RAPPORT DE PROJET

Détection d'anomalies routières à partir d'échanges de données entre systèmes de transports coopératifs

Réalisé par : Antoine BARBET

Tuteur universitaire : Stéphane CORMIER

Projet TER de Master 2 Informatique – du 29/01/2021 au 16/04/2021

https://github.com/Antoine553/projet-master2

Remerciements

Les travaux présentés dans ce document ont été effectués au sein de l'Université de Reims Champagne-Ardenne dans le cadre de la formation Master 2 informatique.

Tout d'abord, je tiens à remercier mon encadrant Mr Stéphane CORMIER et Mlle Juliet CHEBET pour le suivi attentif qu'ils ont apporté au projet ainsi que pour leurs disponibilités et leurs conseils.

Je tiens également à adresser ma profonde gratitude à toute l'équipe pédagogique de l'Université de Reims Champagne-Ardenne pour les connaissances qu'ils m'ont transmises et qui m'ont permis de mener à bien ce projet.

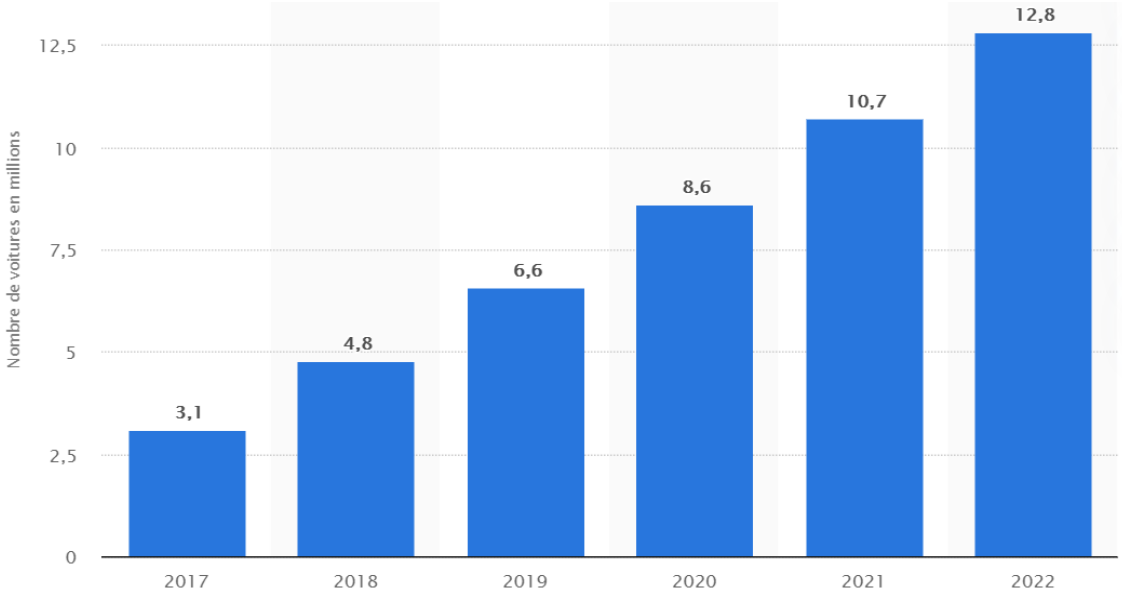
J'adresse mes remerciements à tous ceux qui m'ont aidé lors de la réalisation de ce projet et la rédaction de ce rapport.

Tables des matières

Contexte du projet

Les systèmes de transports intelligents coopératifs sont définis comme des véhicules capables de partager des informations de trafic avec d'autres véhicules et avec des structures. Le but de ces échanges est d'augmenter l'efficacité des déplacements routiers en permettant aux véhicules d'être prévenus à l'avance de tout incident ou événement de la circulation susceptible d'influencer leurs conduites. De cette manière, les structures peuvent s'adapter afin de fluidifier le trafic ou de répondre à des incidents, les véhicules autonomes peuvent adapter leurs conduites et les véhicules connectés peuvent apporter de meilleures informations à leurs usagers.

L'essor des technologies de communication sans fil et des intelligences artificielles a grandement contribué à rendre possible le déploiement de ces véhicules. En grand nombre, leurs observations peuvent rendre un aperçu précis du réseau routier sur lequel ils circulent en temps réel. Cela peut donc nous permettre d'optimiser la circulation mais aussi de détecter des incidents rapidement, de reconstituer des évènements particuliers ou de faire des modèles de prédiction très précis.



Chaque année, de plus en plus de voitures autonomes et connectés sont mises en service chaque année. Devant cette croissance exponentielle, la commission européenne a mis en place des plateformes de développement et des normes pour ces véhicules connectés dès 2010. Mais c'est en 2016 avec le lancement la plateforme C-Roads que l'Union européenne s'est mise à encourager activement une arrivée coordonnée de ces systèmes. Aujourd'hui, plusieurs plateformes relient les activités de déploiement de ces véhicules et vérifient l'interopérabilité de tous les systèmes déployé au sein de l'UE.

Introduction

La détection d'anomalies dans notre cas peut avoir deux utilités. Tout d'abord, un algorithme de détection d'anomalies peut nous permettre de nettoyer les jeux de données des observations incohérentes dues à des problèmes de prélèvement ou de transmission afin de permettre une utilisation finale plus précise de ces données. Ensuite, cette détection peut être utilisée afin de chercher à détecter et signaler des anomalies routières. Ces anomalies peuvent prendre de nombreuses formes mais on peut s'attendre à pouvoir les détecter en observant des changements soudains dans le comportement de plusieurs conducteurs aux mêmes emplacements.

Les données d'observation récupérées quotidiennement auprès des véhicules connectés représentent une quantité massive de données. Leur exploitation est donc très intéressante mais aussi limitée. En effet, si un véhicule génère plusieurs observations par seconde, Il est impensable d'analyser les données des véhicules sur tout le territoire français avec un seul modèle. L'origine de ces données pose aussi un problème de fiabilité. Les données sont relayées par un grand nombre d'agents et l'apparition d'anomalies lors de ces séries de relais est possible. De plus, les capteurs étant positionnés sur chaque véhicule, une anomalie peut donc être due à des anomalies routières (accidents, embouteillages, etc…) mais aussi à des capteurs défaillants ou à une manipulation anormale du véhicule ou des capteurs. On peut très bien imaginer qu'une voiture réalisant un « Dyno test » en garage génère des observations de véhicule roulant à grande vitesse sans changer de position. Une personne pourrait avoir endommagé ou déréglé un capteur par inadvertance lors d'une intervention mécanique sur son véhicule.

Ce projet porte donc sur la réalisation d'un système de détection d'anomalie dans des jeux de données transmis par des véhicules connectés. Notre système se concentrera sur la détection d'anomalies routières uniquement. Dans un premier temps nous analyserons les modèles des jeux de données transmis par ces véhicules et certains exemples d'anomalies. Dans un second temps nous étudierons les axes d'analyse possible, les méthodes de détections les plus adaptées et les choix d'algorithmes disponibles. Enfin nous simulerons l'arrivée en temps réel de nos données et étudierons la viabilité de l'utilisation de ces techniques en termes de performances de calcul et de détection.

1 - Détection d'anomalies

Le but de la détection d'anomalies est de repérer des données qui ne sont pas conformes à ce à quoi l'on peut s'attendre. Ces anomalies sont donc des observations qui dévient considérablement du reste des autres observations. Ces procédés sont utilisés dans le prétraitement de données pour en supprimer celles qui sont anormales. Cela entraîne une augmentation significative de la précision d'un algorithme d'apprentissage. Ils sont aussi utilisés dans de nombreux domaines afin de détecter des données manipulées ou ayant subi une interférence externe. Par exemple : la détection de fraude bancaire, la surveillance de l'état d'un système, la détection de cyber-attaque ou encore la détection d'anomalies routières.

Une détection d'anomalies doit donc classifier ce qui est une observation normale ou anormale. Cette tâche est particulièrement compliquée dans notre cas car la définition d'un comportement anormal n'est pas établie. Des anomalies d'origine routière peuvent prendre beaucoup de formes différentes. De plus, la même situation anormale peut entraîner une multitude de comportements différents selon chaque conducteur. Par exemple, dans le scénario d'une voiture arrivant à grande vitesse au niveau d'un trou dans la chaussée repéré tardivement, certains conducteurs vont brusquement ralentir, tandis que d'autres chercheront à l'éviter. Mais d'autres encore se résigneront à encaisser la secousse sans marquer le moindre changement dans leurs comportements.

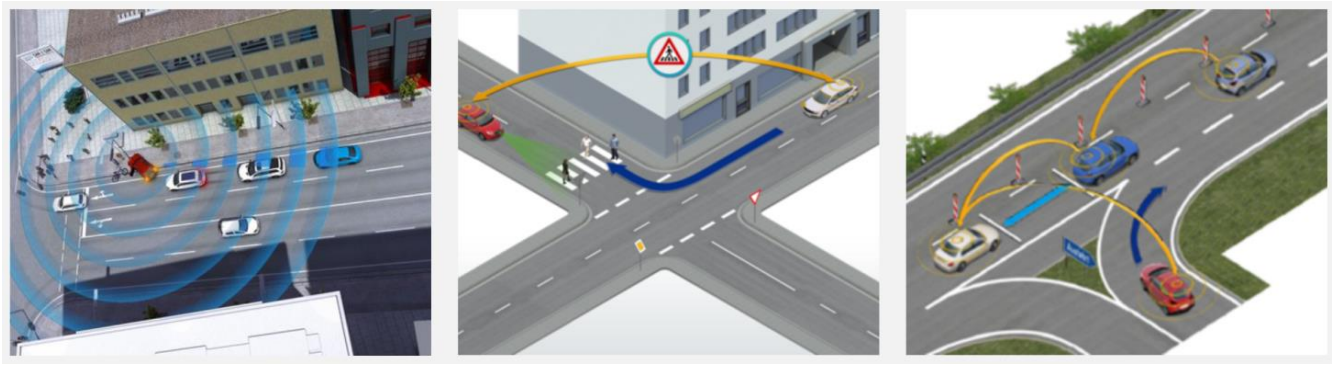
2 - Données transmises par les véhicules

Les véhicules connectés ne possèdent pas tous les mêmes degrés évaluation. Il est possible de les classer de manière simple en trois catégories, selon leurs capacités à transmettre des données plus ou moins complexe :

1) La majorité des véhicules connectés ne font que des observations basiques sur leur propre état tel que la vitesse, la position GPS et la direction.

2) Les véhicules les plus récents poussent leurs observations plus loin : à l'analyse des fonctions basiques, ils ajoutent celle de leurs environnements. Ainsi, à l'aide de capteurs supplémentaires, ils peuvent détecter la présence d'obstacles, de signalisations, de véhicules effectuant des manœuvres, etc. Cela veut dire que ce type de véhicules analysent eux-mêmes la présence potentielle d'anomalies sur la route sans forcément être impacté par celle-ci.

3) Enfin, les véhicules autonomes peuvent aussi fournir leurs intentions de déplacement et d'interactions avec les autres véhicules.



Dans notre cas, nous nous concentrerons sur des observations simples faites sur des véhicules avec des identifiants uniques et l'heure de l'observation. Ces observations concernent la vitesse, la position GPS et la direction des véhicules. Ce type d'observations est généré par les trois catégories de véhicules vus précédemment. Notre système de détection aura donc l'avantage de pouvoir utiliser les données transmises par un maximum de véhicules connectés.

Choix de notre méthode de détection

Il existe trois méthodes de détection d'anomalie. Les méthodes de détections d'anomalies supervisées, les méthodes de détections d'anomalies non-supervisées et les méthodes de détections d'anomalies semi-supervisées.

Il est possible d'utiliser une méthode supervisée à partir du moment où nous possédons un jeu de données étiquetées pour entraîner notre machine en lui fournissant des exemples qu'elle va apprendre à reconnaître. Elle va ainsi ajuster ses paramètres de détection au fil des entraînements afin de réduire sa marge d'erreur et d'acquérir peu à peu la capacité de reconnaître et de classer de nouveaux cas d'anomalies similaires aux cas d'anomalies d'entraînement. Dans notre cas, nous possédons bien des jeux de données étiquetées. En revanche, ces jeux de données ont été créés et étiquetés dans des cadres bien définis. Comme vus précédemment, le nombre de réactions et comportements anormaux possible est très grand. Reproduire et annoter suffisamment de ces instances dans notre jeu de données d'entrainement pour estimer notre modèle comme complet est impossible. Notre modèle ne serait pas capable d'évoluer et de classer correctement de nouveaux cas qui n'auront pas été étiquetés dans ce même jeu. D'autres facteurs tels que la météo, le trafic et l'heure de la journée impactent également le comportement des conducteurs. Les standards de bonne conduite varient selon les régions où les données sont générées. Nos données sont bien trop diverses, imprévisible et contextuelles pour avoir recours à un algorithme supervisé.

Dans une méthode non-supervisée, l'apprentissage par la machine se fait de manière totalement autonome. Les données sont fournies à la machine sans qu'elle soit étiquetée et celle-ci va chercher à regrouper ces données par similitude. Dans ce cas, la machine peut faire face à l'apparition d'un nouveau cas d'anomalie ou aux comportements très varié des conducteurs rencontrant des anomalies. Les jeux de données étiquetées disponibles peuvent servir à évaluer notre machine.

La méthode semi-supervisée est une approche qui consiste à entraîner une machine en combinant une petite quantité de données étiquetées avec une grande quantité de données non étiquetées. Cette méthode permet d'utiliser des données non étiquetées pour compléter un apprentissage non-supervisée. En revanche, dans ce cas, comme pour les algorithmes supervisés, le problème de diversité des anomalies est toujours présent.

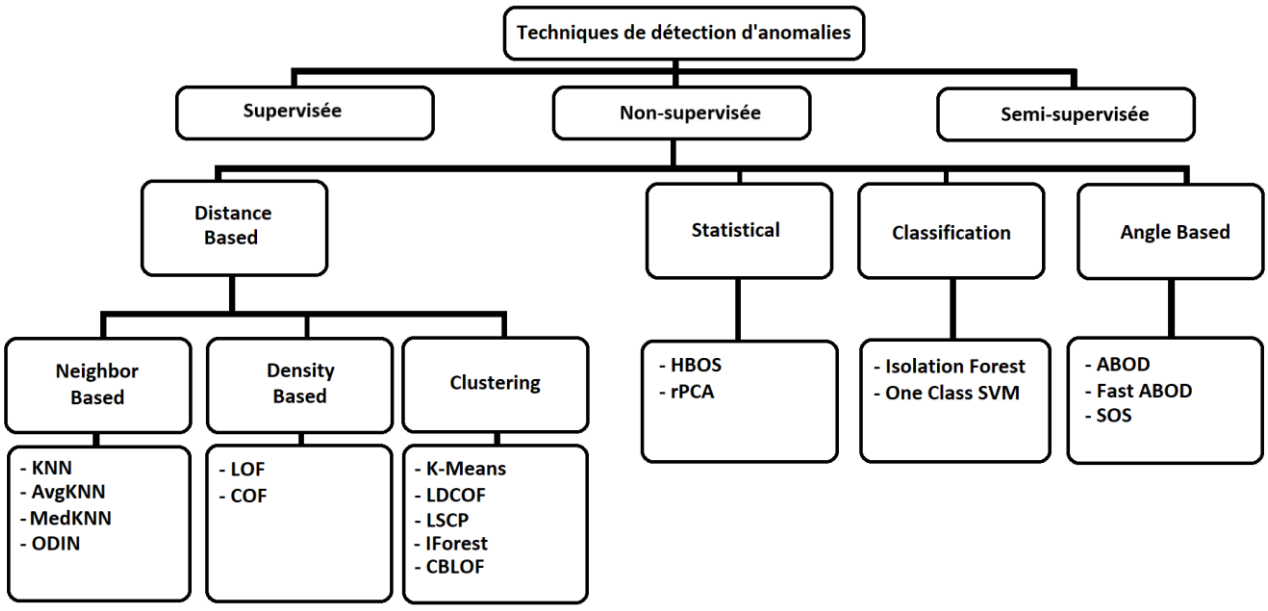
Afin de pouvoir prendre en compte la diversité et les évolutions possibles des données, nous avons donc choisi d'avoir recours à un algorithme non supervisé.

Choix de notre algorithme

Notre choix s'est porté sur un système de détection basé sur une méthode non-supervisé. Mais, il existe beaucoup d'algorithmes de détection d'anomalies non-supervisés et il faut en choisir un adapté à nos contraintes. Le paquet « PYOD » m'offre à lui seul les choix d'algorithme suivant :

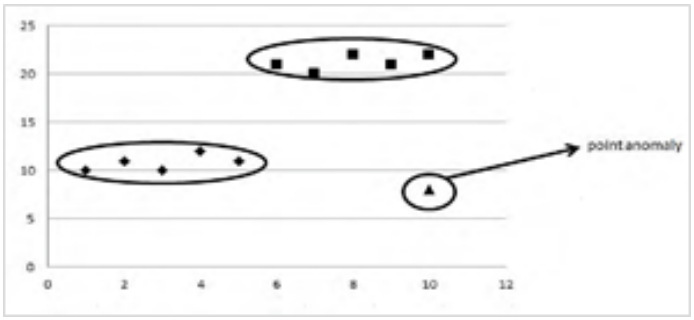
Type	Abbr	Algorithm
Linear Model	PCA	Principal Component Analysis (the sum of weighted projected distances to the eigenvector hyperplanes)
Linear Model	MCD	Minimum Covariance Determinant (use the mahalanobis distances as the outlier scores)
Linear Model	OCSVM	One-Class Support Vector Machines
Proximity-Based	LOF	Local Outlier Factor
Proximity-Based	COF	Connectivity-Based Outlier Factor
Proximity-Based	CBLOF	Clustering-Based Local Outlier Factor
Proximity-Based	LOCI	LOCI: Fast outlier detection using the local correlation integral
Proximity-Based	HBOS	Histogram-based Outlier Score
Proximity-Based	kNN	k Nearest Neighbors (use the distance to the kth nearest neighbor as the outlier score
Proximity-Based	AvgKNN	Average kNN (use the average distance to k nearest neighbors as the outlier score)
Proximity-Based	MedKNN	Median kNN (use the median distance to k nearest neighbors as the outlier score)
Proximity-Based	SOD	Subspace Outlier Detection
Probabilistic	ABOD	Angle-Based Outlier Detection
Probabilistic	FastABOD	Fast Angle-Based Outlier Detection using approximation
Probabilistic	SOS	Stochastic Outlier Selection
Outlier Ensembles	IForest	Isolation Forest
Outlier Ensembles		Feature Bagging
Outlier Ensembles	LSCP	LSCP: Locally Selective Combination of Parallel Outlier Ensembles
Outlier Ensembles	XGBOD	Extreme Boosting Based Outlier Detection **(Supervised)**
Neural Networks	AutoEncoder	Fully connected AutoEncoder (use reconstruction error as the outlier score)
Neural Networks	SO_GAAL	Single-Objective Generative Adversarial Active Learning
Neural Networks	MO_GAAL	Multiple-Objective Generative Adversarial Active Learning
Outlier Ensembles		Feature Bagging
Outlier Ensembles	LSCP	LSCP: Locally Selective Combination of Parallel Outlier Ensembles
Combination	Average	Simple combination by averaging the scores
Combination	Weighted Average	Simple combination by averaging the scores with detector weights
Combination	Maximization	Simple combination by taking the maximum scores
Combination	AOM	Average of Maximum
Combination	MOA	Maximum of Average

Ce choix doit être fait de manière à prendre en compte le processus voulu, nos contraintes techniques et les résultats attendus.

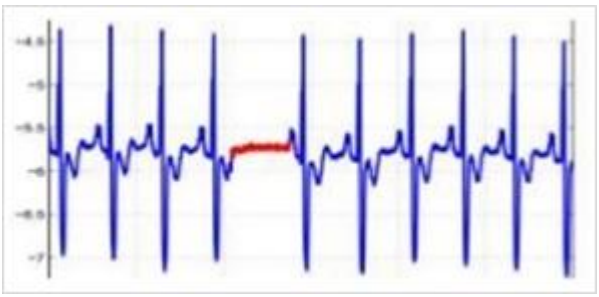


Afin de choisir notre type d'algorithme. Le type d'anomalie est important à définir. Les anomalies d'un jeu de données peuvent prendre plusieurs formes :

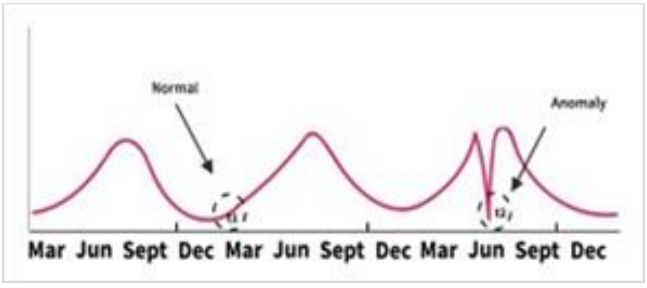
1) Des instances d'observation anormales uniques dans un grand jeu de données :



2) Des instances d'observations anormales groupées :



3) Des instances d'observations anormales uniques qui peuvent sembler normales mais sont anormales dans un contexte donné :



Il est donc nécessaire de réaliser une analyse de nos jeux de données et de certains cas d'anomalies. Ainsi nous pourrons déterminer quelles formes prennent nos anomalies et quel type d'algorithme serait adapté à nos jeux de données.

Jeux de données

Nous possédons plusieurs jeux de données à notre disposition. Ces jeux de données ont des structures similaires, possèdent des anomalies et sont étiquetés. Nos jeux de données comportent donc des observations décrites par les variables : Temps (en seconde), Identifiant, Latitude, Longitude, Vitesse (kms/h), Direction (sur un angle de 0 à 360 degrés).

1 - Analyse statistique

Les intervalles de temps espaçant les observations sont irréguliers et détecter une anomalie à partir de ces variables n'est pas évident. En revanche, Il nous est possible d'enrichir ce jeu de données en rajoutant des variables calculant les variations entre une observation et celle qui l'a précédé. Ainsi, nous pouvons étudier l'ampleur d'un changement de vitesse, de direction ou de position.

Voici un extrait de notre jeu de données après l'ajout des variables de comparaisons :

	ID	Time	CarId	Longitude	Latitude	Speed	Heading	Time diff	Position diff	Speed diff	Heading diff	Class
0	0	594.182212	118457	49.261743	4.056850	8.43	264.4	0.0	0.00000	0.0	0.0	0
1	1	594.282212	118457	49.261740	4.056839	8.37	260.5	0.1	0.00014	0.6	39.0	0
2	2	594.382212	118457	49.261737	4.056829	8.37	256.6	0.1	0.00013	0.0	39.0	0
3	3	594.482212	118457	49.261733	4.056820	8.36	250.6	0.1	0.00013	0.1	60.0	0
4	4	594.582212	118457	49.261727	4.056813	8.44	241.8	0.1	0.00013	0.8	88.0	0

Voici une analyse statistique de notre jeu de données :

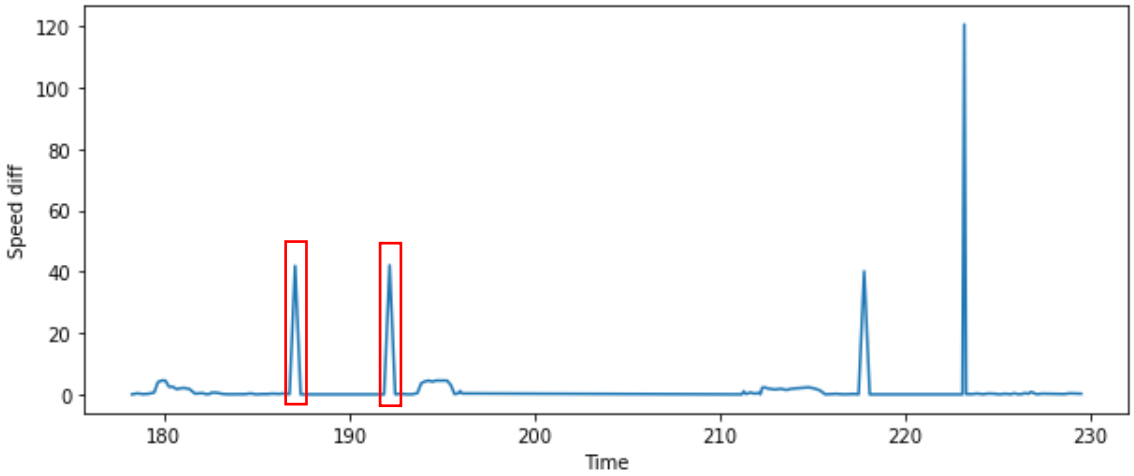
	ID	Time	Longitude	Latitude	Speed	Heading	Time diff	Position diff	Speed diff	Heading diff	Class
count	6341.00000	6341.000000	6341.000000	6341.000000	6341.000000	6341.000000	6341.000000	6341.000000	6341.000000	6341.000000	6341.000000
mean	3170.00000	428.525575	49.260828	4.056626	9.786789	127.626652	0.250654	0.000130	1.648751	12.579527	0.018451
std	1830.63336	261.141013	0.000793	0.000153	4.959000	94.330976	0.120390	0.000047	8.614061	46.861506	0.134587
min	0.00000	30.960276	49.259629	4.056383	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1585.00000	177.400747	49.260098	4.056486	5.280000	6.200000	0.200000	0.000113	0.000000	0.000000	0.000000
50%	3170.00000	457.684540	49.260696	4.056600	12.420000	175.200000	0.300000	0.000140	0.166667	0.000000	0.000000
75%	4755.00000	666.240830	49.261591	4.056769	13.930000	186.200000	0.300000	0.000153	0.800000	10.500000	0.000000
max	6340.00000	864.364589	49.262306	4.056999	16.220000	359.900000	1.000000	0.000500	139.700000	1752.000000	1.000000

2 - Analyse des anomalies

En observant notre jeu de données et des anomalies, nous pouvons nous faire une idée de ce qui caractérise une anomalie.

232	186.4627931	195061	49.260638	4.056615	13.73	8.1	0
233	186.7627931	195061	49.260675	4.056624	13.75	8.1	0
234	187.0627931	195061	49.260711	4.056633	1.18	0.2	1
235	187.3627931	195061	49.260748	4.056641	1.18	0.2	0
236	187.6627931	195061	49.260785	4.05665	1.18	0.2	0
237	187.9627931	195061	49.260822	4.056659	1.18	0.2	0
238	188.2627931	195061	49.260858	4.056668	1.18	0.2	0
239	188.5627931	195061	49.260895	4.056676	1.18	0.2	0
240	188.8627931	195061	49.260932	4.056685	1.18	0.2	0
241	189.1627931	195061	49.260968	4.056693	1.18	0.2	0
242	189.4627931	195061	49.261005	4.0567	1.18	0.2	0
243	189.7627931	195061	49.261042	4.056707	1.18	0.2	0
244	190.0627931	195061	49.261079	4.056714	1.18	0.2	0
245	190.3627931	195061	49.261116	4.056721	1.18	0.2	0
246	190.6627931	195061	49.261153	4.056728	1.18	0.2	0
247	190.9627931	195061	49.26119	4.056735	1.18	0.2	0
248	191.2627931	195061	49.261227	4.056741	1.18	0.2	0
249	191.5627931	195061	49.261264	4.056748	1.18	0.2	0
250	191.8627931	195061	49.261301	4.056755	1.18	0.2	0
251	192.1627931	195061	49.261338	4.056762	13.81	6.2	1
252	192.4627931	195061	49.261375	4.056769	13.8	6.2	0

Dans l'exemple ci-dessus nous pouvons voir une anomalie définie à partir de plusieurs observations caractérisées par une variation importante de la vitesse et une légère variation de l'orientation du véhicule jusqu'à un retour proche aux valeurs initiales. Ce type d'anomalies est représentatif de la majorité des cas d'anomalies de notre premier jeu de données. Notre variable de variation de vitesse nous permet de faire ressortir très distinctement cette anomalie.

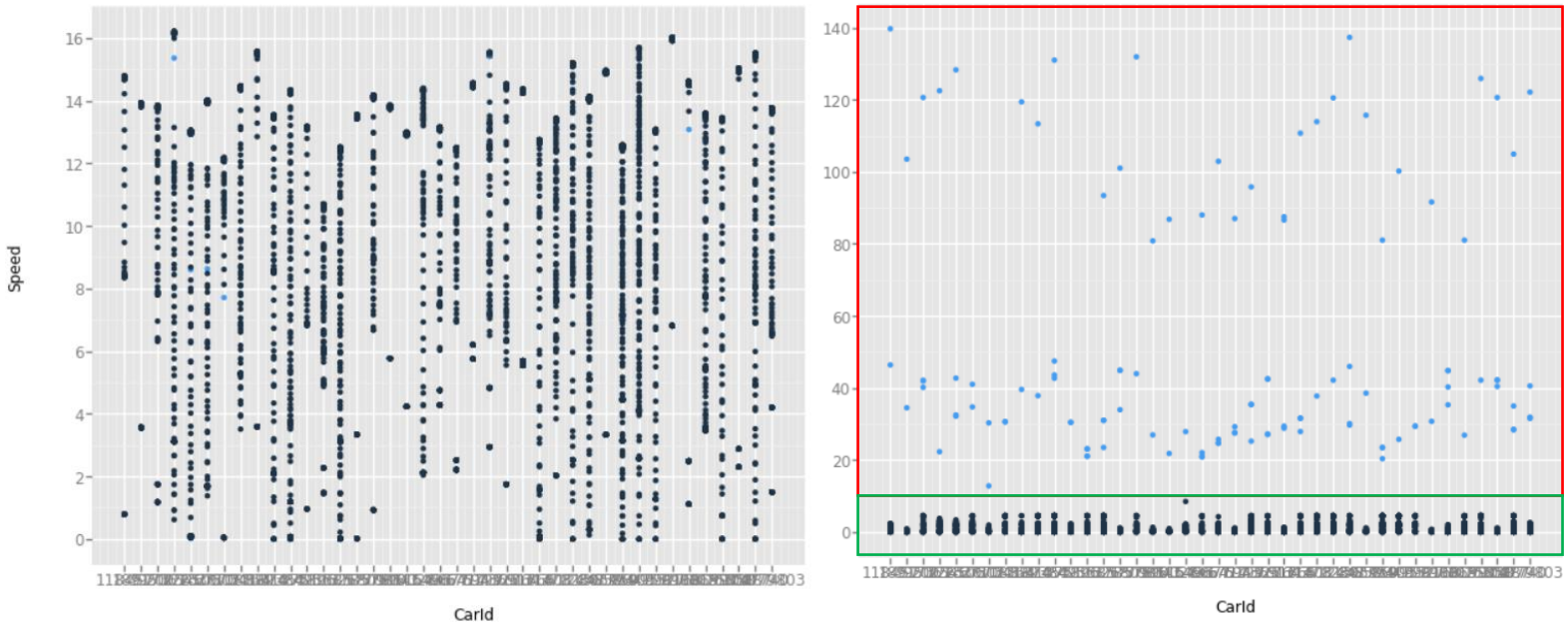


Nos anomalies correspondent donc davantage à des observations anormales uniques dans un grand jeu de données. Etant donné notre type de données, notre type d'anomalies et notre technique de détection, nous baserons notre système sur un algorithme de Clustering.

3 – Visualisations et interprétations des données

Effectuer une analyse graphique peut nous indiquer quelles sont les dimensions les plus intéressantes à étudier. Etant donné notre exemple précédent et que nos données concernent des véhicules en déplacement, nous pouvons nous attendre à voir la plupart des anomalies exposées avec des variations de vitesse et d'orientation.

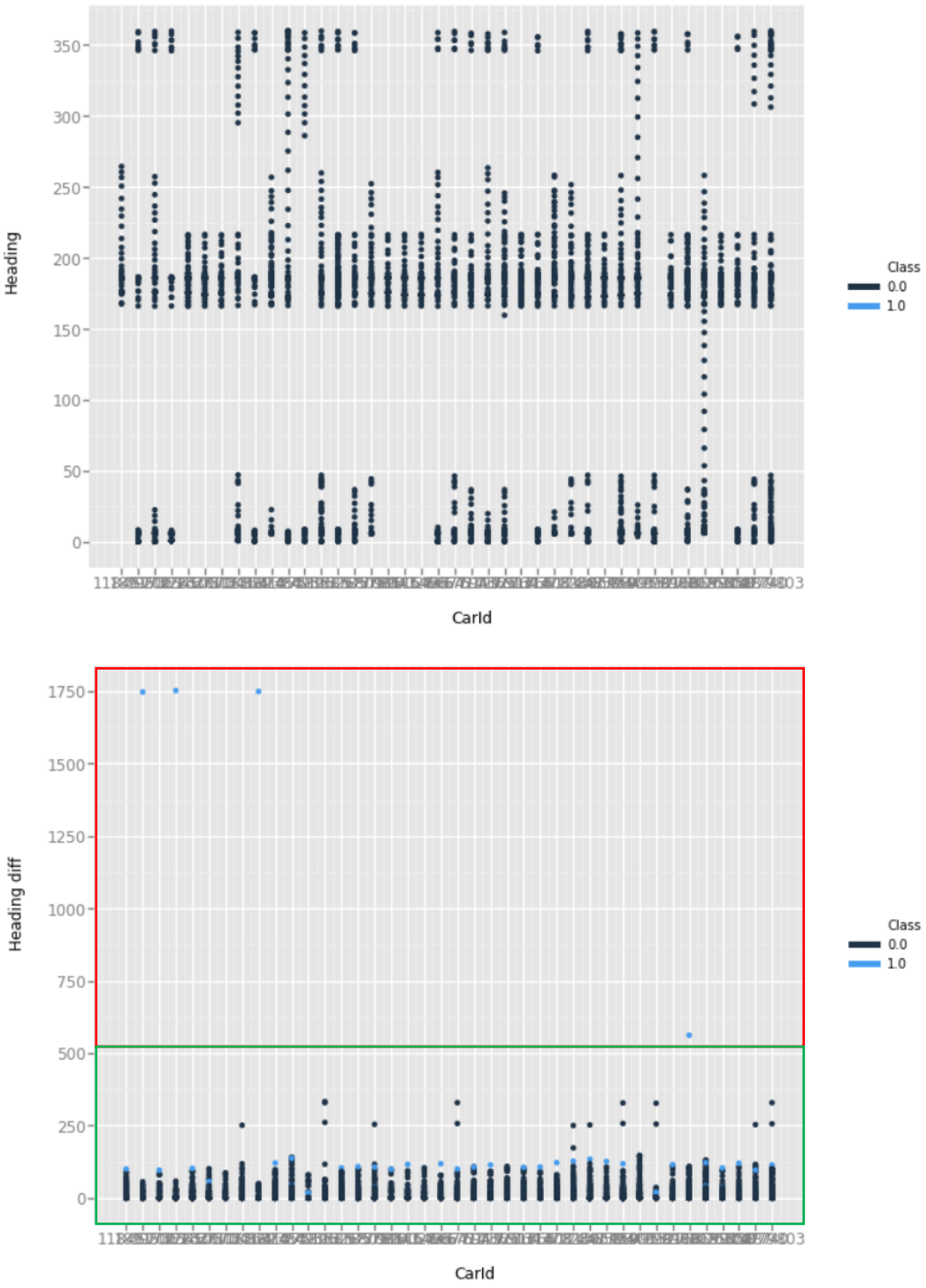
3.1 - Vitesse et accélération du véhicule



Sur le graphique de gauche, nous représentons en abscisse l'identifiant du véhicule et sa vitesse en ordonnée. L'étiquetage des observations est décrit par la coloration des points. Nous pouvons voir qu'aucune anomalie n'est isolée avec cet axe d'analyse.

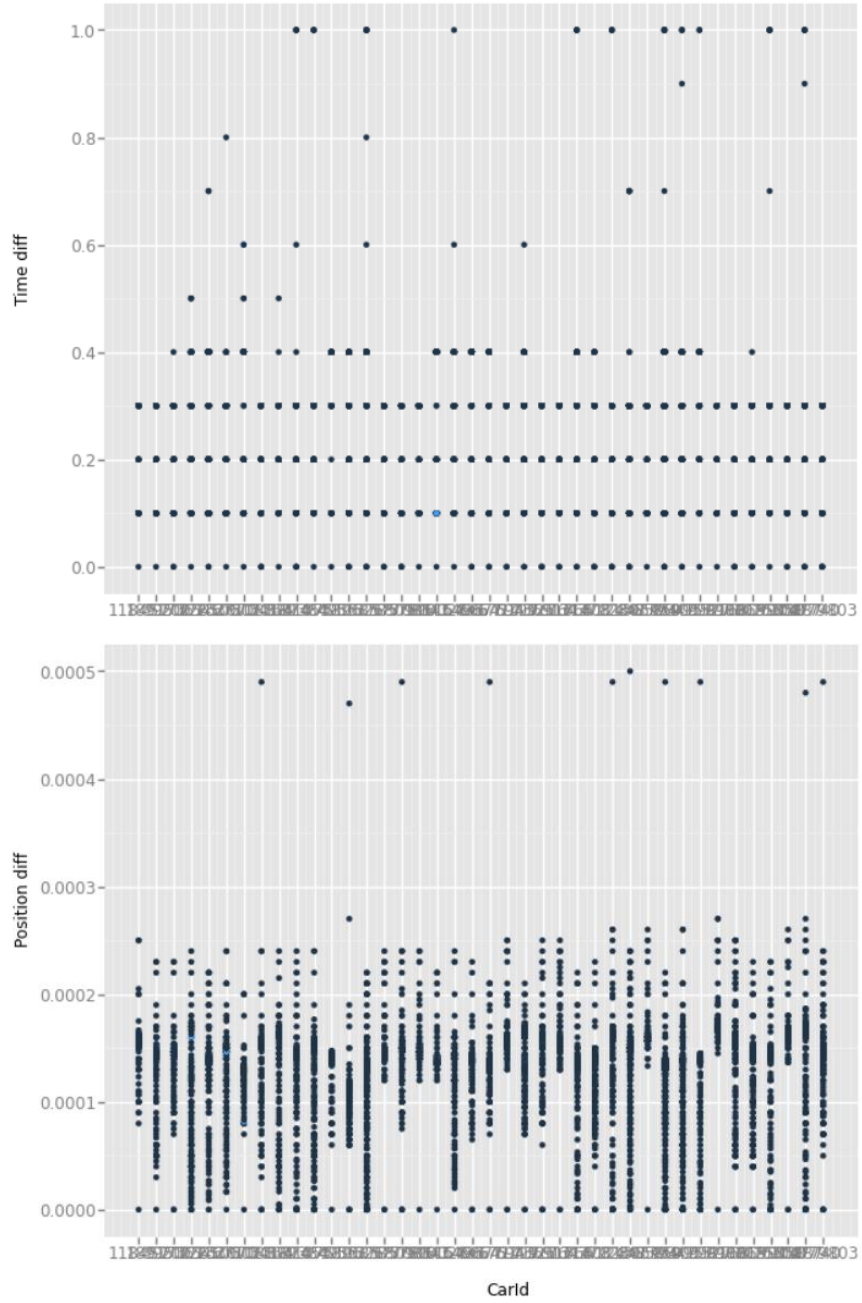
Sur le graphique de droite, nous représentons la variation de vitesse en ordonnée. Nous pouvons voir qu'un bon nombre d'anomalies peuvent être isolées du reste de notre jeu de données à partir de cette dimension. Dans l'idéal notre système de détection devra pouvoir redessiner cette zone rouge et définir les données qui y sont comme des anomalies. Cela nous confirme que les nouvelles variables comparatives calculées avec les observations précédentes sont bien des variables pertinentes à générer et que la variation de vitesse est un axe particulièrement intéressant à analyser.

3.2 - Direction et changement de direction du véhicule



Selon le même principe, nous pouvons voir que la variation de direction des véhicules peut nous permettre d'isoler des anomalies. En revanche, la direction seule n'est pas un axe de détection pertinent. Il est à noter que la direction d'un véhicule est mesurée entre 0 et 360 degrés par rapport au nord. Le calcul d'obtention de la variation d'orientation doit prendre ce paramètre en compte.

3.3 - Intervalles de temps et de position



Nous n'avons pas d'anomalie liée à des changements de position ou des intervalles de temps trop élevés entre deux observations. En revanche on peut imaginer que ces données seraient probablement intéressantes dans le cadre d'une recherche d'anomalie liée aux capteurs. Les dimensions temps, longitude et latitude semblent également plus utiles à la détection d'anomalies dues aux capteurs et transmissions plutôt qu'à la détection d'anomalies routières.

Evaluation des algorithmes

Il existe plusieurs méthodes pour évaluer la classification d'un algorithme. La courbe PR (précision et rappel) est particulièrement intéressante dans notre cas étant donné le déséquilibre entre le nombre de données normales et anormales. La courbe ROC est également un indicateur utile bien que plus adapté aux jeux de données ayant un nombre égal de données normales et anormales.

Courbe PR (Précision et rappel)

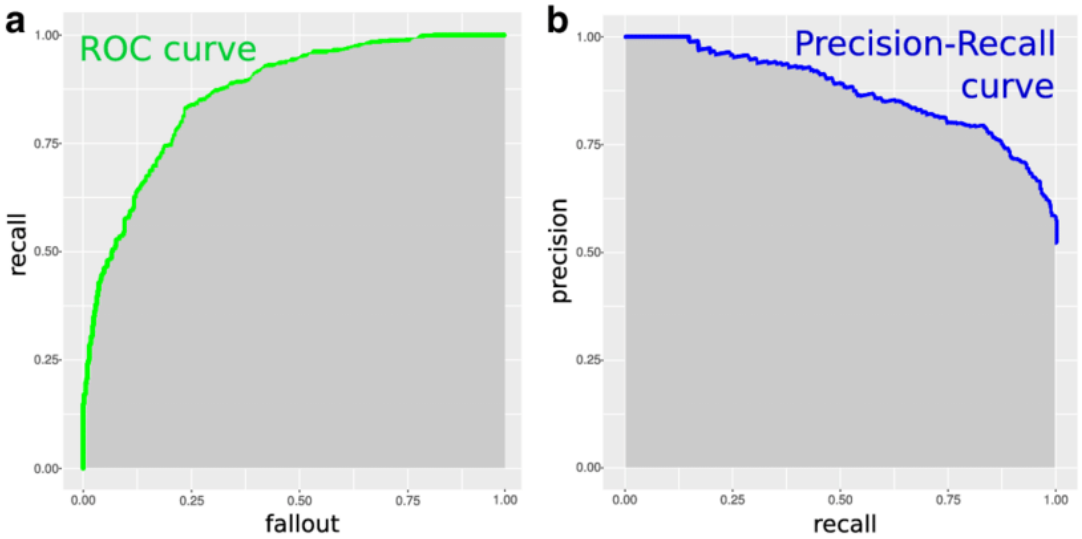
Analyser la précision et le rappel à la fois nous permet d'évaluer les performances d'un modèle de façon complète. Malheureusement, la précision et le rappel sont fréquemment en tension car l'amélioration de la précision se fait généralement au détriment du rappel et réciproquement.

La précision mesure le pourcentage d'observations identifiées comme anormales ayant été classifiées correctement. Le rappel mesure le pourcentage d'observations réelles ayant été classifiées correctement.

Courbe ROC (receiver operating characteristic)

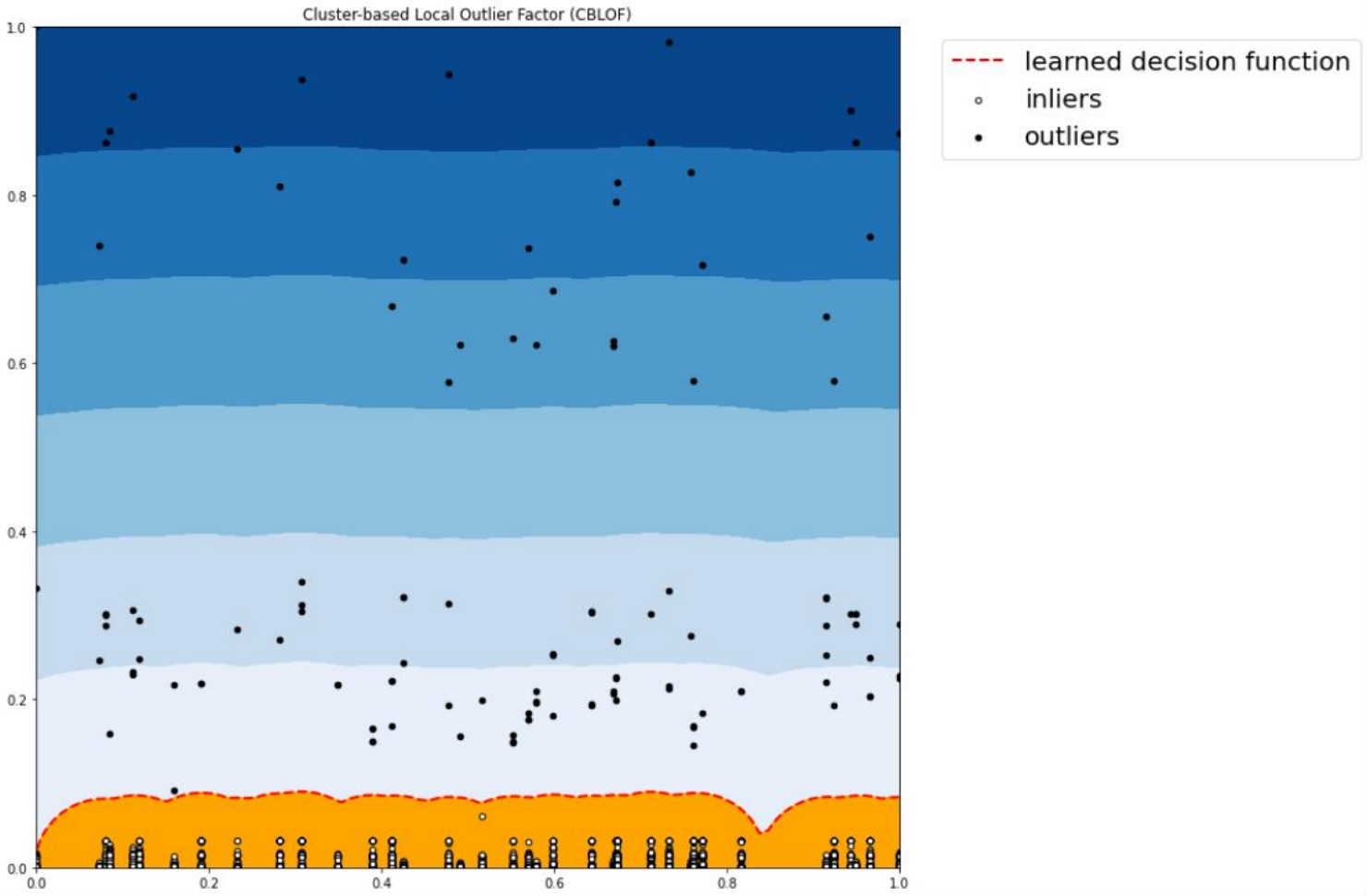
Une courbe ROC est une courbe graphique qui représente les performances d'un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs.

AUC signifie "aire sous la courbe ROC". Cette valeur mesure l'intégralité de l'aire située sous l'ensemble de la courbe ROC. On peut interpréter l'AUC comme une mesure de la probabilité pour que le modèle classe un exemple positif aléatoire au-dessus d'un exemple négatif aléatoire.



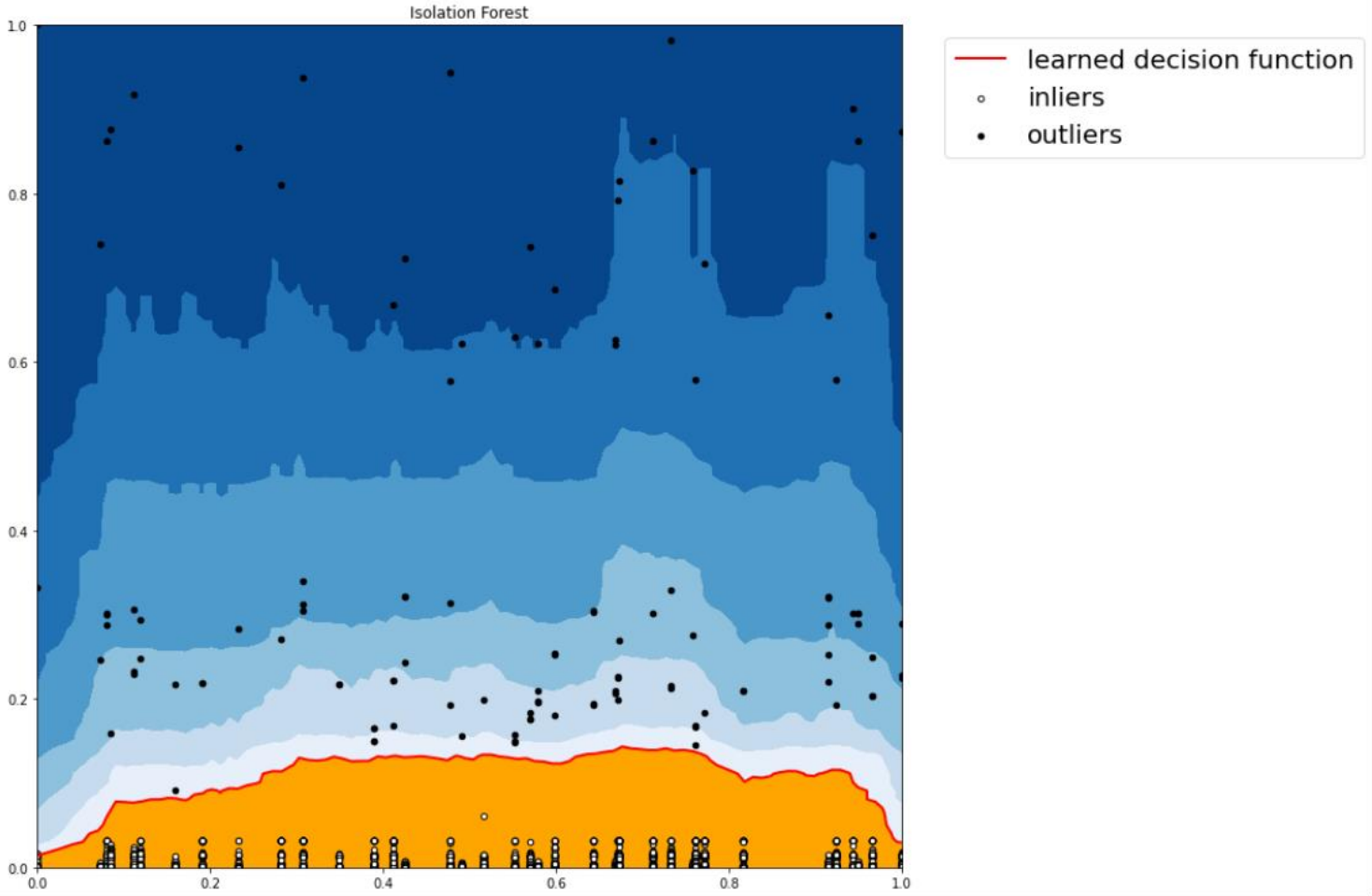
Tests des algorithmes présélectionnés sur la variation de vitesse

1) CBLOF (Clustering-Based Local Outlier Factor)



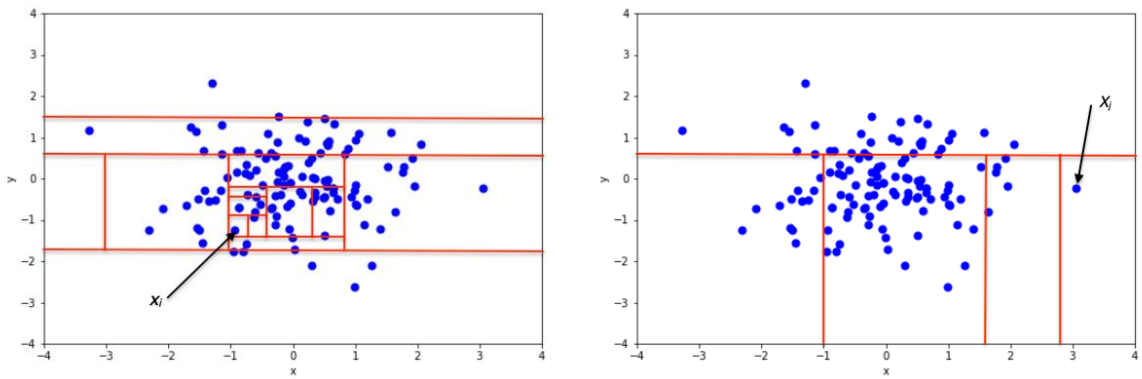
L'algorithme CBLOF prend comme entrée l'ensemble des données et une estimation de clusters à générer. Il classe les clusters en petits et en grands clusters. Le score d'anomalies est ensuite calculé en fonction de la taille du cluster auquel appartient le point ainsi que de la distance par rapport au grand cluster le plus proche. Dans notre cas, l'estimation du nombre de grands clusters peut être définie dans notre algorithme comme étant égale au nombre d'identifiants de véhicule. De cette manière nous aurons au minimum un grand cluster par véhicule.

2) IForest (Isolation Forest)

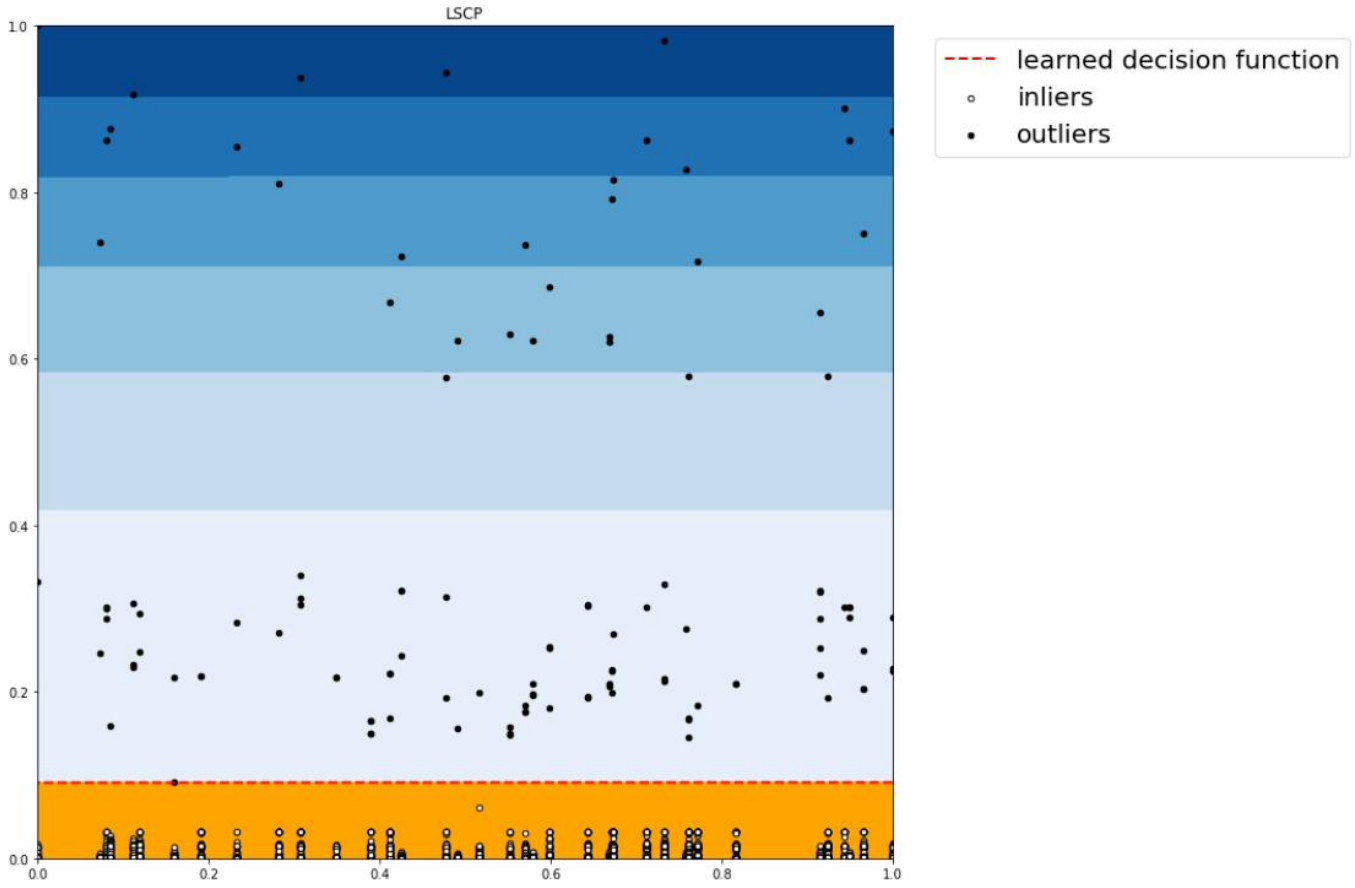


L'algorithme IForest prend chaque observation une par une et calcule son score d'anomalie en cherchant à les isoler de manière récursive. L'algorithme choisit un descripteur et un seuil de coupure au hasard, puis il évalue si cela permet d'isoler la donnée en question. Si tel est le cas, l'algorithme s'arrête, sinon il choisit un autre descripteur et un autre point de coupure aléatoirement jusqu'à ce que la donnée soit isolée. Plus la donnée est isolée rapidement, plus son score d'anomalie est élevé.

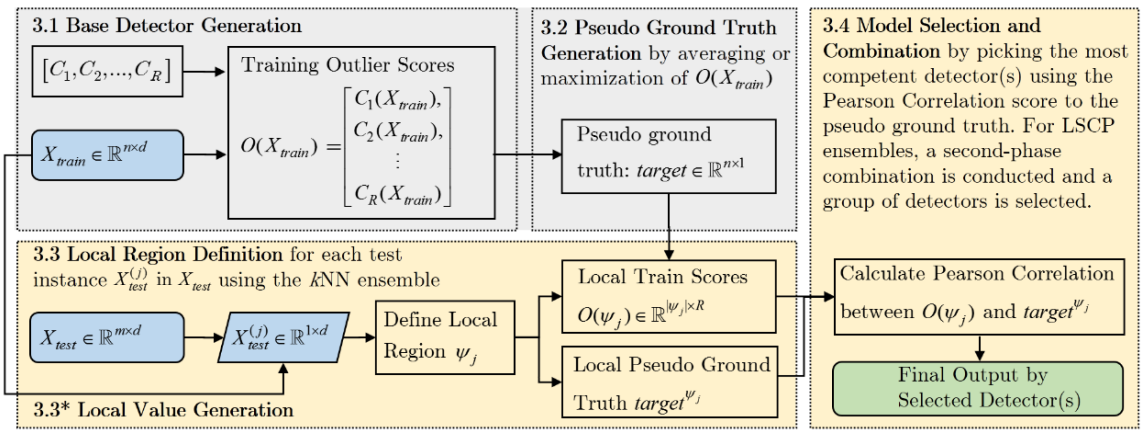
Exemple:



3) LSCP (Locally Selective Combination of Parallel Outlier Ensembles)



L'algorithme LSCP utilise un ensemble d'algorithmes de détections d'anomalies non-supervisé afin de calculer un score d'anomalie. Tout d'abord, pour chaque instance, une région locale est définie comme un lot de données composées des données les plus proches. Ensuite, pour chaque région locale, un « seuil de vérité » est défini et une corrélation est calculée entre les scores d'anomalies de chaque détecteur et le seuil de vérité défini. Un histogramme est ensuite construit à partir des scores obtenus et les détecteurs les plus compétents sont sélectionnés. Enfin, le score moyen des détecteurs compétents sélectionnés est considéré comme le score final de l'observation.



Voici le tableau comparatif des évaluations de nos algorithmes :

	Temps (s)	AUROC	AUPR	# Observation	# Dimension
CBLOF	0,05952	1	0,99999	6341	2
Iforest	0,41902	1	0,99999	6341	2
LSCP	4,42147	1	0,99999	6341	2

Nous pouvons voir que les résultats sont tous excellents mais avec un temps d'exécution tout de même plus long pour l'algorithme LSCP.

Bien que ces résultats soient encourageants, ces résultats ont été obtenus lorsque les algorithmes ont reçu l'intégralité du jeu de données en une seule fois et avec un paramètre d'estimation du nombre d'anomalie définie précisément. Afin de répondre à notre objectif initial, nous devons maintenant les évaluer sur une arrivée progressive des données afin de simuler une analyse sur des données prélevées et réceptionnées en temps réel.

Tests avec réception des données progressives

Voici le tableau comparatif des évaluations de nos algorithmes lors d'une arrivée progressive de nos données :

	Temps (s)	AUROC	AUPR	# Observation	Iterations/s
CBLOF (Window_size=200)	18	1	0,9999	6341	350
Iforest (Window_size=200)	214	0,9991	0,9705	6341	30
LSCP (Window_size=200)	57	0,9992	0,9584	6341	100
CBLOF (Window_size=1000)	21	1	0,9999	6341	300
Iforest (Window_size=1000)	218	0,9999	0,9997	6341	30
LSCP (Window_size=1000)	244	1	0,9999	6341	25

Nous pouvons voir que le temps de calcul est significativement plus élevé pour nos trois algorithmes jusqu'à devenir problématique pour les algorithmes IForest et LSCP. De plus, pour ces deux algorithmes, il faut un nombre élevé de données comparatives afin d'obtenir de bons résultats et cela rend une analyse très précise très coûteuse en temps et en puissance de calcul.

L'algorithme CBLOF en revanche nous fournit toujours d'excellents résultats avec moins de données comparatives en mémoire et un temps de calcul très court. Nous choisirons donc l'algorithme CBLOF afin de construire notre système de détection.

Système de détection basé sur l'algorithme CBLOF

Nos précédents tests ont été réalisés sur notre variable de variation de vitesse uniquement. Nous avons pu voir dans nos analyses préliminaires que les axes d'analyse les plus pertinents pour la détection d'anomalies dans notre jeu de données sont la variation de vitesse et la variation de direction. Notre système va donc analyser chaque nouvelle observation reçue et combiner les scores d'anomalies obtenus en un seul score. Si notre système accorde un plus grand score aux données anormales qu'aux données normales, alors il est efficace.

Tests sur différents jeux de données

Les deux premiers tests ont été réalisés sur des jeux de données annotées fourni par Mlle Juliet CHEBET. Notre système est parvenu à attribuer un score d'anomalie significativement supérieur aux observations anormales des deux jeux de données par rapport aux observations normales. Un tableau comportant les données et leurs scores respectifs d'anomalie est disponible sur le dépôt Github du projet.

Un troisième test à été réalisé sur un jeu de données obtenus sur la plateforme Figshare (Dact : Dataset of Annotated Car Trajectories https://figshare.com/articles/dataset/DACT_Dataset_of_Annotated_Car_Trajectories/5005289). Ce jeu de données ayant été créé et annoté dans un autre but que celui de la détection d'anomalie, il n'est pas possible d'évaluer les résultats aussi précisément qu'avec nos premiers jeux de données. En revanche, une analyse manuelle des résultats a bien démontré que toutes les données les plus anormales ont obtenu un score significativement plus élevé que les données normales.

Comparaison aux travaux existants

Beaucoup de travaux similaires ont été réalisés, tous soulignent le problème apporté par la grande diversité des anomalies, le grand nombre de facteurs impactant le comportement des conducteurs et le caractère contextuel de la normalité du comportement d'un conducteur. La grande majorité des travaux similaires basent aussi leurs systèmes sur des méthodes non-supervisées mais certain utilisent également des méthodes d'apprentissage par renforcement.

Plusieurs articles [2],[4] sont parvenus à optimiser les résultats de leurs systèmes en filtrant les données qu'ils prenaient en entrée. Pour cela, leur système ne sélectionnait que les données provenant des milliers de taxis parcourant les villes de Beijing et Harbin jour et nuit. Ces taxis sont conduits par des professionnels qui ont bien plus d'expérience au volant que les autres conducteurs. Ils connaissent bien ces villes et utilisent leurs GPS en permanence. On peut étendre ce raisonnement en sélectionnant certaines catégories d'acteurs professionnels de la route et en excluant d'autres acteurs plus susceptibles d'avoir une conduite anormale. Par exemple, on peut sélectionner les données provenant des taxis, des poids lourds internationaux et de certains services de livraison, tout en excluant les données provenant des véhicules d'urgence dont le travail amène souvent à des comportements anormaux sur la route.

En plus de filtrer les données entrantes en fonction de leurs provenances, les deux travaux cités précédemment focalisent leurs analyses sur des régions particulières. D'autres travaux [5],[1] concentrent également leurs analyses sur des régions ou routes particulières. L'idée de concentrer son analyse sur une région particulière permet de faire des analyses sur des observations ayant toutes été réalisées dans le même contexte. Ainsi la définition de conduite normale est bien plus précise.

Enfin, dans l'article « Anomaly Detection in Roads with a Data Mining Approach » [8], les auteurs de l'article sont parvenus à démontrer qu'il était possible d'étoffer davantage le concept. Une fois ces anomalies détectées, il est possible d'avoir recours à un système d'apprentissage supervisé afin de déterminer la nature des anomalies.

Conclusion

Nos travaux ont pu démontrer que la détection d'anomalies routières à partir de données récupérées en temps réel auprès de véhicules connectés est possible et présente beaucoup d'intérêt. Nos expérimentations ont montré d'excellents résultats et des temps de calcul relativement faibles.

On peut s'attendre à ce qu'une application réelle de ce système sur un territoire rencontre plus de difficultés et ne soit pas aussi précis dans ses évaluations à cause du caractère contextuel des données. En revanche, beaucoup de solutions ont été apportées à ces problèmes dans de nombreux travaux de recherche. Focaliser son analyse sur des régions distinctes permet de faire une recherche d'anomalie sur des données générées dans le même contexte et fournies donc des analyses bien plus précises et contextuelles. Sélectionné les catégories de véhicules à analyser selon des critères de fiabilité ou de professionnalisme permettront aussi d'optimiser les analyses du système.

Enfin, ce type de système de détection peut être enrichi afin d'aussi chercher à détecter le type précis d'anomalie rencontré. Ainsi, il serait possible d'anticiper les moyens requis pour résoudre un incident ou avertir les services d'urgence au moment même où un accident se produit.

Sources :

https://www.diva-portal.org/smash/get/diva2:1261957/FULLTEXT01.pdf [1]
Detection and classification of anomalies in road traffic using spark streaming.
NATHAN ADOLFO, CONSUERGA RENGIFO.

https://www.mdpi.com/1424-8220/17/3/550/htm [2]
Road Traffic Anomaly Detection via Collaborative Path Inference from GPS Snippet
HONGTAO WANG, HUI WEN, FENG YI, HONGSONG ZHU LIMIN SUN.

https://arxiv.org/pdf/2004.05958.pdf [3]
Anomaly Detection in Trajectory Data with Normalizing Flows
MADSON DIAS, CESAR MATTOS, TICIANA SILVA, JOSE MACEDO, WELLINGTON SILVA

https://www.hindawi.com/journals/mpe/2015/809582/ [4]
Detecting Traffic Anomalies in Urban Areas Using Taxi GPS Data
WEIMING KUANG, SHI AN, HUIFU JIANG.

https://i.cs.hku.hk/~yzyu/publication/GPSvas-vast2010.pdf [5]
Anomaly Detection in GPS Data Based on Visual Analytics
ZICHENG LIAO, YIZHOU YU, BAOQUAN CHEN.

https://www.researchgate.net/publication/336272457_Traffic_signal_detection_from_in-vehicle_GPS_speed_profiles_using_functional_data_analysis_and_machine_learning [6]
Traffic Signal Detection from in-vehicle GPS Speed Profiles using Functional Data Analysis and Machine Learning
YANN MENEROUX, GUILLAUME SAINT PIERRE, MOHAMMAD GHASEMI.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6891262 [7]
Data Anomaly Detection for Internet of Vehicles Based on Traffic Cellular Automat and Driving style.
NAN DING, HAOXUAN MA, CHUANGUO ZAO, YANHUA MA, HONGWEI GE.

https://repositorium.sdum.uminho.pt/bitstream/1822/50851/1/cs_centeris_silva.pdf [8]
Anomaly Detection in Roads with a Data Mining Approach
NUNO SILVA, JOAO SOARES, VAIBHAV SHAH, MARIBEL SANTOS, HELENA RODRIGUES