UNIVERSITÉ CÔTE D'AZUR

Forêts aléatoires à partir de données mixtes, quantitatives et textuelles :
application à l'orientation des étudiants du supérieur

Contacts :

- Christel Dartigues-Pallez : Christel.DARTIGUES-PALLEZ@univ-cotedazur.fr
- Gaëtan Rey : Gaetan.REY@univ-cotedazur.fr
- Johan Montagnat: johan.montagnat@univ-cotedazur.fr

Description générale du sujet

L'orientation des jeunes bacheliers est un point qui chaque année génère un très grand stress chez les bacheliers eux même, mais également pour leurs parents et, dans une moindre mesure, chez les responsables de formation du supérieur (en effet, trop de formations luttent contre un taux d'échec trop élevé). Le projet que nous présentons dans cette demande de financement de bourse de thèse est parti de ce constat et de la certitude que nous avons que ce processus peut (doit) être amélioré.

Pour cela, nous nous appuyons sur nos travaux de recherche dans le domaine de l'apprentissage supervisé, et notamment sur les approches basées sur les forêts aléatoires (Random Forests, RF). Nous utilisons les RF pour traiter les données provenant de capteurs portés par des sujets ou de flux vidéo illustrant des personnes réalisant des actions et nous déterminons de manière pertinente quelle action une nouvelle personne est en train de faire [5, 7].

Les forêts aléatoires peuvent aussi s'appliquer à d'autres domaines et à d'autres types de données que des données vidéo ou les capteurs portés. Nous avons choisi de les utiliser pour apprendre les résultats probables d'étudiants qui sont inscrits dans une formation du supérieur (réussite ou échec) à partir des informations connues sur leur cursus au lycée. Plus concrètement, nous considérons les données suivantes :

- Les notes obtenues par des étudiants à une formation donnée (dans notre cas le premier semestre du département Informatique de l'IUT de Nice Côte d'Azur),
- Les données relatives à la scolarité antérieure de ces étudiants
    - Lycée d'origine,
    - Filière suivie,
    - Notes obtenues par les étudiants en première et terminale,
    - Pour chaque matière suivie par l'étudiant au lycée, moyenne de la classe, plus haute et plus basse moyenne de la classe,
    - Notes obtenues aux épreuves anticipées du baccalauréat.

Nous avons pour le moment réalisé un prototype en nous basant sur les étudiants inscrit en première année du DUT Informatique de l'IUT de Nice Côte d'Azur. À l'aide de toutes ces informations, et après un pré-traitement pour anonymiser les données et pour ajouter des calculs de moyennes et de variations de notes qui nous paraissaient pertinentes, nous avons utilisé les RF ainsi que d'autres modèles d'apprentissage (SVN, Réseaux de neurone, …) pour apprendre des modèles capables de prédire la réussite ou l'échec d'un étudiant en première année de DUT Informatique. Notre prototype a donné des résultats très encourageant (plus de 85% de bonne classification), alors même que nous ne pouvons travailler pour le moment que sur un jeu de données assez restreint, la promotion en première année d'IUT étant limitée à une centaine d'étudiants. Sur les différents modèles qui ont été utilisés pour ce prototype, les RF ont donné les résultats les plus intéressants. Cette étape de notre travail est en cours de publication. Le sujet de thèse que nous présentons a pour ambition d'aller plus loin dans nos apprentissages, et de nous attaquer à des problèmes plus spécifiques qui nous restreignent pour le moment dans nos résultats.

Défis de la thèse

Une partie des informations que nous utilisons pour notre projet correspond à des informations qui sont quantitatives (les notes). Dans le cadre de notre projet, pour lequel nous collaborons avec le Rectorat de Nice, nous avons pour objectif de récupérer également des informations qualitatives sur le niveau des lycées ainsi que les commentaires écrits par les professeurs dans les bulletins scolaires. Si les données permettant de représenter le niveau des lycées peuvent facilement être transformées en données qualitatives en prenant notamment en compte le taux de réussite au bac, le nombre de mentions, etc. Cela semble plus complexe pour ce qui est de l'analyse des commentaires écrits par les professeurs.

Concernant ce dernier point, nous pourrons utiliser des techniques de traitement de texte que nous avons déjà pu maîtriser dans le cadre de la thèse de Mme Ameni Bouaziz (« Méthodes d'apprentissage interactif pour la classification des messages courts » soutenue le 09/06/2017). Ce travail nous a permis d'acquérir des compétences multiples qui seront nécessaires au traitement de ces commentaires : Latent Semantic Analysis [8], stopwords, k-grammes [9], etc. Ces techniques sont utiles pour supprimer tous les mots dénués de sens dans les commentaires (comme certains mots de liaison), de reconnaître des mots ou des groupes de mots, de remplacer les mots reconnus par leur racine (ainsi un verbe conjugué sera remplacé par la racine du verbe à l'infinitif).

Outre l'utilisation de techniques plus ou moins élaborées de traitement du texte que nous venons de citer, nous avons également proposé lors de la thèse de Mme Ameni Bouaziz une nouvelle approche de forêt aléatoire qui intègre la sémantique des mots à la fois dans la modélisation du texte mais également dans le processus même de création de la forêt [6]. En effet, le jeu de données utilisé lors de cette thèse était composé de phrases très courtes. Nous avons donc utilisé des techniques basées sur des dictionnaires externes de mots et sur des regroupements sémantiques pour enrichir le jeu de données. Une fois le jeu de données enrichi, nous avons modifié le principe de base des RF, le Random Feature Selection, qui consiste à choisir la meilleure caractéristique pour un nœud d'un arbre parmi un sous ensemble de caractéristiques choisies au hasard. Nous avons ainsi considéré non seulement

des caractéristiques choisies au hasard mais également d'autres caractéristiques avec un lien sémantique.

Un des défis que nous allons avoir à traiter lors de cette nouvelle thèse consistera à élaborer une approche de RF qui puisse combiner ces informations hétérogènes qui sont à la fois quantitatives (les notes) et qualitatives (les commentaires, les niveaux des lycées). Une étude complète de toutes ces caractéristiques et de leur influence doit être réalisée. Cette étude doit non seulement nous indiquer les caractéristiques à prendre en compte mais également la manière dont nous devrons en tenir compte, en utilisant notamment des pondérations lors du processus de sélection des caractéristiques, comme le montrent les travaux présentés dans [10].

Une autre problématique que nous avons à prendre en compte est liée au fait que la répartition des étudiants dans les différentes catégories que nous considérons peut ne pas être équitable : les classes que nous considérons pour la classification peuvent être réussite/échec, ou réussite/résultats moyens/échec. Dans les 2 cas les étudiants ne se répartissent pas uniformément dans ces différentes catégories. En apprentissage supervisé on parle de jeu de données mal équilibré (unbalanced dataset). Or, ce type de déséquilibre dans les classes qui servent de base aux algorithmes altère grandement les performances de ces mêmes algorithmes. Nous devons donc adapter notre approche pour pallier ce type de problème. Pour cela différentes solutions peuvent être étudiées, comme de sous-échantillonner les classes sur-représentées, sur-échantillonner les classes sous-représentées, d'utiliser des RF dédiés à ce type de problème (Weigthed Random Forest [12]). Une approche que nous souhaitons étudier est celle qui consiste à affecter des poids variables aux éléments de notre dataset. En effet, le nombre d'étudiants en échec étant, fort heureusement, moins élevé que le nombre d'étudiants qui réussissent, nous souhaitons leur donner un poids plus important dans le processus d'apprentissage.

Enfin, la nature même de notre approche implique de prendre en compte des données qui se renouvellent chaque année à mesure que de nouveaux étudiants intègrent les formations du supérieur. Une approche basique consisterait à recréer chaque année un modèle en intégrant les nouvelles données aux données d'apprentissage des forêts aléatoires. Une autre approche plus fine consiste à ne pas recréer toute la forêt à chaque nouvelle vague d'étudiants mais de l'affiner en conservant les arbres qui classifient correctement et en remplaçant les arbres qui se trompent trop souvent. Un travail similaire avait été réalisé dans la thèse de Mme Bouaziz [11]. Lors de ce travail, des métriques fines étaient utilisées pour caractériser les arbres qui se trompaient trop et lors de différentes itérations ceux-ci étaient remplacés par de nouveaux arbres (accuracy, erreur de classification, F1-mesure, inégalité de Hoeffding, etc.). La manière dont le modèle va se mettre à jour peut aussi se faire de manière différenciée. Ainsi, pour les données issues d'une nouvelle année universitaire, au lieu de considérer toutes les données sur le même plan nous pourrions décider de donner plus de poids aux données pour lesquelles notre modèle se serait trompé.

Attendus du projet

Une meilleure prise en compte de données qualitatives telles que les commentaires des professeurs dans les bulletins nous semble un point essentiel pour affiner notre connaissance de la réussite ou non des étudiants.

Si nous nous plaçons au niveau des acteurs des formations du supérieur qui vont recruter des futurs étudiants, ceux-ci sont souvent confrontés à un fort taux d'échec dans leurs formations. Avoir un outil qui permette d'identifier, dès le début de l'année universitaire, les étudiants qui pourraient rencontrer des difficultés permet aux responsables d'adapter leur pédagogie afin d'anticiper ces problèmes. Cette pédagogie adaptée peut prendre différentes formes : mise en place de modules de mise à niveau, parcours adaptatif, heures de soutien, mise en place de tutorat, etc.

Du point de vue des professeurs principaux et psychologues de l'Education Nationale spécialistes de l'orientation officiant dans les lycées, un tel outil serait de notre point de vue très utile. En effet, lorsqu'un lycéen va demander des conseils, ces personnes vont pouvoir avoir une idée objective de ses chances de réussite dans les formations qu'il vise en fonction de ses notes, commentaires dans les bulletins, etc. Il pourra ainsi aider le lycéen à confirmer ses choix ou au contraire l'alerter sur d'éventuelles difficultés qu'il pourra rencontrer.

Du point de vue des lycéens enfin, un tel système pourra, comme nous venons de le dire, lui permettre d'envisager en toute connaissance de cause de candidater à certaines formations du supérieur. De plus, cela pourra apporter une aide non négligeable à la fois à des lycéens qui surestimeraient leur capacité à s'adapter à certaines formations mais également des lycéens qui s'autocensureraient en pensant à tort que certaines formations leur seraient interdites.

Références bibliographiques

1. Gomes, Heitor Murilo & Bifet, Albert & Read, Jesse & Barddal, Jean Paul & Enembreck, Fabricio & Pfahringer, Bernhard & Holmes, Geoff & Abdessalem, Talel. (2017). Adaptive random forests for evolving data stream classification. Machine Learning. 1-27. 10.1007/s10994-017-5642-8.
2. V. T. N. Chau and N. H. Phung, "Imbalanced educational data classification: An effective approach with resampling and random forest, " in Proc. of the 2013 IEEE RIVF Int. Conf., 2013.
3. Spoon K, Beemer J, Whitmer JC, Fan J, Frazee JP, Stronach J, Bohonak AJ, Schmidt-Thieme L (2016) Random forests for evaluating pedagogy and informing personalized learning. J Educ Data Min 8(2):20–50
4. Lu Thi Kim Phung and Thi Ngoc Chau Vo and Hua Phung Nguyen , Extracting Rule RF in Educational Data Classification: From a Random Forest to Interpretable Refined Rules, 2015 International Conference on Advanced Computing and Applications (ACOMP), 2015, p. 20-27.
5. Halim, C. Dartigues-Pallez, F. Precioso, M. Riveill, A. Benslimane, S. Ghoneim, "Human action recognition based on 3D skeleton part-based pose estimation and temporal multi-resolution analysis", Proc. ICIP, pp. 3041-3045, Sep. 2016.
6. Ameni Bouaziz, Christel Dartigues-Pallez, Célia Da Costa Pereira, Frédéric Precioso, Patrick Lloret. Short Text Classification Using Semantic Random Forest. Data Warehousing and Knowledge Discovery, Sep 2014, Munich, Germany. 8646, 2014, Lecture Notes in Computer Science
7. Luis Gioanni, Christel Dartigues-Pallez, Stéphane Lavirotte, Jean-Yves Tigli. Opportunistic Human Activity Recognition: a study on Opportunity dataset 13th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, Nov

2016, Hiroshima, Japan. Proceeding of the 13th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services

8. Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. Journal of the American society f
9. Huan Liu and Hiroshi Motoda. Feature extraction, construction and selection : A data mining perspective. Springer Science & Business Media, 1998. (Cité en page 25.)
10. Computational Linguistics and Chinese Language Processing Vol. 13, No. 4, December 2008, pp. 387-404 The Association for Computational Linguistics and Chinese Language Processing
11. Ameni Bouaziz. Méthodes d'apprentissage interactif pour la classification de messages courts. Thèse de l'Université de Nice Côte d'Azur soutenue le 09/07/2017.
12. Winham SJ, Freimuth RR, Biernacka JM. A Weighted Random Forests Approach to Improve Predictive Performance. *Stat Anal Data Min*. 2013;6(6):496–505. doi:10.1002/sam.11196