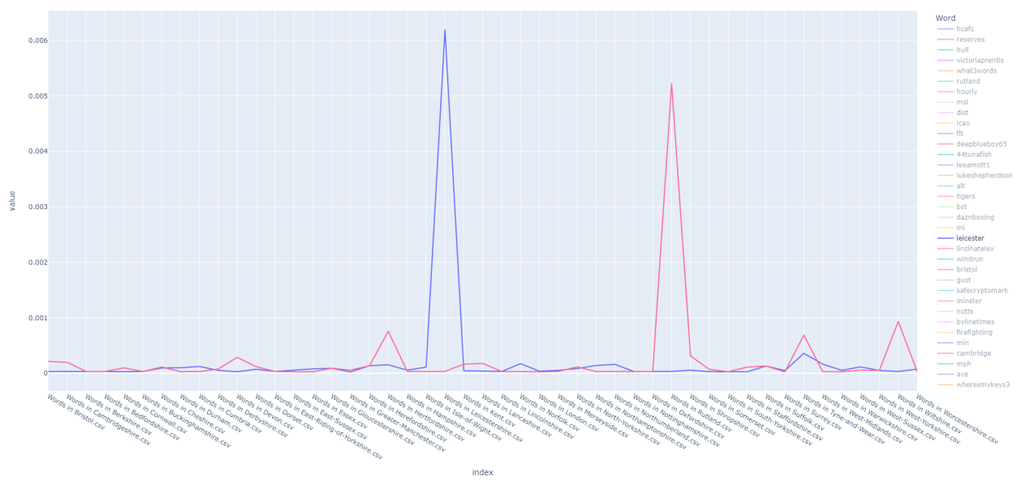

FIG 3:Entropy ranked location indicative works from out corpus