Méta-algorithme de classement. Application à la sécurité routière

Zaïd Ouni [1,2]

[1] *LAB, 132, rue des Suisses, 92000 Nanterre*
[2] *Modal'X, Université Paris Ouest Nanterre, 200, av. de la République, 92000 Nanterre*
zaid.ouni@lab-france.com

Résumé. Chaque année, les données BAAC (Bulletin d'Analyse des Accidents Corporels) incluent tous les accidents de la circulation sur la voie publique française impliquant un ou plusieurs véhicules légers et blessant au moins un des occupants. Chaque véhicule léger se voit associer une "classe générationnelle" (CG), qui en donne une description sommaire. A contextes accidentels donnés, deux véhicules légers de CGs distinctes n'offrent pas nécessairement la même sécurité à leurs passagers. L'objectif de ce travail est d'évaluer dans quelle mesure les nouvelles générations de véhicules légers sont plus sûres que les anciennes à partir des données BAAC.

Nous procédons par "scoring" : nous cherchons une fonction de score qui associe à tout contexte et toute CG un nombre réel ; plus ce nombre est petit, plus la CG est sûre dans le contexte accidentel donné. Une meilleure fonction de score est apprise à partir des données BAAC par validation croisée, sous la forme d'une combinaison convexe optimale de fonctions de scores produites par une librairie d'algorithmes de classement par scoring. Une inégalité oracle illustre les performances du méta-algorithme ainsi obtenu. Nous implémentons ce méta-algorithme, l'appliquons, et en montrons quelques résultats.

Collaboration avec A. Chambaz (Modal'X, Université Paris Ouest Nanterre), C. Chauvel (LAB), C. Denis (LAMA, Université Marne-la-Vallée).

Mots-clés. Classement, Inégalité oracle, Scoring, Sécurité automobile, Super-Learning, Validation croisée

Abstract. Each year, the BAAC (Bulletin d'Analyse des Accidents Corporels) data set includes all traffic accidents on the French public roads involving one or several ligth vehicles and injuring at least one of the passengers. Each light vehicle is associated with its "generational class" (GC), which gives a raw description of the vehicle. In two given contexts of accident, two light vehicles with two different GCs do not necessarily offer the same level of safety to their passengers. The objective of this study is to assess to which extent more recent generations of light vehicles are safer than older ones based on the BAAC data set.

We rely on "scoring" : we look for a score function that associates any context and any GC with a real number ; the smaller is this number, the safer is the GC in the given context

of accident. A better score function is learnt from the BAAC data set by cross-validation, under the form of an optimal convex combination of score functions produced by a library of ranking algorithms by scoring. An oracle inequality illustrates the performances of the resulting meta-algorithm. We implement it, apply it, and show some results.

Joint work with A. Chambaz (Modal'X, Université Paris Ouest Nanterre), C. Chauvel (LAB), C. Denis (LAMA, Université Paris-Est Marne-la-Vallée).

Keywords. Car safety, Cross-validation, Oracle inequality, Ranking, Super-Learning, Scoring

1 Introduction

En 2015, les accidents de la route restent une priorité mondiale, européenne et française pour l'ensemble des acteurs de la sécurité routière. Les automobiles sont un des acteurs principaux de l'activité routière. L'amélioration de l'activité routière passe donc notamment par une analyse des caractéristiques accidentologiques des automobiles. Les modèles de véhicule sont développés en bureaux d'études et validés en laboratoires. C'est néanmoins la réalité accidentologique qui permet de vraiment cerner les niveaux qu'ils offrent en matière de sécurité active (grâce aux systèmes d'aide à la conduite, qui assurent, par exemple, une meilleure tenue de route et un meilleur freinage) et de sécurité passive (grâce, par exemple, aux ceintures, airbags, structures à déformation programmée). Dans ce cadre, les experts en accidentologie du LAB (Laboratoire d'accidentologie, de biomécanique et du comportement conducteur) souhaitent disposer d'un outil statistique leur permettant, en interne, de suivre l'évolution au cours du temps de la sécurité offerte par les générations des véhicules.

Nous nous appuyons sur les données BAAC (Bulletin d'Analyse des Accidents Corporels). Elles répertorient, chaque année, tous les accidents de la route ayant eu lieu sur la voie publique française et ayant conduit à au moins un blessé léger. Les bulletins sont établis par les forces de l'ordre. Ils décrivent les conditions générales de l'accident. Ils nous permettent de déterminer :

- quand (date, horaire), où (localisation géographique), dans quelles conditions (nature et état de la route, type de collision) ont eu lieu les accidents ;
- une description du ou des conducteurs impliqués (âge, sexe, catégorie socio-professionnelle, alcoolémie, validité du permis de conduire) ;
- une description des éventuels passagers (âge, sexe, catégorie socio-professionnelle) ;
- quelles ont été les conséquences de l'accident pour les personnes impliquées (indemne ou blessé léger, blessé grave ou tué).

En complément de ces données nationales, des données de flottes automobiles permettent d'associer une classe générationnelle (CG) à chacun des véhicules impliqués. Constituée de sept variables (segment, année de conception et cinq autres variables), la CG d'un

véhicule le décrit sommairement. Dans la suite, les données BAAC sont associée avec les données des CGs.

2 Modélisation

Nous utilisons les données BAAC 2011 pour l'apprentissage et les données BAAC 2012 pour la validation. Nous nous restreignons aux accidents qui n'ont impliqué qu'un ou deux véhicules légers, soit plus de 15 000 accidents pour plus de 30 000 personnes impliquées en 2011 ainsi qu'en 2012. Parce qu'un ou deux véhicules sont impliqués, et parce que nous adoptons le point de vue individuel des occupants des véhicules, les données viennent par "clusters".

Notons $\mathbb{O}^1, \dots, \mathbb{O}^n$ les n observations d'un jeu de données BAAC. Nous les modélisons comme des variables aléatoires indépendantes, identiquement distribuées selon la loi $\mathbb{P}$. Considérons la $i$ème observation $\mathbb{O}^i$ correspondant au $i$ème accident.

- Si un seul véhicule léger est impliqué dans l'accident, alors $\mathbb{O}^i = \mathbb{O}^i_1$.
- Si deux véhicules sont impliqués dans l'accident, alors $\mathbb{O}^i = (\mathbb{O}^i_1, \mathbb{O}^i_2)$.
- Pour $k = 1, 2$, $\mathbb{O}^i_k$ se décompose en $\mathbb{O}^i_k = (O^i_{k1}, \dots, O^i_{kJ_k})$, où J_k est le nombre d'occupants du $k$ème véhicule.
- Pour $k = 1, 2$ et $j = 1, \dots, J_k$ fixés, la variable $O^i_{kj} = (Y^i_{kj}, Z^i_{kj})$ décrit l'accident du point de vue du $j$ème occupant du $k$ème véhicule. La variable $Y^i_{kj} = (W^i_{kj}, X^i_{kj})$ contient les données contextuelles W^i_{kj} et la CG du véhicule X^i_{kj}. La variable $Z^i_{kj} \in \{0, 1\}$ décrit quant à elle le degré de sévérité des conséquences de l'accident sur l'occupant (indemne ou blessé léger, blessé grave ou tué).

Nous avons déjà dit que nous nous intéressons tout particulièrement au point de vue des occupants des véhicules. Nous pouvons désormais formaliser cette affirmation comme suit : nous nous intéressons tout particulièrement à la loi commune des O^i_{kj} qui, sous des hypothèses raisonnables, s'écrit

$$P = \sum_{J=1}^{J_{\max}} \mathbb{P}(K = 1, J_1 = J)\mathbb{P}_{K=1,J_1=J} + \sum_{J=1}^{J_{\max}} \mathbb{P}(K = 2, J_1 = J)\mathbb{P}_{K=2,J_1=J},$$

où nous notons $\mathbb{P}_{K=\kappa,J_1=J}$ la loi conditionnelle commune des J composantes de $\mathbb{O}_1 = (O_{11}, \dots, O_{1J})$, l'accident décrit du point de vue du premier véhicule, sachant $K = \kappa$ (κ véhicules sont impliqués dans l'accident) et $J_1 = J$ (le premier véhicule transporte J occupants). Nous démontrons un lemme qui nous permet, quel que soit le paramètre d'intérêt Ψ, d'apprendre $\Psi(P)$ à partir de l'échantillon $\mathbb{O}^1, \dots, \mathbb{O}^n$ qui est tiré sous $\mathbb{P}$ et non sous P. Ceci s'applique par exemple à $y \mapsto \Psi(P)(y) = P(Z = 1|Y = y)$.

3 Elaboration par validation croisée d'un méta-algorithme de classement par "scoring"

Nous élaborons un algorithme de classement fondé sur le principe du "scoring". Il existe déjà dans la littérature de telles procédures de classement, voir par exemple [Freund et al., 2004, Clémençon et al., 2008, Clémençon et al., 2009]. La nôtre se démarque de l'état de l'art à deux titres : d'une part, elle est capable de s'adapter à la dépendance existant entre les composantes de chacune des observations indépendantes et identiquement distribuées ; d'autre part, elle s'appuie sur le principe de la validation croisée pour élaborer un méta-algorithme de classement à partir d'une librairie d'algorithmes de classement dont on ne sait pas à l'avance lequel va s'avérer être le plus performant.

3.1 Principe du "scoring"

Pour mettre en place le classement, nous procédons par "scoring" :

1. nous construisons une fonction de "scoring" $s : \mathcal{Y} \to [0, 1]$,
2. nous attribuons un score $s(w, x)$ à toute combinaison (w, x) d'un contexte d'accident w associé à une CG x,
3. nous décidons que la combinaison (w, x) est plus sûre que la combinaison (w', x') si $s(w, x) \leq s(w', x')$.

Notons r_s une telle procédure fondée sur la fonction de score s. La performance statistique de r_s est évaluée en termes de risque de "ranking" $R(r_s) = E_{P^{\otimes 2}}(L(r_s, O, O'))$, pour la perte $L(r_s, O, O')$ caractérisée par

$$L(r_s, O, O') = \mathbf{1}\{(Z - Z')(s(Y) - s(Y')) < 0\})$$

avec $O = (Y, Z)$ et $O' = (Y', Z')$. Il est bien connu [Freund et al., 2004] que la règle de classement r_π associée à $y \mapsto \pi(y) = P(Z = 1|Y = y)$ est optimale. Précisément, il apparaît que, quelle que soit la règle de "scoring" r_s, on a

$$0 \leq R(r_s) - R(r_\pi) \leq 2E_P(|\pi(Y) - s(Y)|).$$

L'optimalité est en fait même valable sur la classe de toutes les règles de classement, et pas seulement sur celle de classement par "scoring".

3.2 Construction d'un méta-algorithme de classement par validation croisée

Il y a autant de règles de "scoring" que d'estimateurs de la fonction π.

Soit $\widehat{\Psi}^1, \ldots, \widehat{\Psi}^K$ K algorithmes d'estimation de π. Soit $k = 1, \ldots, K$ et $I \subset \{1, \ldots, n\}$ fixés arbitrairement. Le sous-ensemble $\{\mathbb{O}^i : i \in I\}$ est associé

— à la loi empirique $\mathbb{P}_I = (\text{card}(I))^{-1} \sum_{i \in I} \text{Dirac}(\mathbb{O}^i)$,
— donc, via l'algorithme $\widehat{\Psi}^k$, à la fonction (empirique) $\widehat{\Psi}^k(\mathbb{P}_I) : \mathcal{Y} \to [0,1]$ qui a vocation à approcher π,
— et donc, par plug-in, à la règle de classement r_s où $s = \widehat{\Psi}^k(\mathbb{P}_I)$.

Au lieu d'identifier et de sélectionner un unique, meilleur algorithme, nous construisons une combinaison convexe des algorithmes que nous appelons un méta-algorithme. Il prend la forme

$$\widehat{\Psi}_n^* = \sum_{k=1}^{K} \alpha_n^k \widehat{\Psi}^k$$

pour un optimal $\alpha_n \in [0,1]^K$ tel que $\sum_{k=1}^K \alpha_n^k = 1$ déterminé par validation croisée. Cette façon de procéder est connue sous le nom de Super Learning dans la littérature. Elle a été introduite par [van der Laan et al., 2007].

Nous démontrons grâce à une inégalité oracle que la règle de classement fondée sur $\widehat{\Psi}_n^*$ est presqu'aussi performante pour établir des classements que la meilleure des règles de classement fondées sur $\widehat{\Psi}^1, \ldots, \widehat{\Psi}^K$. La mesure de performance s'appuie sur le risque R introduit dans la section précédente.

4 Application

Le méta-algorithme de classement r_s, $s = \widehat{\Psi}_n^*(\mathbb{P}_n)$, est construit à partir de $K = 49$ algorithmes individuels et des $n = 16\ 877$ accidents qui ont eu lieu entre un ou deux véhicules légers en 2011. Le nombre de personnes impliquées s'élève à $\sum_{i=1}^n \sum_{k=1}^{K^i} J_k^i = 37\ 721$. L'échantillon de validation correspond aux 15 852 accidents qui ont eu lieu en 2012 entre un ou deux véhicules légers, pour un total de 35 636 impliqués. Les données contextuelles W_{kj}^i regroupent 25 variables. Rappelons que les CGs X_k^i regroupent sept variables.

Dans l'application, nous évaluons la performance de notre méta-algorithme à l'aide de l'aire sous la courbe ROC (AUC). Sa valeur est estimée à 80% avec $[79\%, 81\%]$ comme intervalle de confiance à 95%.

La validation industrielle de notre approche repose notamment sur la comparaison de CGs de véhicules de différentes générations au sein du même segment. Il est attendu qu'au sein d'un même segment, un véhicule d'une génération plus récente surclasse un véhicule d'une génération plus ancienne en termes de sécurité passive. Pour confronter notre méta-algorithme à cet a priori industriel, nous extrayons ainsi un contexte d'accident de chaque entrée des données BAAC 2012 (dédiées à la validation). Pour chaque segment, nous élaborons une CG pour chaque génération. Nous classons ensuite les CGs dans tous les contextes accidentels. Nous dénombrons le nombre de fois où l'ordre attendu est effectivement observé. Selon le nombre de générations comparées, les pourcentages de

classements observés en accord avec l'a priori industriel varient de 27% (cinq générations) à 99% (deux générations). Lorsque nous comparons plus de deux générations, les performances sont dégradées, mais les comparaisons deux à deux sur lesquelles s'appuie le classement global sont bonnes.

Nous évoquerons plus en détails ces résultats, ainsi que d'autres qui confirment que notre procédure est performante.

5 Discussion

Les CGs et les données contextuelles ont vocation à être enrichies, par exemple avec les dimensions, silhouettes, systèmes de sécurité embarqués, vitesses d'impact. Cet enrichissement ne remettra en cause ni la théorie ni les algorithmes que nous avons développés.

Bibliographie

[1] S. Clémençon and N. Vayatis(2009), True-based ranking methods. IEEE Trans. Inform. Theory, 55(9) :4316-4336.

[2] S. Clémençon, G. Lugosi, and N. Vayatis (2008), Ranking and empirical minimization of U-statistics, Ann.statist., 36(2) :844-874, 2008.

[3] Y. Freund, R. Iyer, R. E. Shapire and Y. Singer (2004), An efficient boosting algorithm for combining preferences, J. Mach. Learn. Res., 4(6) :933-969.

[4] M. J. van der Laan, E. Polley and A. E. Hubbard (2007), Super learner, Stat. Appl. Genet. Mol. Biol., 6 : Art. 25.