Défi EGC 2017: Modélisation Cost-Sensitive et Enrichissement de données

Vincent Levorato*,**, Michel Lutz***,****
Matthieu Lagacherie*

*OCTO Technology
vlevorato@octo.com, mlagacherie@octo.com
**LIFO, Université d'Orléans
vincent.levorato@univ-orleans.fr
***TOTAL, Digital Corporate Team
michel.lutz@total.com
****Ecole des mines de Saint-Etienne

Résumé. La conférence EGC'2017 propose un défi dont le contexte est la gestion des espaces verts pour la ville de Grenoble, et notamment des arbres qui y sont présents. L'objectif est de proposer un modèle basé sur des données fournies qui permettrait de prédire au mieux les arbres malades, ainsi que la localisation potentielle de la maladie. Après avoir obtenu quelques résultats intéressants avec des modèles standards, notre approche utilisant un modèle Cost-Sensitive One Against All (CSOAA) nous permet d'obtenir une exactitude de 0,86, une précision de 0,88, et un rappel de 0,91 sur la prédiction unilabel, et une précision/rappel micro de 0,82/0,74 ainsi qu'une précision/rappel macro de 0,66/0,46 pour la prédiction multilabel. L'extraction de connaissances pour la tâche 2 nous a permis de mettre en relief l'intérêt de l'ajout de données sur la nature des maladies et la concentration de la pollution dans la ville.

1 Introduction et Données

Le travail présenté dans cet article répond au Défi EGC 2017, dont l'enjeu est, pour la ville de Grenoble, de prédire l'apparition de maladies parmi les arbres du parc urbain. Une partie de l'article expliquera notre approche sur les 2 types de prédiction proposés dans la tâche 1 (unilabel et multilabel). Une seconde partie proposera un enrichissement des données, suivie d'une conclusion. Avant cela, nous étudierons comment nous avons préparé les données afin de les intégrer à nos modèles.

1.1 Données disponibles

Chaque arbre possède 29 variables qui le décrivent, et 5 variables à prédire, dont une qui reflète si l'arbre est sain ou pas ('Default or not' prenant les valeurs 0 ou 1), et 4 variables spécifiant les endroits atteints ('Collet', 'Houppier', 'Racine', 'Tronc' prenant également les valeurs 0 ou 1). Parmi les données fournies, la géolocalisation de chaque arbre était présente