Using National Electronic Health Care Registries to Analyse and Predict Alcoholic Liver Disease

Dhouha Grissa*, Ditlev Nytoft Rasmussen**
Aleksander Krag**, Søren Brunak*, Lars Juhl Jensen*

*Novo Nordisk Foundation Center for Protein Research,
University of Copenhagen, Denmark.
{dgrissa, soren.brunak, lars.juhl.jensen}@cpr.ku.dk
**Department of Gastroenterology and Hepatology, Odense University Hospital.
{Ditlev.Nytoft.Rasmussen2, Aleksander.Krag}@rsyd.dk

1 Analysis of health registry data

Alcoholic liver disease (ALD) is a common chronic liver disease worldwide Mathurin and Bataller (2015). It progresses from fatty liver to alcoholic liver fibrosis (ALF)then to cirrhosis (ALC). Unfortunately, the clear majority of patients with ALD are diagnosed so late that they already have irreversible damage to their livers. It is then important to find an effective way to detect ALD at an early stage, by identifying alcohol over-using patients at risk of developing ALD. To examine this problem, we use the Danish National Patient Registry (NPR) comprising diagnoses covering the entire population of Denmark during $\sim$19 years (1996–2014). The data set covers 6.6 million patients with a total of 153 million clinical encounters, and contains $\sim$101 million unique assignments of primary and secondary diagnoses coded in the International Classification of Diseases 10th Revision. These diagnoses are combined with data on sex and birth/death dates Jensen et al. (2014). Accordingly, we use machine-learning techniques to discover relevant upstream diagnoses helping to give a new avenue of understanding how ALD develops and progresses in heavy drinker patients.

The analysis of NPR data lead to the identification of a cohort of 33,391 patients with ALD, from which we derive groups of patients with ALC and ALF. From each group, we extract 2-year upstream time window of the first discharge date of ALC/ALF diagnosis, and obtain three sets of patients: (i) $10,831$ of patients with ALC; (ii) 12 patients with ALF but not ALC; and (iii) 23 patients with ALF at least 6 months before ALC. These numbers underline the main motivation for this work, namely that the clear majority of ALD patients are discovered only after ALF has developed into irreversible ALC.

2 Learn from ALC and test on ALF

Based on the group of 10,831 patients with ALC, we perform a matched case-control study where we randomly select for each case, 5 matched controls not suffering from ALD if available. We then transform their NPR data into a binary dataset of the form 52,926 patients

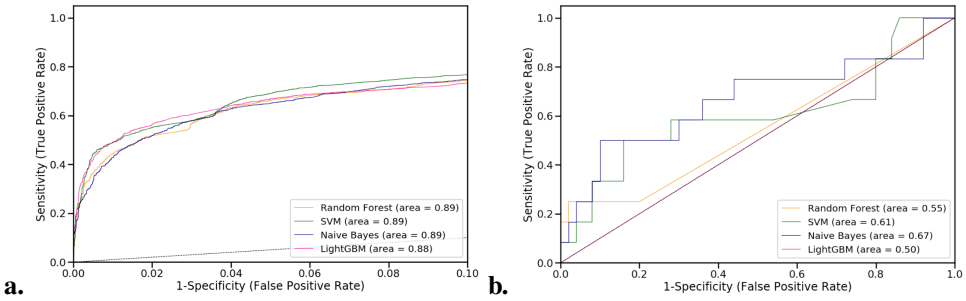

FIG. 1 – *The performance of the four classification models: RF, SVM, LightGBM and NB according to the ROC curve.* ***[a.]*** *presents the ROC curve based on the set of ALC patients and their matched controls.* ***[b.]*** *presents the ROC curve based on the four models trained on the set of ALC patients, and tested on the set of ALF patients and their matched controls.*

and 969 diagnoses. The entire binary dataset is then sampled into a training set (75% of the samples), and a test set (25% of the samples), on which we apply 4 classifiers (RF, SVM, LightGMB and NaiveBayes). Their performance is evaluated via the ROC curve of Figure 1a.. Accordingly, there is no unique classifier that highly outperforms all the others (area under ROC=0.88 to 0.89). When testing the same trained models on the small set of ALF patients and their matched controls (60 patients $\times$ 969 diagnoses), their performance obviously drops a lot. This drop reflects that ALF patients are much less sick and thus harder to identify from upstream diagnoses. Figure 1b. presents the NaiveBayes as the best performing classifier (area under ROC = 0.67). To extract the set of features that could be predictive of ALD diagnoses, we compute the feature importance for each classifier. Respectively, we sort the features according to their scores and select the top-10 from each classification model. The obtained results underscore small groups of statistically significant upstream comorbidities ($P < 0.001$) that accurately detect the set of patients with ALC and that could be promising in predicting ALF. Some of these groups are conditions either caused by alcohol-overuse as disorder behavior due to alcohol or to complications to cirrhosis, such as oesophageal varices. Others are comorbidities related to trauma and life style, as fracture and open wounds.

References

Jensen, A., P. Moseley, T. Oprea, S. Ellesøe, R. Eriksson, H. Schmock, P. Jensen, L. Jensen, and S. Brunak (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications 5*(4022).

Mathurin, P. and R. Bataller (2015). Trends in the management and burden of alcoholic liver disease. *Journal of Hepatology 62*(1), 256–260.

Acknowledgements This work was supported by the *European Union's Horizon 2020* (grant 668031) and *Novo Nordisk Foundation* (grant NNF14CC0001). It was approved by the *Danish Data Protection Agency* (ref: SUND-DT-2017-57) and the *Danish Health Authority* (ref: FSEID-00003092).

Apprentissage Ensembliste Multicouche pour la prévision à court terme de variables agro-climatiques

Jade Eva Guisiano*, Raja Chiky**
Julien Orensanz***, Shohreh Ahvar****

* ** ****Institut supérieur d'électronique de Paris, France
jade.guisiano,raja.chiky,shohreh.ahvar@isep.fr
***Cap2020, Gradignan, France
Julien.orensanz@cap2020.fr

1 Introduction et motivations

La fiabilité des prévisions météorologiques est nécessaire dans de nombreux domaines dont les activités dépendent des conditions climatiques. L'agriculture est l'un de ses domaines pouvant être particulièrement impacté par les événements météorologiques tels que les températures extrêmes, le vent, la grêle, la pluie, etc. Ces événements peuvent causer des dommages conséquents et durables sur les récoltes en entraînant la perte totale ou partielle de la production. Outre les dégâts, certaines conditions climatiques peuvent aussi affecter les opérations culturales comme par exemple la possibilité de traitement limitée en cas de vent, mais aussi l'impraticabilité du sol avec les engins agricoles en cas de pluie. L'enjeu pour les agriculteurs est donc de prendre connaissance suffisamment tôt des risques climatiques à venir afin de pouvoir mettre en place des plans d'action pour minimiser les dégâts potentiels. Pour cela, ils consultent généralement plusieurs fois par jour les prévisions météorologiques. Conscients que ces dernières ont une fiabilité limitée, les agriculteurs consultent 3, 4, parfois davantage, sources de prévisions météorologiques et arbitrent entre ces sources de façon subjective. En effet, ces sources ne performent pas forcément toujours de la même manière. Certains fournisseurs de prévisions météorologiques vont parfois "sur performer" en fournissant des prévisions proches des valeurs réellement observées et d'autres "sous performer" avec des prévisions plus éloignées des valeurs réellement observées. La précision des prévisions météorologiques de chaque fournisseur peut varier selon la période, le type de climat ou encore la zone géographique. Face à la variabilité de la fiabilité des prévisions météorologiques et aux multiples fournisseurs de prévisions dont disposent les agriculteurs, leur prise de décision n'est pas facilitée et reste de ce fait souvent imprécise. Nous proposons donc une méthode de prévision capable de fournir des prévisions de la température et de l'humidité plus fiables que les fournisseurs de prévisions météorologiques pour les 1 à 7 heures à venir pour 2 sites agricoles distants de quelques kilomètres. Notre cadre expérimental se base tout au long de l'étude sur diverses données météorologiques collectées chaque heure sur la période du 1er mars au 26 août de l'année en cours. Notre méthode, inspirée du domaine de l'apprentissage ensembliste,

permet aux agriculteurs de ne consulter plus qu'une seule source de prévisions, plus performante que celles dont ils disposent, et ainsi de ne plus avoir à arbitrer entre les divers fournisseurs de prévisions habituels.

2 Notre approche

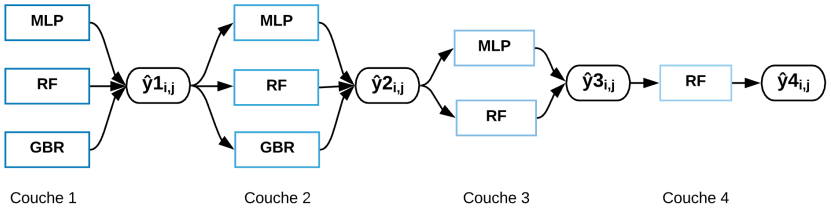

FIG. 1 : Architecture du modèle Ensembliste Multichouches.

Notre modèle d'Apprentissage Ensembliste Multichouche, représenté par la figure 1, repose principalement de la technique de Stacking. En effet, chaque couche du modèle est composée d'un ensemble de différents algorithmes d'apprentissage supervisés permettant de réaliser des régressions tels que les Perceptrons Multichouche (MLP), les Forêts Aléatoires (RF) et enfin l'Amélioration du Gradient en régression (GBR). Ces 3 algorithmes possèdent l'avantage de ne pas être sensibles aux changements d'échelle (dispensent de l'étape de normalisation et donc de la perte du sens initial de l'information) et ne requièrent pas d'hypothèses statistiques particulières. Un large éventail d'algorithmes de prévisions a été testé avec notamment les Régressions à vecteurs de support (SVR) et les Réseaux de Neurones récurrent (LSTM). Seuls les algorithmes ayant permis de fournir les prévisions les plus proches des valeurs réellement observées sont présents dans chaque couche de notre modèle. Dans notre cas d'application, notre modèle prend en entrée de sa première couche diverses variables exogènes telles que la couverture nuageuse, la pression, l'intensité du vent, etc. Puis la variable endogène de température ou d'humidité selon ce que nous voulons prévoir. Chaque algorithme de la couche première couche reçoit ces mêmes données réparties en 75% de données d'entraînement et 25% de données de test. Les prévisions $\hat{y}1_{i,j}$ issues de ces 3 algorithmes constituent les variables endogènes en entrée des algorithmes de la seconde couche toujours réparties sur le même ratio de données d'entraînement et de test. Ce processus se répète de la même manière de la couche 1 jusqu'à la couche 4. Au fur est à mesure de l'avancement dans les couches, le volume de données diminue progressivement, c'est pourquoi le choix du nombre de couches de notre modèle dépend directement du volume de données initial. La justesse de prévision des algorithmes de chaque couche est évaluée quantitativement par l'erreur quadratique moyenne (RMSE) pour prédire la température et l'humidité des 7 heures à venir. En moyenne, sur ces 7 heures de prévisions, le RMSE des prévisions de notre modèle est nettement plus faible que le RMSE des prévisions des fournisseurs météorologiques.