Learning an Optimal Policy for Police Resource Allocation on Freeways

Brian Jackson Taylor Howell Ola Shorinwa

Abstract

Anomaly detection is well studied and has applications in surveillance, driver assistance, and navigation. In this project, we develop an anomaly detection system using Principal Component Analysis and a Mixture of Gaussians. Using this system we identify feature distributions for anomalous drivers, from the NGSIM highway dataset, and then use these distributions in a simulator to learn an optimal policy for police allocation to apprehend anomalous vehicles. The allocation problem is formulated as a Markov decision process and solved using reinforcement learning. The optimal policy learns to allocate police only for drivers with higher citations.

1. Introduction

Each year over 32,000 driving-related deaths and 2 million injuries occur on US roads (Sauber-Schatz, 2016). We posit that automatic traffic surveillance and efficient allocation of police to apprehend anomalous drivers can improve road safety. Surveillance systems that automatically detect anomalous drivers can act as useful tools to help police more efficiently evaluate hundreds of drivers and efficiently focus on the ones that "stand out". In this project, we apply the EM algorithm to fit a mixture of Gaussians to a real highway dataset (NGSIM) and identify anomalous drivers from the resulting Gaussians. We use reinforcement learning (RL) in obtaining an optimal policy for police allocation.

For the CS229 project we focused on anomaly detection and reinforcement learning and for AA228 we focused on formulating the MDP.

2. Related Work

The problem of detecting anomalous driver is a well-studied problem given its valuable applications in surveillance, crisis prevention, driver assistance, and navigation. (Chandola et al., 2009) gives an overview of various approaches to anomalous behavior detection, which are categorized into the following classes: classification-based, parametric or non-parametric statistical, nearest neighbor-based, clustering-based, spectral techniques, and information theoretic. Many studies have taken a similar approach to the one used in the current project. Notably, Morris and Trivedi (Morris & Trivedi, 2008) propose an algorithm for extracting trajectory information from camera data, compressing the feature subspace using Principal Component Analysis (also used in the current project), and then classifying the data using a variety of unsupervised learning techniques including K-means, fuzzy-C-means, and neural networks. New trajectories are assigned a score rating on how anomalous they are, based upon the distance to the nearest cluster. Other approaches have included support-vector machines (SVMs) (Piciarelli et al., 2008), Hidden-Markov Models clustered using K-means (Suzuki et al., 2007), and semi-supervised learning (Sillito & Fisher, 2008). The field of anomaly detection is a broad and very active area of research, and presenting any sort of in-depth analysis of the various approaches is beyond the scope of this project. However, based on the cited papers, SVM techniques are advantageous given their computational efficiency, while the more traditional approach is advantageous given that it builds a distribution that models "typical" driver behavior. The semi-supervised learning approach showed promising results but requires human-in-the-loop training and is susceptible to undetected anomalies. Resource allocation is a well-studied field and the topic of police allocation has received some attention. However, applying state-of-the-art decision making algorithms to police allocation is limited (Adler et al., 2014), and typically based on optimizing deployment location rather than deciding whether or not to deploy limited resources. Our project therefore seems to be a novel application of decision-making processes to learning an optimal policy for deploying police in an uncertain environment with variable rewards.

3. Dataset and Features

The real-world driver data comes from the NGSIM US Highway 101 dataset. This data was captured using a set of cameras set atop a 36-story building and processed into vehicle trajectories that include:width, length, velocity, position, lane, etc. However, the original trajectory data contains significant noise in the position, velocity, and acceleration estimates. We used data post-processed by Stanford Intelligent Systems Laboratory, which improves state

estimates using a Kalman filter. The resulting data consists of 45 minutes of vehicle trajectories, including about 6000 unique vehicle trajectories. From this filtered data, we have constructed a training set consisting of features averaged over 15-second time intervals with 5 second spacing, resulting in about 64,000 unique driver examples. Each examples contains six features: average velocity, average acceleration, maximum velocity, number of lane changes, average deviation from lane centers, and standard deviation from lane centers. The velocity, acceleration, and deviation from lane centers were all given by the filtered dataset. The averages, maximums and standard deviation were taken over the 15-second time window of interest. The number of lane changes was determined by counting the number of times the lane deviation varied by more than a fixed threshold in a given time step (in either direction). The resulting features provide a depiction of driver behavior over an appropriate time window to identify anomalous behavior.

4. Methods

The project is divided into three principal components: 1) preprocessing: analyzing and compressing real-world highway-driving data, 2) anomaly detection: identifying anomalous drivers from the dataset, and 3) the MDP: learning an optimal policy for deploying police to cite anomalous drivers and maximize ticket revenue. We assume that it's possible to detect anomalous driving behavior that corresponds with traffic offenses on freeways and that it's possible to indirectly make the freeway safer by maximizing ticket revenue. We use the NGSIM data set, which tracks vehicle trajectories in congested morning traffic along a half mile section of Highway 101, to create a training set; model highway driving behavior as a mixture of Gaussians (fitted using the EM algorithm) and characterize atypical driving behavior as lying outside the modeled distributions; and solve the police allocation problem using model-based reinforcement learning.

5. Experiments/Results/Discussion

EM Algorithm

We assumed that each example came from a joint distribution $p(x, z) = p(x|z)p(z)$ consisting of four Gaussians (the number of latent distributions was selected by balancing the number of anomalous drivers we wanted to consider in the police allocation problem). Thus, each $x^{(i)}$ was obtained by randomly choosing $z^{(i)} \in \{1, 2, ..., k\}$ and then sampling $x^{(i)}$ from the corresponding Gaussian distribution. We intend to learn the parameters of the joint distribution $p(x, z; \theta)$ for the training examples where $z^{(i)}$ represents the latent variables and $\theta = \{\phi, \mu, \Sigma\}$. Finding a closed form solution for the maximum likelihood estimates of θ is difficult; hence, we used the EM algorithm for estimating θ. The EM algorithm works by constructing a lower bound on the log-likelihood of the training set $\ell(\theta)$ before optimizing the lower bound. The algorithm is given below.
E-step: For each i,

$$Q_i(z^{(i)}) \coloneqq p(z^{(i)}|x^{(i)}; \theta)$$

M-step: Set

$$\theta \coloneqq \underset{\theta}{\operatorname{argmax}} \sum_{i}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

We use the estimated parameters in fitting a mixture of Gaussians to the NGSIM data set and identify outliers using a threshold $p(x) < 0.0025$.

Principal Component Analysis

We used principal component analysis as a dimensionality reduction algorithm to map the extracted features from 6 features to different feature spaces ranging from sizes 2 to 5. The first k principal components of the data were selected to maximize the variance of the projections where the first principal component satisfies the optimization problem:

$$u \coloneqq \underset{u:u^T u=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^{m} ({x^{(i)}}^T u)^2$$

We apply the EM algorithm to each of the resulting feature spaces to obtain a model for classifying anomalous drivers. The features of the anomalous drivers flagged by each model was analyzed to evaluate the performance of each model.

Formulating the MDP

We have formulated the task of allocating police on freeways as a discounted infinite-horizon Markov decision process. The state is defined by two variables: the police state and the driver state. The police state is the vector $P = [p_1 \dots p_{n_p}] \in \mathbb{Z}^{n_p}$ of police car states $p \in \{0, 1, ..., p_{max}\}$. Each integer p_i represents the number of time steps until the police car i is again available for allocation. $p_i = 0$ indicates that police i is ready for allocation. After being allocated, $p_i = p_{max}$ and decrements by one each time step. The driver state is the vector $D = [d_1 \dots d_{n_d}] \in \mathbb{Z}^{n_d}$ of driver states $d \in \{1, ..., d_{max}\}$. The integer values d_j correspond to the state of the anomalous vehicle j, defined as by Table I.

d	Reward	Description	Conditions
1	50	No citation	Not 2 or 3
2	100	Speeder	Speed $>$ 17 m/s
3	150	Weaver	# lane changes $>$ 3
4	250	Both	Both 2 & 3

Table 1. Definition of anomalous vehicle states

The size of the state space is $(p_{max})^{n_p} \times (d_{max})^{n_d}$. For our final solution, we set $n_p = 1, n_d = 2,$, $p_{max} = 5$, and $d_{max} = 4$. ($n_p = 2, n_d = 4,$, $p_{max} = 5$, and $d_{max} = 4$).

The action space defines the allocation of police cars in P. The action $a = \{0, 1, \ldots, n_p\} \in \mathbb{Z}$ corresponds to the number of citations to deliver. It is assumed that the vehicles in D with the highest reward will be cited.

The rewards are deterministically assigned: each vehicle state p_j is assigned a particular reward according to the simulator which queries the anomaly detector, as shown in Table I. If $a > 0$, the resulting reward is the sum of the a highest rewards corresponding to states in D. If $a > 0$ and $p_i \neq 0, \forall p_i \in P$ (no police available), then a penalty of -100 was assigned.

Because the NGSIM data set only provides a finite amount of experience, we created a simulator that samples from distributions learned from the NGSIM data. The simulator maintains an internal "scene" of 10 anomalous vehicles with each specified by the features from section 3, in addition to a vehicle ID and time stamp. The time stamp is decremented at each time step and vehicles with negative time stamps are removed from the scene. The features are sampled from independent distributions. However, only n_d of the 10 vehicles are selected from each scene, according to highest reward. This is a fair assumption since it would clearly be sub-optimal to issue a citation to a vehicle known to yield less reward (given our previous assumption that the rewards are deterministic). At each time step, new vehicles are added to keep the total number of vehicles in each scene constant. New vehicles were always initialized with a time stamp of p_{max} (described above). The time stamp was included with the purpose of modeling the fact that drivers can only be observed for a fixed amount of time and that the police only have a finite amount of time to decide whether or not to apprehend the vehicle. After adding the new vehicles, they are assigned a state value as a deterministic function of their features (maximum velocity and number of lane changes), according to the conditions listed in Table I. The distribution of vehicle state values used in the simulation is shown in Figure **??**. It should be noted that the threshold for speeding (17 m/s, or 38 mph) is lower than the actual speed limit (65 mph). The average maximum speed of the drivers in the data set was found to be 11.176 m/s which was expected as the data was recorded during heavily congested morning traffic. Thus, the chosen speed limit reflects anomalous behavior among the population of drivers in the data set.At each time step, the simulator accepts a state/action tuple and returns the new state and reward. Based on the input state and action, it issues citations to the vehicles in D with the highest reward and removes them from the scene. It then updates the scene by decrementing the time stamps and adding new vehicles.

Solving the MDP

We use model-based reinforcement learning to find an optimal policy for the MDP. The agent gains experience by interacting with the simulator for $n = 100$ time steps before updating the transition $T(s'|s, a)$ and reward $R(s, a)$ models using maximum likelihood estimates. Asynchronous value iteration is performed on the Bellman equation (Eq. 5).

$$U^*(s) \coloneqq \max_a \left(R(s, a) + \gamma \sum_{s'} T(s'|s, a)U^*(s')\right)$$

An update to the value function was assumed to have converged when $||U_k - U_{k-1}||_\infty < 0.005$. The learning procedure was repeated until 10 consecutive updates of the the value function converged on the first iteration. The optimal policy $\pi^*(s)$ was obtained from the optimal value function.

$$\pi^*(s) \coloneqq \text{argmax}_a \left(R(s, a) + \gamma \sum_{s'} T(s'|s, a)U^*(s')\right)$$

To evaluate the optimal policy, we calculate the cumulative rewards per allocation over a 1000 step episode (and average the results over 5 episodes) for the optimal policy and similarly compare it to a randomly-send policy that randomly assigns police vehicles to apprehend offending drivers if police are available and an always-send policy that assigns the maximum number of available police vehicles at each time step.

6. Results

Anomaly Detection

New feature sets of sizes ranging from 2 to 4 were constructed from the highway data set using PCA. These feature sets were evaluated against the baseline model with six features. The anomaly detection model was created by fitting four Gaussians to each feature set. Anomalous drivers were identified using a threshold $p(x) < 0.0025$. Table 2 shows the percentage of anomalous drivers flagged by each model. The number of anomalous drivers decreases with the size of the feature space.The baseline model flagged about 30% of drivers as atypical while the two-feature set model identified a substantially lower percentage of atypical drivers, 0.25%. The three-feature and four-feature models classified about 4% and 6% of drivers as atypical respectively.

Size of feature space	Percentage of anomalous drivers (%)
2	0.249
3	3.648
4	5.356
5	29.514

Table 2. Percentage of Anomalous Drivers for each model

The features of the flagged drivers were analyzed to further evaluate the effects of dimensionality reduction using PCA. Figure 1 shows the histogram of the number of lane

changes of outliers identified by the two-feature and baseline models. The two-feature model flags drivers having a high number of lane changes with high probability compared to the baseline model. The drivers with zero lane changes which were classified as anomalous exceeded the acceptable limits on other features.

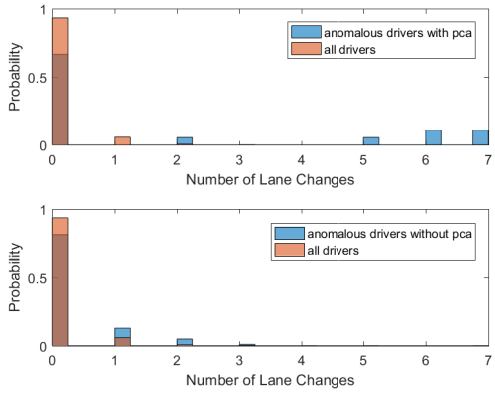

Figure 1. Histogram of number of lane changes of anomalous drivers for the two-feature and baseline models

Figure 2 shows the normalized histograms of the maximum velocity of the anomalous drivers identified by the two-feature and baseline models. The two-feature model performs remarkably well in identifying drivers moving at high velocities in addition to slow-moving drivers who might have been distracted. The baseline model seems more susceptible to noise, flagging a greater proportion of drivers with acceptable maximum velocities as anomalous.

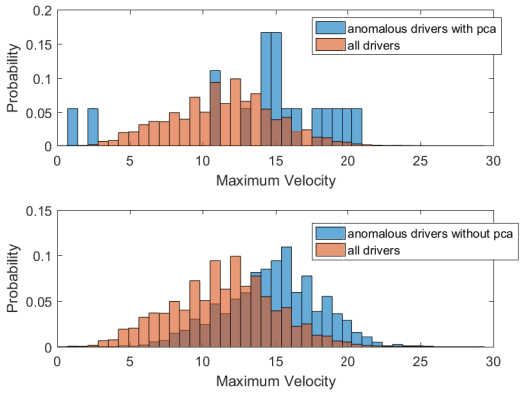

Figure 2. Histogram of the maximum velocity of anomalous drivers for the two-feature and baseline models

Figure 3 shows the normalized histograms of the average acceleration of anomalous drivers identified by the two-feature and baseline models. The two-feature model performs considerably well in identifying drivers with extreme accelerations, flagging such drivers as anomalous with a high probability. In contrast, drivers with acceptable accelerations have a higher probability of being flagged anomalous by the baseline model compared with drivers with extreme accelerations. The results show that the application of PCA helped reduce the susceptibility of the model to noisy observations, ultimately improving the accuracy of the anomaly detection algorithm. The PCA model retained the valuable variations among the feature attributes while discarding noisy variations which resulted in the model flagging a lower percentage of drivers as anomalous. The application of PCA also provided computational benefits by reducing the size of the feature sets. Such computational benefits scale with the size of the training set.

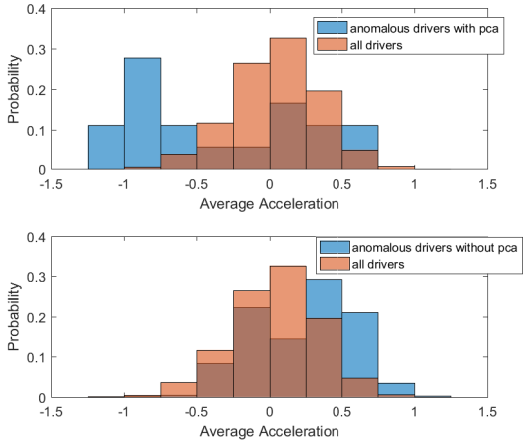

Figure 3. Histogram of average acceleration of anomalous drivers for the two-feature and baseline models

The simulator for the MDP generates new anomalous vehicles by sampling from independent distributions over each feature. The distributions over these features were created from the anomalous drivers identified by the two-feature model. The normalized histograms of the features are multi-modal; hence, the features could not be modeled as a unimodal Gaussian distribution. We estimated the distribution using a normal kernel distribution with different bandwidths. The bandwidth parameter controls the smoothness of the probability density curve. Figure 4 shows the fitted probability distributions over the features for different bandwidths. The default bandwidth parameter shown in the figure results from optimizing the mean integrated square error. A bandwidth parameter of one provided better estimations of the probability density curves of the number of lane changes, average velocity, and maximum velocity, showing the major peaks in the probability density curve for these features. The kernel distribution using the theoretically optimal bandwidth parameter provided a better estimate of the probability density curve of the average acceleration and was used by the simulator.

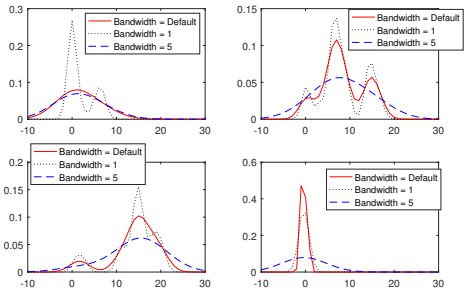

Figure 4. Kernel distribution over the (a) number of lane changes, (b) average velocity, and (c) maximum velocity and (d) average acceleration for different bandwidths

Optimal Policy

For the simple MDP with one police and two anomalous drivers, solving for the optimal policy took 2135.18 seconds, converging after 21,000 epochs (100 steps of experience per epoch). An ϵ-greedy strategy was used with ϵ = 0.1 and a discount factor γ = 0.9. For any state with no police to allocate (i.e., $\sum_{i=1}^{n_p} p_i = 0$) the police learned to do nothing (i.e., action = 0) which is optimal since taking any other action in this state will result in a negative reward. To better understand the policy, we examine the policy at critical states where police are available for allocation (i.e., p_i = 0). For these states, we find that there is some threshold such that drivers with low reward values are not cited. We found that drivers with $d_i < 3$ were not cited. This demonstrates that our agent was able to learn an intelligent policy that enables it to delay immediate reward for greater future reward. We compare the optimal policy with a randomly-send policy (if $\sum_{i=1}^{n_p} p_i > 0$ then a random number of available police are allocated, else no police are allocated) and always-send policy (number of police allocated = $\sum_{i=1}^{n_p}$. Because we are interested in efficiently allocating resources, we compare the average cumulative reward per allocation for each policy for a 1000 step episode for 5 trials and then average the results. The results of the learned optimal policy for the during training versus the "expert" policies are shown in Figure 5. The actual learned policy (and value function $U*$) for key states is shown in Table 3 against the expert policies.

State			Value	Policy		
d_1	d_2	p_1	U*	Optimal	Always	Random
1	1	0	333.209	0	1	0
2	1	0	341.488	0	1	0
3	1	0	372.812	1	1	1
4	1	0	454.712	1	1	1
2	2	0	341.113	0	1	0
3	2	0	377.063	1	1	1
4	2	0	475.914	1	1	0
3	3	0	377.997	1	1	1
4	3	0	478.403	1	1	0
4	4	0	496.187	1	1	1

Table 3. Policies for critical states

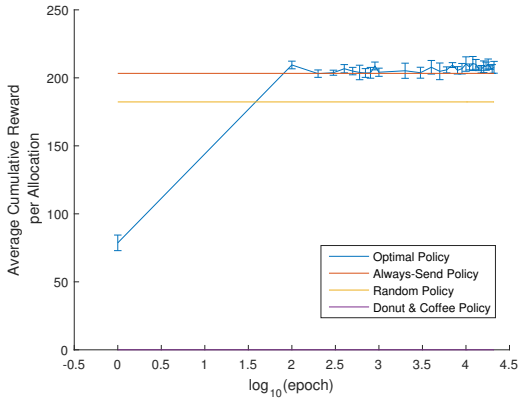

Figure 5. Learning Curve

7. Conclusion/Future Work

Driving behavior on freeways can be modeled by fitting a mixture of Gaussians to the trajectory data of drivers. Anomalous drivers can then be identified using the resulting model with a given threshold $p(x) < \epsilon$, indicating the probability of a driver's behavior lying outside the acceptable range. We applied PCA as a dimension-reduction algorithm before fitting the model and our results indicate that the application of PCA improved the accuracy of the algorithm. PCA also reduced the number of flagged drivers to more realistic proportions between 0.2% and 0.5% suggesting that this technique improved resilience to sensor and driver noise. We also learned an optimal policy for police allocation which does not allocate police for anomalous drivers with low citation rewards, preferring to reserve police for the future when drivers with higher citations may be identified. Consequently, our agent was able to maximize the citation rewards with a minimal number of police allocations.

The features of anomalous drivers are not independent as assumed in our generation of new anomalous drivers. To address this, we intend to train a generative adversarial network to generate new anomalous drivers for the MDP simulator. We also expect the optimal policy (reserving police vehicles for drivers with higher citations) to scale to a larger state space with more police vehicles and anomalous drivers and plan to test this hypothesis.

8. Contributions

Jackson designed and coded the simulator. Howell coded the EM algorithm and MDP solver. Shorinwa coded the PCA algorithm and analyzed the anomaly detection stack.

References

Adler, Nicole, Hakkert, Alfred Shalom, Kornbluth, Jonathan, and Sher, Mali. Location-allocation models for traffic police patrol vehicles on an interurban network. *Springer*, pp. 9–31, 2014.

Chandola, Varun, Banerjee, Arindam, and Kumar, Vipin. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

Morris, Brendan Tran and Trivedi, Mohan Manubhai. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE transactions on circuits and systems for video technology*, 18(8):1114–1127, 2008.

Piciarelli, Claudio, Micheloni, Christian, and Foresti, Gian Luca. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for video Technology*, 18(11):1544–1554, 2008.

Sauber-Schatz, Erin K. Vital signs: motor vehicle injury preventionunited states and 19 comparison countries. *MMWR. Morbidity and mortality weekly report*, 65, 2016.

Sillito, Rowland R and Fisher, Robert B. Semi-supervised learning for anomalous trajectory detection. In *BMVC*, volume 27, pp. 1025–1044, 2008.

Suzuki, Naohiko, Hirasawa, Kosuke, Tanaka, Kenichi, Kobayashi, Yoshinori, Sato, Yoichi, and Fujino, Yozo. Learning motion patterns and anomaly detection by human trajectory analysis. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pp. 498–503. IEEE, 2007.