Conserver la meilleure précision en réduisant le nombre de dimension

Kerryghan Relot[1] Xavier Coupé[1]
Clément Hibon[1] Ambre Collard[1]
(1) Le Mans Université, LIUM, 72000 Le Mans, France
{ Kerryghan.Relot.Etu, Xavier.Coupe.Etu, Clement.Hibon.Etu,
Ambre.Collard.Etu }@univ-lemans.fr

RÉSUMÉ

Lors du traitement de données liées à la sécurité routière, nombre d'analyses sont possibles: classement de dangerosité des routes, analyse de la variation de l'accidentogénité au fil du temps ou encore évaluation de la gravité. C'est sur ce dernier point que se concentre cet article. Ce travail présente la mise en œuvre de l'optimisation de nos données dans le corpus sur la sécurité routière. Nous avons pour tâche préliminaire de faire un prétraitement des données afin de les uniformiser. Nous avons ensuite réduit le nombre de dimension de nos données en sélectionnant de manière adéquate les caractéristiques ainsi qu'en effectuant une analyse en composante principale. Pour évaluer et quantifier le gain ou la perte d'information, nous avons utilisé le modèle Random Forest tout au long de nos mesures, ceci afin d'avoir une base commune de comparaison sans introduire de variabilité due au modèle. Cet article vise à rapporter l'impact des modifications effectuées sur les données sur la précision de notre modèle de prédiction.

ABSTRACT

Conservation of the best accuracy while reducing number of dimensions.
When processing road safety data, number of analyses are possible : road hazard ranking, analysis of variations in accident rates over time or assessing severity. This article focuses on the latter. This work presents the implementation of our data optimization in the road safety dataset. Our preliminary task is to pre-process the data in order to standardize it. We then reduced the number of dimensions in our data by selecting the appropriate features as well as performing a principal component analysis. To evaluate and quantify the gain or loss of information, we used the Random Forest model throughout our measurements, in order to have a common basis for comparison without introducing model variability. Variability due to the model. The aim of this article is to report on the impact of data modifications on the accuracy of our model.

MOTS-CLÉS : Accident, Données, Gravité, Descripteur, Classes, Accident de la route, Prétraitement, ACP, Random Forest

KEYWORDS : Road accident, Gravity, Features, Data, Preprocessing, PCA, Random Forest

1 Introduction

Lors d'un accident corporel beaucoup d'informations sont collectées par les forces de l'ordre pour le caractériser. Un accident corporel est un accident survenu sur une voie ouverte à la circulation publique, impliquant au moins un véhicule et ayant fait au moins une victime ayant nécessité des soins.

Notre questionnement original se portait sur le niveau d'importance de chaque information dans la caractérisation de la gravité de l'accident. Autrement dit, nous avons tenté de rechercher quels pouvaient être les éléments prédominants – si ils existent – pour déterminer la gravité d'un accident corporel.

Notre travail se portant en priorité sur le prétraitement des données, nous avons donc décidé de ne choisir qu'un seul modèle pour l'entraînement ainsi que l'évaluation, et ceci afin de pouvoir facilement comparer nos résultats. Le modèle que nous avons choisi est RandomForest. Concernant le prétraitement plus particulièrement, nous avons mis en place plusieurs méthodes : Retypage et Nettoyage de données (data cleaning), Sélection de caractéristiques (features selection), Analyse en Composante Principale (PCA)

Après avoir présenté l'état de l'art du domaine, nous présenterons comment et pourquoi nous avons nettoyé nos données. Ensuite, nous décrirons nos différentes démarches et approches pour la sélection de caractéristiques, avant de finir en explicitant la réalisation de notre analyse en composante principale et comment elle nous a permis de réduire la dimensionnalité de nos données.

2 Travaux connexe

Il est à noter que notre travail s'inscrit dans un cadre plus général non seulement sur des techniques de Machine Learning mais aussi sur le thème de l'accidentologie routière.
En effet, bien qu'il soit allé plus loin que nous dans sa démarche, l'article de recherche par Miaomiao et Yindong[i] présente une problématique similaire à la notre, ce qui montre que notre approche n'est pas dénuée de sens et est même un sujet bien étudié.
Quant à la pertinence du choix du modèle RandomForest dans notre étude, nous pouvons constater dans cet article de Santos, Dias et Amado[ii], que les auteurs – après avoir effectué un revue de 51 articles scientifiques – nous montrent que RandomForest semble être le modèle prodiguant les meilleurs résultats pour prédire la gravité d'un accident de la route.

3 Méthodologie

3.1 Données Utilisées

Pour cette étude nous nous sommes basés sur les données d'accidents corporels mis à disposition par le Ministère de l'Intérieur[iii]. Plus particulièrement nous nous sommes concentrés sur les données pour l'année 2022 afin de ne pas nous éparpiller et d'avoir une analyse basée sur des données récentes. Au moment de nos premières recherches, les données les plus récentes étaient les données pour l'année 2022 ; aussi, depuis 2019 la catégorisation des blessés a changé dans le jeu de données, c'est une raison de plus qui nous a convaincu dans le

choix de données récentes, afin de pouvoir nous baser sur des caractéristiques qui ne changent pas selon les années. Nos données sont réparties en quatre fichiers distincts: *usagers.csv*, *lieux.csv*, *vehicules.csv*, *caractéristiques.csv*. Nous verrons donc ci-dessous que nous avons dû faire une jointure pour les réunir en un seul dataset afin de pouvoir exécuter notre modèle dessus.

3.2 Sélection de caractéristiques

Dans un premier temps, nous avons analysé nos données et la documentation qui était fournie avec afin de déterminer quelles étaient les features redondantes ou qui ne pourraient pas nous apporter d'information. Après avoir effectué notre analyse, il y a donc 14 caractéristiques que nous avons pu supprimer :

- **id_usager:** cette colonne n'apparaît que dans le fichier `usagers` et ne contient que des valeurs uniques (permettant d'identifier de manière unique chaque usager).
- **id_vehicule:** Pareil ici, `id_vehicule` n'apporte aucune information sur l'accident. Cet attribut nous servira juste de clé primaire pour joindre les tables relatives aux usagers et aux véhicules. Mais une fois cette opération faite, nous pourront enlever cette colonne.
- **num_veh:** ceci est un identifiant propre à chaque accident pour identifier chaque véhicule impliqués. Mais il est repris entre différents accidents, on peut donc le supprimer également car il ne sert à rien pour notre analyse.
- **occutc:** est null la très grande majorité du temps puisque cette variable ne concerne que les transports en commun. Il y a trop de valeurs nulles pour qu'on puisse vraiment tirer quelques choses de concluant avec cette feature.
- **voie**, **V1**, **V2**, **pr**, **pr1**, **adr**: les valeurs ne sont pas forcément uniques pour chaque route (en effet, les numéro de voies, les adresses et autres peuvent être communs à plusieurs lieux). Ces caractéristiques ne donnent donc aucune information permettant de catégoriser de manière unique un accident.
- **lartpc**: définie la largeur du terre-plein central mais n'est déclarée sur seulement 0.0005% des données.
- **an:** l'année ne nous apporte aucune information puisque nous avons préalablement sélectionné les données relatives à 2022 uniquement.
- **lat**, **long:** La latitude et longitude ne nous seront d'aucune utilité ici car ce sont des valeurs continue assez difficile à discrétiser correctement. Nous prenons donc la décision de nous passer de cette information géographique pour notre étude. De plus, cela nous permettra de faire une évaluation de la gravité indépendamment de la localisation de l'accident.

3.3 Nettoyage de données

Il y a aussi quelques caractéristiques que nous avons pu modifier :

- **an_nais:** cette colonne ne contient que des années à 4 chiffres, on peut donc convertir les années de naissance en age puisque nous nous concentrons sur les données de 2022. Et les années non renseignées peuvent être misent à -1.
- **actp:** il s'agit d'une valeur à un chiffre en hexadécimale entre -1 et B. On peut simplement convertir cela en entier pour s'éviter d'avoir à manipuler des chaines de caractères.
- **larrout:** nous convertirons la largeur de la route en centimètres afin d'avoir des valeurs entière uniquement et non des nombres flottant.

- **nbv:** plusieurs types de données sont mélangées pour cette caractéristique. Nous convertirons donc toutes ces caractéristiques en entier.
- **hrmn:** nous effectuerons une discrétisation sur cette caractéristique. Cela servira plusieurs but. Le premier d'entre eux est de réduire le nombre de valeurs différentes prises par la feature. Cela nous permettra également de convertir ces valeurs en entiers. Enfin, pour effectuer cette discrétisation nous découperons la journée en bloc de 12 minutes, nous compterons donc le nombre de fois ou 12 minutes se sont écoulées depuis minuit. Nous avons choisi cet intervalle de 12 minutes car c'est le plus petit intervalle de temps qui nous permettent d'effectuer notre discrétisation sur un entier 8bit uniquement (il y a 120 fois 12 minutes dans une journée). De plus, 12 minutes nous semble être une précision assez raisonnable pour notre étude.
- **dep**, **com:** nous réduirons l'intervalle de valeurs prises par ces deux caractéristiques en rendant ces dernières contiguës.

Suite à ces modifications nous avons réduit toutes les caractéristiques (à l'exception de 3 d'entre elles) à des entiers 8bits. Les trois caractéristiques qui n'ont pas été converties en int8 sont :

- **Num_Acc:** entier 64 bits
- **com:** entier 16 bits
- **larrout:** entier 16 bits

3.4 Jointure des données

Nous avons joint les données *usagers* et *vehicules* sur la clé primaire *id_vehicule*. Une fois cette jointure faites, nous avons pu supprimer la caractéristique id_vehicule qui ne nous servira plus.

Ensuite, nous avons joint cette nouvelle table aux tables *lieux* et *caractéristiques* sur la clé primaire *Num_Acc*.

3.5 Analyses en Composante Principale

Enfin, nous avons aussi réaliser une Analyse en Composante Principale, avec des données centrées réduites et des données non-centrées réduites pour comparer l'impact de la normalisation de nos données. Pour ce faire nous avons utilisé le module '*StandardScaler*' de scikit-learn.

4 Résultats / Discussion

Après avoir pré-traité nos données et effectué nos Analyses en Composantes Principales nous nous sommes penchés sur deux métriques en particulier : La Variance cumulée et L'Accuracy de notre modèle. Comme mentionné précédemment, nous avons donc choisi comme modèle « Random Forest » et plus précisément, nous avons utilisé l'implémentation de RandomForest fournie par le module python Scikit-learn. Il en est de même pour l' Analyses en Composantes Principales.

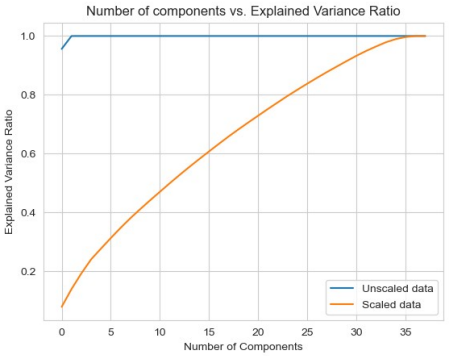

Figure 1: Comparaison de l'information cumulée porté par les n premiers axes

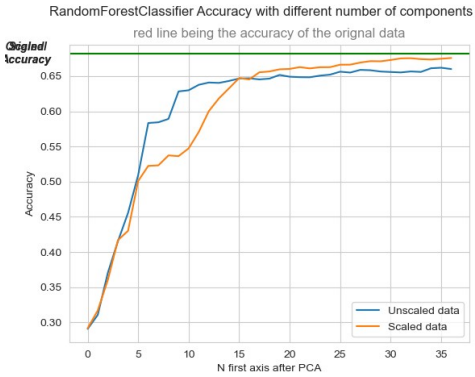

Figure 2: Comparaison de l'Accuracy en fonction du nombre de composants choisi

Comme nous pouvons le constater en Figure 1, sur des données non-centrées-réduites, la majeure partie de la variance est porté par le premier axe de l'ACP. En effet, rajouter plus d'axe par la suite n'ajoute que peu d'information. Tandis que les données préalablement centrées-réduites ont une répartition de la variance plus homogène. Cela nous indique que l'échelle de nos différentes caractéristiques est très hétérogène. En effet le premier axe principal suffit à capturer 95,59 % de la variance totale. C'est pourquoi l'on préfère généralement utiliser des données centrées-réduites lorsque l'on effectue une ACP, afin qu'une seule caractéristique ne domine pas la variance dû à son échelle disproportionnée.

Dans un second temps, nous avons comparé l'Accuracy de notre modèle en fonction du nombre d'axes principaux gardés. Bien entendu plus on rajoute d'axes plus l'accuracy augmente. Nous pouvons cependant observer un phénomène intéressant, à partir de 15 axes l'accuracy de notre modèle avec les données centrées-réduites dépasse celle du modèle avec les données non-standardisées, ce qui montre bien l'intérêt d'un tel processus.

5 Conclusion

Pour conclure, nous avons montré dans cet article qu'il était possible de réduire le nombre de descripteurs tout en permettant d'obtenir des métriques optimales.

Nous avons en premier lieu effectué une analyse en composante principale qui nous a permis de faire ressortir les axes principaux dans le corpus et ceci permet d'obtenir une accuracy correcte avec le modèle RandomForest.

Enfin l'analyse en composante principale aurait pu être plus efficace si le corpus avait moins de données manquantes et avec des données plus conséquentes.

Références

[i] MIAOMIAO Y. & YINDONG S. (2022). Traffic Accident Severity Prediction Based on Random Forest. Sustainability, 14, 1729. DOI : 10.3390/su14031729

[ii] SANTOS K., DIAS J. P., AMADO C. (2021). A literature review of machine learning algorithms for crash injury severity prediction. Elsevier, 254-269. DOI : 10.1016/j.jsr.2021.12.007

[iii] Ministère de l'Intérieur (màj 2024). Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2023. data.gouv.fr (*consulté le 20/09/2024*)